
بسم الله
الرحمن الرحیم

دانشکده علوم ریاضی

رشته آمار، گرایش آمارریاضی

پایان‌نامه کارشناسی ارشد

ناهمگنی فضایی در مدل‌های تعیین قیمت محصول با استفاده از یک رهیافت رگرسیون موزون موضعی تکراری

نگارنده: رویا پیرواولیاء

استاد راهنما

دکتر حسین باغیشنی

بهمن ۱۳۹۷

تقدیم به دستان مادر و خاطرات پدرم

سپاس گزاری...

سپاس خدای بزرگ را که مرا یاری رساند تا بتوانم این مقطع تحصیلی را به پایان رسانده و گامی در راستای اعتلای علم بردارم. از استاد راهنمای گران قدرم جناب آقای دکتر حسین باغیشنی که وجودشان همیشه قوتی برای انجام کارهایم بوده است و بدون شک انجام این پایان نامه بدون کمک و راهنمایی های ارزنده ایشان امکان پذیر نبود، کمال تشکر را دارم. از اساتید گرامی جناب آقای دکتر محمدرضا ربیعی و سرکار خانم دکتر نگار اقبال که زحمت داوری این پایان نامه را داشتند نیز سپاس گزارم. متشکرم از اساتید گروه آمار دانشگاه صنعتی شاهرود، آقایان دکتر داوود شاهسونی، دکتر احمد نزاکتی و دکتر محمد آرشی، که توفیق دانشجویی در محضرشان را داشتم. در پایان از تمامی عزیزانی که در طول انجام این پروژه مرا یاری کرده اند کمال تشکر و قدردانی را ابراز می نمایم.

رویا پیرواولیاء
بهمن ۱۳۹۷

تعهد نامه

این‌جانب **رویا پیرواولیاء** دانشجوی کارشناسی ارشد رشته **آمار** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **ناهمگنی فضایی در مدل‌های تعیین قیمت محصول با استفاده از یک رهیافت رگرسیون موزون موضعی تکراری** ، تحت راهنمایی دکتر **حسین باغیشنی** متعهد می‌شوم:

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

رویا پیرواولیاء
بهمن ۱۳۹۷

مالکیت نتایج و حق نشر

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

چکیده

بسیاری از مطالعات تجربی، مستلزم استفاده از متغیرهایی است که تحت تأثیر موقعیت مکانی مشاهدات هستند. در این موارد میزان و نحوه تأثیر ساختار فضایی داده‌ها دارای اهمیت است و نادیده گرفتن آن موجب خطا در برآورد و از دست رفتن اطلاعات مهمی می‌شود. در برخورد با چنین داده‌هایی مدل محبوب و پرکاربرد رگرسیون خطی پاسخگو نیست؛ زیرا عامل موقعیت فضایی موجب نقض پذیره‌های همگنی و ناهمبستگی این مدل می‌شود. بنابراین در چنین مواردی باید از مدل‌های متناسب با ساختار فضایی استفاده شود. برای داشتن یک مدل فضایی منعطف، لازم است وابستگی و ناهمگنی فضایی با هم در نظر گرفته شوند. این در حالی است که ساختار وابستگی و ناهمگنی موجود در داده‌ها نامعلوم است و اغلب هیچ‌گونه اطلاع پیشینی هم در مورد این ساختار در دسترس نیست. در این پایان‌نامه از یک رهیافت رگرسیون موزون تکراری برای تعیین ناحیه‌های همگن تحت عنوان رژیم‌های فضایی استفاده شده است. سپس این رژیم‌ها در دسته‌ای از مدل‌های اتورگرسیو فضایی معروف به مدل‌های اقتصادسنجی فضایی، اعمال و نتیجه با عنوان مدل‌های اقتصادسنجی با رژیم فضایی درونی به داده‌ها برازش داده شده‌اند. این مدل‌ها بر اساس معیار اطلاع آکاییک با مدل رگرسیون خطی و مدل‌های اقتصادسنجی فضایی مقایسه شده‌اند؛ نتایج بیانگر عملکرد بهتر مدل‌های اقتصادسنجی با رژیم فضایی درونی هستند.

کلمات کلیدی: ناهمگنی فضایی، وابستگی فضایی، رژیم فضایی، مدل‌های اقتصادسنجی، مدل‌های اقتصادسنجی با رژیم فضایی درونی.

پیش‌گفتار

تحلیل داده‌های فضایی، بر اساس یک نمونه مشاهده‌شده در ناحیه جغرافیایی تحت مطالعه در عمل می‌تواند دشوار باشد؛ چرا که شکل وابستگی و ساختار همگنی موجود در داده‌ها نامعلوم هستند. یکی از روش‌های مدل‌بندی داده‌های فضایی، استفاده از مدل‌های اتورگرسیو فضایی است که ساختار وابستگی داده‌ها را از طریق ضریب اتورگرسیو فضایی در مدل رگرسیونی لحاظ می‌کنند. این مدل‌ها در مطالعات اقتصادی نظیر علوم منطقه‌ای، اقتصاد شهری و اقتصاد کشاورزی کاربرد دارند و با عنوان مدل‌های اقتصادسنجی فضایی نیز شناخته می‌شوند.

اولین بار ویتل (۱۹۵۴) با فرض این که خطاهای هر موقعیت فقط به موقعیت‌های مجاور وابسته است، از این روش برای مدل‌بندی وابستگی فضایی استفاده کرد. بسج (۱۹۷۴)، اورد (۱۹۷۵)، انسلین (۱۹۸۸a)، هاینینگ (۱۹۹۰)، کرسی (۱۹۹۲)، لی‌سیج و پیس (۲۰۰۹) و اسمیت (۲۰۱۶) نیز به تحلیل و مدل‌بندی داده‌های فضایی با استفاده از مدل‌های اتورگرسیو فضایی پرداخته‌اند. این مدل‌ها بدون توجه به ناهمگنی فضایی برای کل فضای مورد مطالعه، یک ساختار وابستگی واحد ارائه می‌دهند.

یکی دیگر از روش‌های مدل‌بندی ساختار فضایی، که ناهمگنی و وابستگی فضایی را با ارائه یک مدل موضعی برای هر موقعیت فضایی توأم در نظر می‌گیرد، مدل رگرسیون موزون جغرافیایی (فضایی) است. از این مدل در برخی متون (غیرفضایی) با عنوان رگرسیون موزون موضعی نیز یاد می‌شود. مدل رگرسیون موزون جغرافیایی به‌طور کامل توسط فودرینگ‌هام و همکاران (۲۰۰۳) تشریح شده است. آندرانو و همکاران (۲۰۱۶) و بیل و همکاران (۲۰۱۷) دو روش مشابه برای مدل‌بندی ساختار فضایی داده‌ها ارائه کردند. در این روش‌ها وابستگی و ناهمگنی فضایی در قالب رژیم‌های فضایی مختلف مدل‌بندی می‌شوند.

با این مقدمه، با هدف معرفی یک مدل ایده‌آل برای مدل‌بندی ساختار داده‌های فضایی، پایان‌نامه خود را در چهار فصل و یک پیوست ارائه می‌کنم.

۱. فصل اول به مروری بر رگرسیون خطی، پذیره‌ها و راه‌کارهای موجود برای برخورد با عدم برقراری پذیره‌های ناهمبستگی و همگنی در آن، اختصاص دارد. همچنین به دلیل اهمیت اثر عامل موقعیت فضایی در نقض این پذیره‌ها به مقدمه‌ای از آمار فضایی و همچنین ارائه آزمون‌هایی برای بررسی حضور ناهمگنی و وابستگی فضایی، پرداخته شده است.

۲. در فصل دوم، چند روش معمول مدل‌بندی ساختار وابستگی و ناهمگنی فضایی، روش‌های برآورد پارامترها و معایب هر کدام از آن‌ها بیان شده‌اند.

۳. در فصل سوم، مدل پیشنهادی بیل و همکاران (۲۰۱۷) برای مدل‌بندی هم‌زمان ساختار وابستگی و ناهمگنی فضایی ارائه و تشریح شده است.

۴. در فصل چهارم، کارایی بهتر مدل پیشنهادی فصل سوم نسبت به مدل‌های معرفی‌شده در فصل دوم در قالب دو مثال از داده‌های واقعی تعیین قیمت مسکن نمایش داده شده است.

۵. در پیوست نیز بسته‌افزارها و کدهای نرم‌افزار R برای تحلیل داده‌ها و برازش مدل‌های پیشنهادی، به‌صورت مرحله به مرحله، ارائه شده‌اند.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مفاهیم و مقدمات

۱.۱ مقدمه

تجربه نشان داده است که تحلیل مسائل اقتصادی اغلب با چالش‌هایی مواجه است. چرا که داده‌ها به شکل‌های ناشناخته‌ای ناهمگن[۱] و به یکدیگر وابسته‌اند. در موارد متعددی این ناهمگنی و وابستگی داده‌ها حاصل از فضا و موقعیت قرارگیری آن‌ها نسبت به یکدیگر است. به این نوع داده‌ها، داده‌های فضایی[۲] می‌گویند (انسلین، ۱۹۸۸a). دیدگاه‌های مختلفی برای مدل‌بندی داده‌های فضایی پیشنهاد شده‌اند. یکی از دیدگاه‌ها مدل‌های اقتصادسنجی فضایی[۳] است که به‌طور گسترده برای مدل‌بندی ساختار وابستگی داده‌ها معرفی و استفاده می‌شوند. اما برخورد مستقیم با مشکل ناهمگنی[۴] داده‌های فضایی معمولا کنار گذاشته و نادیده گرفته می‌شود. در این پایان‌نامه، قصد داریم که هر دو این ویژگی‌های واقعی داده‌ها را در یک مدل رگرسیونی لحاظ کنیم. این کار مستلزم معرفی برخی مفاهیم و تعاریف اولیه است که در این فصل ارائه می‌کنیم.

---

[1] Heterogeneous
[2] Spatial data
[3] Spatial econometric models
[4] Heterogeneity

۲.۱ مدل رگرسیون خطی

ساده‌ترین و کاربردی‌ترین مدل رگرسیونی، رگرسیون خطی است که معادله آن در حالت کلی به‌صورت

$$y_i = \beta_۰ + \beta_۱ X_{i۱} + \beta_۲ X_{i۲} + \cdots + \beta_p X_{ip} + \varepsilon_i \quad i = ۱, ۲, \cdots, n \tag{۱.۱}$$

است (مونت‌گومری و همکاران، ۲۰۱۲). همچنین می‌توان معادله (۱.۱) را به صورت ماتریسی زیر بازنویسی کرد:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{۲.۱}$$

که در آن $\mathbf{y} = (y_۱, \cdots, y_n)^T$ بردار متغیرهای پاسخ، $\mathbf{X}$ ماتریس متغیرهای تبیینی با بعد $n \times (p+۱)$، و $\boldsymbol{\beta}$ بردار ضرایب نامعلوم رگرسیونی با بعد $(p+۱) \times ۱$ است که بیان‌گر چگونگی رابطه خطی بین متغیرهای تبیینی و متغیر پاسخ است. همچنین $\boldsymbol{\varepsilon}$ بردار $n \times ۱$ خطاهای تصادفی است.

در رگرسیون خطی، برای اعتبار استنباط‌های آماری انجام‌شده، پذیره‌هایی را در نظر می‌گیرند. این پذیره‌ها شامل موارد زیر هستند:

۱ـ رابطه بین $\mathbf{X}$ و $\mathbf{y}$ خطی بوده و متغیرهای تبیینی مستقل از یکدیگرند.

۲ـ میانگین جمله خطا برابر صفر است، یعنی $\mathrm{E}(\boldsymbol{\varepsilon}) = ۰$.

۳ـ واریانس جمله خطا ثابت است، یعنی $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^۲\mathbf{I}$. این پذیره بیان می‌کند که واریانس $\boldsymbol{\varepsilon}$ یا $\mathbf{y}$ به مقادیر $\mathbf{X}$ بستگی ندارد. این پذیره به واریانس متجانس[۵] (همگن[۶]) معروف است.

۴ـ جملات خطا ناهم‌بسته‌اند.

۵ـ جملات خطا دارای توزیع نرمال هستند.

تحت پذیره‌های شماره ۴ و ۵، جمله‌های خطای $\varepsilon_۱, \cdots, \varepsilon_n$ مستقل هستند. به‌منظور برآورد ضرایب رگرسیونی، یک روش معمول استفاده از روش کمترین توان‌های دوم معمولی[۷] (OLS) است (مونت‌گومری و همکاران، ۲۰۱۲)، که در ادامه آن را بازگو می‌کنیم.

[5] Homoscedastic
[6] Homogenuos
[7] Ordinary least square

۱.۲.۱ روش کمترین توان‌های دوم معمولی

در این روش، ضرایب رگرسیونی در مدل (۲.۱) از طریق کمینه کردن مجموع توان‌های دوم خطا برآورد می‌شوند. برای این منظور فرض می‌شود که همواره $n > p$ و همه خطاها ناهم‌بسته و دارای میانگین صفر و واریانس ثابت هستند. تابع هدف کمترین توان‌های دوم به‌صورت زیر است:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} \\ &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^T\mathbf{y}-۲\mathbf{y}^T\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

و برآورد کمترین توان‌های دوم ضرایب رگرسیونی از حل معادله زیر نتیجه می‌شود:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -۲\mathbf{y}^T\mathbf{X}+۲\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}=۰.$$

بنابراین

$$\begin{aligned} \mathbf{y}^T\mathbf{X} &= \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X} \\ \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-۱} &= \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-۱} \end{aligned}$$

و در نهایت

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-۱}\mathbf{X}^T\mathbf{y}. \qquad (۳.۱)$$

همان‌طور که گفته شد، برای اعتبار نتایج استنباط در مدل رگرسیون خطی باید چند پذیره پایه‌ای برای جمله خطای مدل برقرار باشند. اما در عمل و کاربرد ممکن است که هریک از این پذیره‌ها برقرار نباشند. به‌طور ویژه ممکن است پذیره همگنی جمله‌های خطا معتبر نباشد. در این صورت برای مثلا برآورد ضرایب، دیگر مجاز به استفاده از روش OLS نخواهیم بود و در صورت عدم توجه به این موضوع از کارایی و صحت مدل کاسته می‌شود. بنابراین ناهمگنی جمله‌های خطا باید در مدل لحاظ شود و این کار در رگرسیون از طریق انجام تبدیلات مناسب یا تغییر روش برآورد پارامترها امکان‌پذیر است.

در ادامه روش‌های معمول برخورد با ناهمگنی واریانس جمله خطا را مطرح می‌کنیم.

۲.۲.۱ راه‌های برخورد با ناهمگنی در مدل رگرسیون خطی کلاسیک

در این بخش به بیان متداول‌ترین روش‌های برخورد با ناهمگنی واریانس در مدل‌های رگرسیون خطی کلاسیک می‌پردازیم.

تبدیل باکس‌ـ کاکس

در صورتی که جمله خطا و در نتیجه $\mathbf{y}$، نانرمال یا دارای واریانس ناهمگن باشد، تعدیل توانی $\mathbf{y}^{\lambda}$ یک راه مناسب برای رفع آن‌ها است. باکس و کاکس (۱۹۶۴) نشان دادند که پارامترهای رگرسیون و λ می‌توانند هم‌زمان، با استفاده از روش درستنمایی ماکسیمم برآورد شوند. اما این تبدیل در حالتی که λ برابر صفر شود، دچار مشکل می‌شود. لذا تبدیل مناسب برای استفاده را به صورت

$$\mathbf{y}^{(\lambda)} = \begin{cases} \frac{\mathbf{y}^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y}\ln\mathbf{y} & \lambda = 0 \end{cases}$$

معرفی کردند، که در آن $\dot{y} = \ln^{-1}[1/n\Sigma\ln y_i]$ میانگین هندسی[8] مشاهدات است و مدل

$$\mathbf{y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

با روش کمترین توان‌های دوم یا درستنمایی ماکسیمم برآورد می‌شود.

تبدیل لگاریتمی

این تبدیل حالت خاصی از تبدیل باکس‌ـ کاکس ($\lambda = 0$) است. با تبدیل لگاریتمی غالبا ناهمسانی واریانس کاهش می‌یابد، زیرا مقیاس‌های اندازه‌گیری متغیرها را تحت تاثیر قرار می‌دهد. اما این تبدیل در مواردی که بعضی از مقادیر X و y صفر یا منفی باشند، کاربرد ندارد. از طرفی این تبدیل می‌تواند منجر به تولید همبستگی ساختگی شود.

کمترین توان‌های دوم تعمیم‌یافته

زمانی که جمله خطا دارای خودهمبستگی و واریانس ناهمگن است، یعنی $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ و

$$\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$$

که در آن $\mathbf{V}$ یک ماتریس متقارن غیرقطری است، می‌توان از روش کمترین توان‌های دوم تعمیم‌یافته[9] (GLS) برای برآورد ضرایب رگرسیونی استفاده کرد. در این روش با تبدیل داده‌ها به یک مجموعه داده جدید که دارای پذیره‌های لازم برای به کارگیری روش OLS است، خودهمبستگی و ناهمگنی در برآورد وارد می‌شود. البته چگونگی ساختار ماتریس $\mathbf{V}$ موجب ایجاد مشکل در این روش می‌شود، اما با فرض این که $\mathbf{V}$ نامنفرد[10] و مثبت باشد، ماتریس نامنفرد و متقارن $\mathbf{K}$ را می‌توان طوری تعیین کرد که $\mathbf{KK} = \mathbf{K}^T\mathbf{K} = \mathbf{V}$ و مدل جدید به‌صورت زیر تعریف می‌شود:

$$\mathbf{z} = \mathbf{B}\boldsymbol{\beta} + \mathbf{g}$$

[8] Geometric mean

[9] Generalized least squares

[10] Nonsingular

که در آن

$$\mathbf{z} = \mathbf{K}^{-1}\mathbf{y}, \qquad \mathbf{B} = \mathbf{K}^{-1}\mathbf{X}, \qquad \mathbf{g} = \mathbf{K}^{-1}\boldsymbol{\varepsilon}.$$

در این مدل جدید پذیره‌های پایه روش OLS برقرار هستند. یعنی

$$\mathrm{E}(\mathbf{g}) = \mathrm{E}(\mathbf{K}^{-1}\boldsymbol{\varepsilon}) = \mathbf{K}^{-1}\mathrm{E}(\boldsymbol{\varepsilon}) = 0$$

و

$$\begin{aligned}\mathrm{Var}(\mathbf{g}) &= \mathrm{Var}(\mathbf{K}^{-1}\boldsymbol{\varepsilon}) = \mathbf{K}^{-1}\mathrm{Var}(\boldsymbol{\varepsilon})(\mathbf{K}^{-1})^T \\ &= \mathbf{K}^{-1}\sigma^2\mathbf{V}(\mathbf{K}^T)^{-1} = \mathbf{K}^{-1}\sigma^2\mathbf{V}(\mathbf{K})^{-1} \\ &= \sigma^2\mathbf{K}^{-1}\mathbf{K}\mathbf{K}\mathbf{K}^{-1} = \sigma^2\mathbf{I}.\end{aligned}$$

بنابراین فرآیند برآورد همانند OLS انجام می‌شود. به عبارت دقیق‌تر

$$\begin{aligned}S(\beta) &= \mathbf{g}^T\mathbf{g} \\ &= \boldsymbol{\varepsilon}^T\mathbf{V}^{-1}\boldsymbol{\varepsilon} \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

و بنابراین

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}. \qquad (۴.۱)$$

کمترین توان‌های دوم موزون

مدل رگرسیون خطی با واریانس خطای ناهمگن را می‌توان با روش کمترین توان‌های دوم موزون[11] (WLS) برازش داد. در این روش برآورد، تفاضل بین مقدار واقعی y_i و مقدار برآوردشده آن در وزن w_i ضرب می‌شود. این روش حالت خاصی از روش GLS است که در آن درایه‌های غیر قطر اصلی صفر هستند. تابع هدف روش WLS به صورت زیر نوشته می‌شود:

$$\begin{aligned}S(\boldsymbol{\beta}) &= \boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{W}\mathbf{y} - \mathbf{y}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^T\mathbf{W}\mathbf{y} - 2\mathbf{y}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

و برآورد کمترین توان‌های دوم موزون از حل معادله زیر نتیجه می‌شود:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{y}^T\mathbf{W}\mathbf{X} + 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{X} = 0.$$

[11] Weighted least squares

بنابراین

$$\mathbf{y}^T\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} = \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$$

و در نهایت

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}. \qquad (۵.۱)$$

برای استفاده از این روش، ماتریس وزن $\mathbf{W}$ باید مشخص شود. معمولا وزن‌ها متناسب با معکوس واریانس جمله خطا تعریف می‌شوند؛ یعنی به مشاهدات با واریانس بزرگ وزن کوچک‌تر و به مشاهدات با واریانس کوچک وزن بزرگ‌تر تخصیص می‌یابد. برای مثال، اگر $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$ آن‌گاه $w_i = n_i$ انتخاب می‌شود. به‌طور کلی، اگر $\mathrm{Var}(\varepsilon_i) = \mathrm{Var}(y_i) = \frac{\sigma^2}{n_i}$ باشد که در آن $\mathbf{V}$ یک ماتریس قطری با بعد n است، آن‌گاه $\mathbf{W} = \mathbf{V}^{-1}$.

یکی دیگر از پذیره‌های پایه‌ای رگرسیون خطی، ناهمبسته بودن جمله خطاست. این پذیره نیز در بسیاری موارد در برخورد با داده‌های واقعی، برقرار نیست؛ یعنی جمله‌های خطا به شکلی به هم وابسته‌اند. این وابستگی در داده‌های سری زمانی (داده‌هایی که در طول زمان جمع‌آوری شده‌اند)، داده‌های طولی و داده‌های فضایی وجود دارد. در این پایان‌نامه، تمرکز بر داده‌های فضایی است؛ لذا برای درک بهتر ساختار وابستگی در این داده‌ها، در بخش بعد به بیان مفاهیم و مقدمات آمار فضایی می‌پردازیم.

۳.۱ مقدمه‌ای بر آمار فضایی

اغلب روش‌های معمولی آمار شامل رگرسیون خطی، OLS و WLS که در بخش قبل مطرح شدند مبتنی بر استقلال مشاهدات است، اما این پذیره در عمل در بسیاری از موارد برقرار نیست. کاربردهای مختلفی وجود دارند که در آن‌ها مشاهدات به یکدیگر وابسته بوده و نادیده گرفتن این وابستگی، سبب از بین رفتن اطلاعات مفیدی می‌شود و در نتیجه، تحلیل‌های آماری دور از واقعیت خواهند بود. در چنین مواقعی روش‌های معمول کارایی خود را از دست داده و باید روش دیگری را برای تحلیل مناسب داده‌ها انتخاب کرد.

این مسأله که مشاهده مربوط به یک موقعیت مکانی به مشاهدات موقعیت‌های مجاور خود وابسته باشد، کاملا منطقی به نظر می‌رسد. این امر در داده‌های سری زمانی نیز به نوعی دیده می‌شود، با این تفاوت که در داده‌های زمانی به دلیل تک بعدی بودن زمان، مشاهدات در طول محور زمان، به‌طور متوالی و به ترتیب جمع‌آوری می‌شوند و در نتیجه وابستگی در مشاهدات پیاپی ظاهر می‌شود. اما گاهی وابستگی داده‌ها به وضعیت قرار گرفتن آن‌ها در فضای جغرافیایی تحت مطالعه بستگی دارد و در واقع وابستگی آن‌ها تابعی از فاصله مکانی آن‌ها از هم است. در برخورد با چنین داده‌هایی با دو دسته از عوامل روبه‌رو هستیم، که به

مکان استقرار متغیرها مربوط می‌شوند و می‌توان آن‌ها را تحت عناوین وابستگی فضایی[۱۲] یا به بیان ضعیف‌تر خودهمبستگی فضایی[۱۳] و ناهمگنی فضایی[۱۴] توضیح داد. به شاخه‌ای از علم آمار که به بررسی و تحلیل این نوع داده‌ها می‌پردازد، آمار فضایی گفته می‌شود. در واقع در آمار فضایی سعی بر این است که وابستگی بین مقادیر مختلف یک متغیر، از حیث فاصله و جهت قرار گرفتن آن‌ها نسبت به هم، در مدل‌بندی لحاظ شود. این ارتباط فضایی که معمولا در قالب روابط ریاضی بیان می‌شود، ساختار فضایی نام دارد.

در ادامه تعاریف و مفاهیم اولیه آمار فضایی را مطرح می‌کنیم. این مطالب از محمدزاده (۱۳۹۱)، انسلین (۱۹۸۸a) و لی‌سیج و پیس (۲۰۰۹) گرفته شده‌اند.

۱.۳.۱ داده‌های فضایی

در آمار فضایی متغیر مورد اندازه‌گیری ممکن است گسسته یا پیوسته و موقعیت مشاهدات نیز ممکن است گسسته یا پیوسته، نقطه‌ای یا ناحیه‌ای، منظم یا نامنظم باشد. مشاهدات فضایی با توجه به انواع موقعیت‌ها به سه گروه تقسیم می‌شوند:

زمین‌آماری[۱۵]: این نوع داده‌ها در موقعیت‌های ثابت و مشخص در ناحیه‌ای پیوسته مشاهده می‌شوند. متغیر مورد بررسی ممکن است گسسته یا پیوسته باشد. در داده‌های زمین‌آماری معمولا پیش‌گویی مقدار متغیر در یک موقعیت جدید مد نظر است.

مشبکه‌ای[۱۶]: این نوع داده‌ها مربوط به مکان‌های ناحیه‌ای هستند، که این مکان‌ها ممکن است منظم یا نامنظم باشند. نواحی می‌توانند شامل شهرهای یک استان، مناطق شهرداری یک شهر و موارد مشابه باشند. به‌طور معمول هدف از تحلیل داده‌های مشبکه‌ای مدل‌بندی احتمالاتی مشاهدات است.

الگونقطه‌ای[۱۷]: در این حالت مکان یا موقعیت مشاهده‌شده خود متغیری تصادفی است. الگوهای نقطه‌ای شامل تعدادی متناهی از مکان‌ها در یک ناحیه‌اند که در آن‌ها یک صفت خاص اندازه‌گیری می‌شود. این داده‌ها به سه دسته به‌طور کامل تصادفی فضایی، منظم و خوشه‌ای تقسیم و به مدل‌بندی آن‌ها اقدام می‌شود.

---

[12] Spatial dependence
[13] Spatial autocorrelation
[14] Spatial heterogeneity
[15] Geostatistical
[16] Lattice
[17] Point pattern

۲.۳.۱ مدل آماری

برای تحلیل داده‌های فضایی نیازمند یک مدل آماری هستیم. در آمار فضایی معمولا از یک میدان تصادفی به‌صورت

$$Z(s) = \mu(s) + \delta(s), \qquad s \in \mathrm{D}$$

برای مدل‌بندی داده‌ها استفاده می‌شود، که در آن D یک زیرمجموعه از فضای اقلیدسی d بعدی، $d \geq ۱$، از $\mathbb{R}^d$ است. مولفه $\mu(\cdot)$ تغییرات بزرگ‌مقیاس[۱۸] یا روند[۱۹] و $\delta(\cdot)$ فرایند خطا یا تغییرات کوچک‌مقیاس[۲۰] نامیده می‌شوند. تغییرات کوچک‌مقیاس ممکن است ناشی از خطای اندازه‌گیری یا تغییر در درون موقعیت مشاهده‌شده و تغییرات بزرگ‌مقیاس ممکن است ناشی از تغییرپذیری بین موقعیت‌های مشاهده‌شده باشد.

در ادامه به معرفی برخی ویژگی‌های میدان تصادفی که موجب ساده‌سازی مسأله می‌شود، می‌پردازیم. باید توجه داشت که با وجود این‌که تمامی این پذیره‌ها موجب ساده‌سازی مسأله می‌شود، اما در صورت عدم برقراری، استفاده از مدل‌های فضایی بر پایه این پذیره‌ها موجب نامناسب شدن برآورد ساختار فضایی میدان و کاهش دقت نتایج حاصل از تحلیل داده‌های فضایی از جمله رگرسیون می‌شود.

تعریف ۱.۳.۱. میدان تصادفی $Z(\cdot)$ ایستای ذاتی[۲۱] نامیده می‌شود، هرگاه میانگین میدان تصادفی مستقل از s باشد، یعنی $\mathrm{E}(Z(s)) = \mu$ و واریانس تفاضل مقادیر $Z(s_۱)$ و $Z(s_۲)$ فقط تابعی از فاصله موقعیت‌های $s_۱$ و $s_۲$ باشد، یعنی

$$\mathrm{Var}(Z(s_۱) - Z(s_۲)) = ۲\gamma(s_۱ - s_۲), \qquad s_۱, s_۲ \in \mathrm{D} \subset \mathbb{R}^d.$$

تعریف ۲.۳.۱. میدان تصادفی $Z(\cdot)$، ایستای مرتبه دوم[۲۲] نامیده می‌شود، هرگاه میانگین آن ثابت و کواریانس $Z(s_۱)$ و $Z(s_۲)$، فقط تابعی از فاصله موقعیت‌ها باشد. تحت ایستایی مرتبه دوم، واریانس میدان تصادفی به موقعیت فضایی بستگی ندارد و تغییرپذیری میدان در همه جا یکسان است. یعنی

$$\mathrm{Var}(Z(s)) = \mathrm{Cov}(Z(s), Z(s)) = C(۰) = \sigma^۲.$$

تعریف ۳.۳.۱. در صورتی که توزیع میدان تصادفی تحت هر فاصله h ثابت بماند، آن میدان را ایستای قوی[۲۳] می‌گوییم. به عبارتی

$$(Z(s_۱), \ldots, Z(s_n)) \overset{\mathrm{D}}{=} (Z(s_۱ + h), \ldots, Z(s_n + h))$$

[18] Large scale
[19] Trend
[20] Small scale
[21] Intrinsic stationary
[22] Second order stationary
[23] Strong stationary

که منظور از $\overset{D}{=}$ هم‌توزیعی بردار متغیرهای تصادفی است.

تعریف ۴.۳.۱. در صورتی که توزیع توام هر تعداد متناهی از متغیرهای میدان تصادفی $Z(\cdot)$ نرمال چندمتغیره باشد، آن را میدان تصادفی گاوسی[۲۴] گویند.

۳.۳.۱ ساختار همبستگی فضایی

در آمار فضایی مفاهیمی مشابه واریانس و کواریانس در آمار کلاسیک، برای بیان ساختار همبستگی فضایی وجود دارند.

تعریف ۵.۳.۱. واریانس تفاضل مقادیر میدان تصادفی در دو موقعیت s و $s+h$ را تغییرنگار[۲۵] می‌نامند و به صورت زیر نمادگذاری می‌شود:

$$۲\gamma(h) = \mathrm{Var}(Z(s+h) - Z(s)).$$

تغییرنگار معیاری برای نشان دادن میزان اثرات متقابل یک موقعیت با سایر موقعیت‌های فضای مورد مطالعه است. بنابراین کوچک بودن تغییرنگار، نشان‌گر وابستگی زیاد و بزرگ بودن آن نشانه وابستگی کم میدان تصادفی است. شیب نمودار تغییرنگار معرف همبستگی فضایی است، یعنی اگر برای تمام مقادیر h تغییرنگار ثابت باشد همبستگی فضایی داده‌ها بسیار ضعیف است.

تغییرنگار دارای سه پارامتر دامنه[۲۶]، ازاره[۲۷] و اثرقطعه‌ای[۲۸] است.

دامنه: بازه تغییرات تغییرنگار که خارج از آن به حالت افقی درآمده و ثابت می‌ماند، دامنه نامیده می‌شود. خارج از دامنه، مشاهدات تقریبا اثری بر هم ندارند، یعنی تقریبا ناهمبسته هستند. از این‌رو داده‌های خارج از دامنه با روش‌های کلاسیک آماری قابل تحلیل‌اند.

ازاره: به‌طور معمول تغییرنگار تابعی افزایشی از تاخیر $\|h\|$ است و ممکن است به کران بالایی منتهی شود. منظور از $\|h\|$، فاصله اقلیدسی بین دو مشاهده فضایی با فاصله h در یک جهت مشخص است. چنین کرانی ازاره تغییرنگار نامیده شده و عبارت است از

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

با توجه به طبیعت داده‌های فضایی که انتظار می‌رود در آن‌ها با افزایش فاصله مکانی، همبستگی داده‌ها کاهش یابد و در نتیجه تغییرنگار ثابت بماند، سعی می‌شود که بیش‌تر

[24] Gaussian random field
[25] Variogram
[26] Range
[27] Sill
[28] Nugget effect

از مدل‌های کران‌دار استفاده شود. ازاره به عنوان یکی از پارامترهای مهم تابع تغییرنگار، از روی مشاهدات برآورد می‌شود. تعیین کران‌دار بودن تغییرنگار از روی مشاهدات مشکل است و در عمل با رسم نمودار تغییرنگار تجربی مقدار c برآورد می‌شود. مقداری که تابع تغییرنگار تجربی بعد از رسیدن به دامنه، حول آن نوسان می‌کند، ازاره را نمایش می‌دهد.

اثر قطعه‌ای: مقدار تغییرنگار در مبدا مختصات $(h = ۰)$، اثر قطعه‌ای نامیده شده و به صورت

$$\lim_{\|h\|\to ۰}\gamma(h) = c_۰$$

تعریف می‌شود. به لحاظ نظری اثر قطعه‌ای باید صفر باشد، زیرا دو نمونه از موقعیت واحد، باید کمیت یکسانی داشته باشند. ولی در عمل تغییرنگارهای تجربی پیرو این وضعیت نیستند. دلایل این موضوع را می‌توان خطای نمونه‌گیری، خطای اندازه‌گیری یا تغییرات شدید کمیت مورد بررسی در موقعیت‌های نزدیک به هم دانست. استفاده از اثر قطعه‌ای می‌تواند معیاری برای تشخیص وجود ساختار فضایی در داده‌ها باشد. هر اندازه نسبت $\frac{c}{c+c_۰}$ کوچکتر از ۰٫۵ باشد، ساختار فضایی داده‌ها ضعیف‌تر است.

یکی از ویژگی‌های تغییرنگار، همیشه منفی شرطی[۲۹] بودن آن است. همچنین ماترون (۱۹۷۱) ثابت کرد که اگر $\|h\| \to \infty$ آن‌گاه $\frac{۲\gamma(h)}{\|h\|^۲} \to ۰$.

مدل‌های معتبر تغییرنگار

مدل‌های معتبر تغییرنگار توابعی هستند که در شرط همیشه منفی شرطی صدق می‌کنند. جورنل و هویج‌برگتس (۱۹۷۸) مدل‌های پارامتری معتبر مختلفی را معرفی کردند. مدل‌های تغییرنگار شامل دو گروه با اِزارِه و فاقد اِزارِه هستند. مدل‌های تغییرنگار همسان‌گرد[۳۰] دارای اِزارِه، به‌طور معمول بعد از دامنه تخت شده و با افزایش فاصله موقعیت‌ها تغییر معنی‌داری در آن‌ها صورت نمی‌پذیرد. همسان‌گردی را در ادامه تعریف خواهیم کرد.

انواع مدل‌های نیم‌تغییرنگار معتبر همسان‌گرد دارای اِزارِه عبارتند از:

۱. **مدل گاوسی:** تابع نیم‌تغییرنگار گاوسی به‌صورت

$$\gamma(h) = c_۰ + c(۱ - e^{\frac{-\|h\|^۲}{a^۲}}), \qquad h \in \mathbb{R}^d, \qquad d \geq ۱$$

است و با سه پارامتر a، c و $c_۰$ مشخص می‌شود، که در آن a دامنه، $c_۰$ اثر قطعه‌ای و $c + c_۰$ اِزارِه است.

[29] Conditional negative definite

[30] Isotropic

۲. **مدل نمایی:** تابع نیم‌تغییرنگار نمایی

$$\gamma(h) = c_\circ + c(1 - e^{\frac{-||h||}{a}}), \qquad h \in \mathbb{R}^d, \qquad d \geq 1$$

دارای سه پارامتر a، c و $c_\circ$ است. دامنه آن به‌طور معمول بسیار بزرگ است. وقتی داده‌ها در محدوده مورد بررسی دارای روند باشند یا نسبت به ابعاد محدوده تحت پوشش نمونه‌برداری دارای دامنه بزرگی باشند، تغییرنگار آن‌ها از مدل نمایی پیروی می‌کند.

۳. **مدل کروی:** ضابطه تابع نیم‌تغییرنگار کروی به‌صورت

$$\gamma(h) = \begin{cases} c_\circ + c(\frac{3}{2}\frac{||h||}{a} - \frac{1}{2}\frac{||h||^2}{a^2}) & \circ < ||h|| \leq a, h \in \mathbb{R}^d, \quad d = 1,2,3 \\ c_\circ + c & ||h|| > a \end{cases}$$

است.

به‌طور معمول مقدار مدل‌های نیم‌تغییرنگار بدون اِزارِه با افزایش فاصله هم‌چنان افزایش پیدا می‌کند. انواع این مدل‌ها عبارت‌اند از:

۴. **مدل توانی:** تابع نیم‌تغییرنگار توانی با سه پارامتر λ، a و $c_\circ$ به‌صورت

$$\gamma(h) = c_\circ + a||h||^\lambda \qquad \circ < \lambda < 2,\ h \in \mathbb{R}^d,\ d \geq 1$$

تعریف می‌شود. به پارامتر a شیب مدل گفته می‌شود. این مدل در حالت $\lambda = 1$ به مدل خطی منتهی می‌شود.

۵. **مدل لگاریتمی:** تابع نیم‌تغییرنگار لگاریتمی به‌صورت

$$\gamma(h) = 3a\ln(||h||)$$

است. نمودار این مدل دارای شیب $3a$ است که a پراکندگی مطلق نامیده می‌شود. چون برای $||h|| < 1$ مقدار $\gamma(h)$ منفی است، این مدل برای نمونه‌گیری‌های با ابعاد کوچک قابل استفاده نیست.

۶. **مدل موجی:** تابع نیم‌تغییرنگار موجی به‌صورت

$$\gamma(h) = c_\circ + c_w(1 - a_w \sin(-||h||/a_w)/||h||),\ h \in \mathbb{R}^d,\ d = 1,2,3$$

است و با سه پارامتر $a_w \geq \circ$، $c_w \geq \circ$ و $c_\circ \geq \circ$ مشخص می‌شود.

۷. **مدل سهمی‌گون:** تابع نیم‌تغییرنگار سهمی‌گون به‌صورت یک سهمی درجه دوم به‌صورت

$$\gamma(h) = \frac{1}{2}a^2||h||^2$$

است. اگر میدان تصادفی $Z(\cdot)$ دارای روند خطی باشد، یعنی

$$Z(s+h) = Z(s) + ah$$

آن‌گاه مدل تغییرنگار آن سهمی‌گون خواهد بود، زیرا

$$۲\gamma(h) = E[(Z(s+h) - Z(s))^۲] = a^۲h^۲.$$

تعریف ۶.۳.۱. کواریانس دو متغیر $Z(s)$ و $Z(s+h)$ را هم‌تغییرنگار[۳۱] می‌نامند و به‌صورت زیر نمادگذاری می‌شود:

$$C(h) = \mathrm{Cov}(Z(s), Z(s+h)).$$

هم‌تغییرنگار میزان شباهت تغییرپذیری دو متغیر $Z(s)$ و $Z(s+h)$ را نشان می‌دهد. بنابراین در صورت وجود همبستگی فضایی، نقاط نزدیک به هم تشابه تغییرپذیری بیشتری دارند و مقدار هم‌تغییرنگار بزرگتر است.

تعریف ۷.۳.۱. ضریب همبستگی بین $Z(s)$ و $Z(s+h)$ همبستگی‌نگار[۳۲] نامیده می‌شود و با شرط $C(۰) > ۰$ به‌صورت

$$\rho(s, s+h) = \rho(h) = \frac{C(h)}{C(۰)} = ۱ - \frac{\gamma(h)}{C(۰)}$$

تعریف می‌شود. همبستگی‌نگار هم مانند تغییرنگار تابعی از فاصله بین موقعیت‌ها است و با افزایش فاصله، مقدار آن کم و با کاهش فاصله مقدار همبستگی‌نگار افزایش می‌یابد.

تعریف ۸.۳.۱. میدان تصادفی $Z(\cdot)$ را همسان‌گرد گویند هرگاه تغییرنگار، هم‌تغییرنگار و همبستگی‌نگار آن همسان‌گرد باشند، یعنی تنها تابعی از $\|h\|$ باشند و به جهت h بستگی نداشته باشند.

تعریف ۹.۳.۱. میدان تصادفی $Z(\cdot)$ که ایستای مرتبه دوم و همسان‌گرد باشد را همگن گویند. در غیر این صورت میدان ناهمگن است.

۴.۳.۱ ناهمگنی فضایی

ناهمگنی فضایی یکی دیگر از اثرات فضایی و صرفا ناایستایی ساختاری است. این ناایستایی به دو شکل بی‌ثباتی در واریانس (واریانس نامتجانس[۳۳]) یا ضرایب مدل (ضرایب متغیر[۳۴]، رژیم‌های فضایی[۳۵])، بروز پیدا می‌کند. این دو جنبه از ناهمگنی کاملا مجزا از یکدیگر هستند

[31] Covariogram
[32] Correlogram
[33] Heteroscedasticity
[34] Variable cofficients
[35] Spatial regimes

(انسلین، ۱۹۸۸a). اولی، نتیجه حذف متغیرهای تبیینی لازم یا سایر شکل‌های مدل‌بندی نادرست، که منجر به بی‌ثباتی واریانس جمله خطا می‌شود، و دومی ناشی از ثابت نبودن ارتباط بین متغیر پاسخ و متغیرهای تبیینی در کل فضای مورد مطالعه است. همان‌طور که در بخش قبل گفته شد، می‌توان این ناهمگنی را با ابزار و تبدیلات مختلف استانداردسازی، سنجید. اما سه دلیل وجود دارند که تایید می‌کنند ناهمگنی را به صراحت در مدل در نظر بگیریم.

اول این‌که ساختار این ناایستایی از جنس فضایی است. به این معنی که مکان مشاهدات در تعیین شکل ناایستایی مهم است. به‌عنوان مثال، برای مدل‌بندی واریانس نامتجانس با طرح گروهی، می‌توان ناحیه مورد مطالعه را به زیرناحیه‌هایی با واریانس متفاوت تقسیم کرد. برای تفهیم بیشتر، یک مجموعه S از N واحد جغرافیایی (مثل استان‌ها، شهرها و سرشماری تراکم) را در نظر بگیرید. این مجموعه را به R زیرمجموعه S_r به هم متصل که هم‌پوشانی ندارند، تقسیم می‌کنیم به طوری که برای هر $r \neq s$، $S_r \bigcap S_s = \varnothing$ و $\bigcup_{r=۱,\dots,R} S_r = S$. با این روش از این پس واریانس نامتجانس گروهی، به صورت واریانس خطای خوشه‌بندی فضایی برای هر مشاهده دنبال می‌شود. یعنی برای هر $i \in S_r$، $\mathrm{Var}(\varepsilon_i) = \sigma^۲$. به‌طور مشابه، تنوع در ضرایب می‌تواند با رژیم‌های فضایی یا زیرمجموعه‌های جغرافیایی داده‌ها که در آن‌ها برای هر $i \in S_r$، $\beta_i = \beta_r$ مشخص شود. به بیان دیگر پارامترهای مدل در زیرمجموعه‌هایی از ناحیه تحت مطالعه مقادیر مختلف اختیار می‌کنند؛ یعنی رابطه بین متغیر تبیینی و متغیر پاسخ در کل فضای مطالعه پایا نیست.

دوم این که به دلیل ساختار فضایی، ناهمگنی فضایی و وابستگی فضایی به صورت مشترک رخ می‌دهند و جداسازی آن‌ها مشکل است. این مسأله در متون تخصصی به‌عنوان مشکل معکوس[۳۶] شناخته می‌شود. همچنین به نشدنی بودن تمایز بین آلودگی[۳۷] واقعی و ظاهری مربوط است (انسلین، ۲۰۱۰).

سوم این که در یک بخش‌بندی کلی، خودهمبستگی فضایی و ناهمگنی فضایی ممکن است معادل در نظر گرفته شوند. یعنی مثلا یک خوشه فضایی از باقی‌مانده‌های کرانگین[۳۸] ممکن است به علت وجود ناهمگنی فضایی (مثلا واریانس نامتجانس گروهی) یا خودهمبستگی فضایی تفسیر شود. لذا لازم است که هر دو جنبه ناهمگنی و وابستگی فضایی با هم در نظر گرفته شوند.

اما اکثرا در تحقیقات و مدل‌های فضایی نظیر مدل میدان تصادفی گاوسی، مدل‌های اتورگرسیو و مدل‌های اقتصادسنجی فضایی، تمرکز بر وابستگی فضایی است و اثر ناهمگنی فضایی به انزوا رفته است (انسلین، ۱۹۹۹). به‌عنوان مثال، نوسانات قیمت مسکن عموما از کلان شهرها شروع شده و در مراحل بعد به سایر شهرها سرریز می‌شود. لذا بحث مجاورت

[36] Inverse problem
[37] Contagion
[38] Extreme

و وابستگی فضایی مطرح می‌شود و در تحلیل‌های فضایی وارد می‌شود؛ اما این موضوع نیز وجود دارد که متغیرهای اثرگذار برقیمت مسکن در شهرهای مختلف دارای تاثیرات متفاوت هستند و از مکانی به مکان دیگر تغییر می‌کنند. هنگامی که این تنوع بیش از حد رخ دهد، نشان از ناهمگنی فضایی است و باید در مدل لحاظ شود. در فصل دوم، مدل رگرسیون موزون جغرافیایی[۳۹] (GWR) را که در آن به این امر توجه شده است و ناهمگنی فضایی نیز در کنار وابستگی فضایی در مدل لحاظ می‌شود، معرفی می‌کنیم.

۴.۱ سایر مفاهیم و مقدمات

در این بخش یک روش مدل‌بندی ساختار وابستگی فضایی و ماتریس وزن فضایی به‌عنوان ابزاری برای این مدل‌بندی مطرح می‌شود. همچنین آزمون‌های کاربردی معمول، برای تشخیص وجود اثرات فضایی در داده‌ها معرفی می‌شوند.

۱.۴.۱ مدل خودهمبستگی

یکی از شاخه‌های جالب و در حال رشد آمار فضایی مربوط به خودهمبستگی فضایی است. خودهمبستگی فضایی مفهومی نسبتا ساده و در حقیقت بسط همین مفهوم در آمار متعارف است. خودهمبستگی، خواه از نوع فضایی باشد یا نه، یک معیار برای بیان تشابه و همبستگی بین مشاهدات همسایه است. مدل

$$\mathbf{u} = \rho \mathbf{W}\mathbf{u} + \varepsilon \qquad (۶.۱)$$

برای بیان خودهمبستگی به‌کار می‌رود که در آن همسایگی‌ها از طریق ماتریس وزن $\mathbf{W}$ معرفی می‌شوند. در این مدل از میانگین وزنی همسایگی‌های مشاهده i، یعنی

$$u_i = w_{i۱}u_۱ + w_{i۲}u_۲ + \cdots + w_{iN}u_N$$

برای برآورد مقدار آن استفاده می‌شود. پارامتر ρ ضریب خودهمبستگی ترکیب خطی مشاهده u_i با همسایگی‌های خود است و متوسط تاثیر مشاهدات همسایه بر مقدار مشاهده i را اندازه‌گیری می‌کند. این موضوع همان‌طور که یک چالش محسوب می‌شود و موجب پیچیدگی آزمون‌های آماری می‌شود، اما با اطلاعات فضایی نیز مطابقت دارد.

از این رو می‌توان مدل خودهمبستگی را با رگرسیون خطی ترکیبی و ساختار فضایی را مدل‌بندی کرد. این مدل‌ها با عنوان مدل‌های اتورگرسیو فضایی[۴۰] شناخته می‌شوند. در این مدل‌بندی جمله خودهمبستگی، که در آن ماتریس $\mathbf{W}$ وزن همسایگی‌های مکانی را نشان

[39] Geographical weighted regression

[40] Spatial autoregressive model

می‌دهد، با توجه به شرایط داده‌ها در متغیر پاسخ، متغیرهای تبیینی، جمله خطا یا ترکیبی از این‌ها قرار می‌گیرد و مدل‌های رگرسیون فضایی مختلفی را تولید می‌کند که در فصل دوم تحت عنوان مدل‌های اتورگرسیو شرطی، اتورگرسیو هم‌زمان و مدل‌های اقتصادسنجی به تشریح آن‌ها می‌پردازیم.

۲.۴.۱ ماتریس وزن فضایی

ماتریس وزن فضایی، ابزاری است که شدت رابطه بین مشاهدات در یک همسایگی را به صورت کمی نشان می‌دهد. این ماتریس طوری تعریف می‌شود که موافق با ساختار همبستگی فضایی است؛ یعنی به همسایه‌های نزدیک وزن بیشتر و به همسایه‌های دورتر وزن کمتری اختصاص داده می‌شود. ماتریس وزن فضایی یک ماتریس مربعی از مرتبه n (تعداد مشاهدات) است که مولفه‌های آن، یعنی w_{ij}، دارای خواص زیر هستند:

۱. نامنفی هستند.

۲. به ازای $i = ۱, \cdots, n$، $w_{ii} = ۰$ و $w_{ij} = w_{ji}$.

به منظور تعیین w_{ij} دو منبع اطلاعاتی موجود است. یکی موقعیت مشاهدات در صفحه مختصات که از طریق طول و عرض جغرافیایی، فاصله هر نقطه در فضا را نسبت به نقاط دیگر نشان می‌دهد. دومی مجاورت و همسایگی است که اغلب در مورد داده‌های مشبکه ای مورد استفاده قرار می‌گیرد و بر این اساس که مناطق i و j حاشیه یا مرز مشترک دارند یا خیر وزن w_{ij} را تعیین می‌کند. در ادامه تعدادی از روش‌های وزن‌دهی بر پایه این دو منبع اطلاع را بیان می‌کنیم.

ماتریس فاصله[۴۱]: در این ماتریس میزان تاثیر و ارتباط هر جفت مشاهده تابعی از فاصله آن‌ها است. یعنی

$$w_{ij} = \begin{cases} d_{ij} & i \neq j \\ ۰ & i = j \end{cases} \qquad (۷.۱)$$

ماتریس معکوس‌ـفاصله[۴۲]: در این حالت شدت رابطه بین دو مشاهده با فاصله آن‌ها از هم نسبت عکس دارد. یعنی

$$w_{ij} = \begin{cases} \frac{۱}{d_{ij}} & i \neq j \\ ۰ & i = j \end{cases} \qquad (۸.۱)$$

[41] Distance matrix

[42] Inverse-distance matrix

ماتریس مجاورت ضلعی[۴۳]: این ماتریس وزن در خصوص داده‌های مشبکه‌ای منظم مورد استفاده قرار می‌گیرد و هرگاه نواحی i و j دارای ضلع مشترک باشند، $w_{ij} = ۱$ و در غیر این صورت $w_{ij} = ۰$.

ماتریس مجاورت رأسی[۴۴]: در این روش وزن‌دهی، دو ناحیه همسایه یکدیگر محسوب می‌شوند و w_{ij} مقدار ۱ می‌گیرد، اگر دارای یک رأس مشترک باشند. این روش نیز در خصوص داده های مشبکه‌ای منظم به کار می‌رود.

ماتریس مجاورت شعاعی[۴۵]: در این روش که برای داده‌های مشبکه‌ای منظم به کار می‌رود، دایره‌ای به مرکز ناحیه i و شعاع طول ضلع نواحی رسم می‌شود. به نواحی‌ای که دایره از آن‌ها عبور می‌کند یعنی نواحی‌ای که دارای ضلع و/یا راس مشترک با ناحیه i هستند، وزن ۱ و در غیر این صورت وزن صفر تخصیص داده می‌شود.

در خصوص داده‌های مشبکه‌ای نامنظم، نواحی‌ای که دارای مرز مشترک هستند معمولا همسایه محسوب می‌شوند.

۳.۴.۱ آماره آزمون موران نوع I

متداول‌ترین آزمون حضور خودهمبستگی فضایی توسط یک آماره که مشابه آن اولین بار توسط موران (۱۹۴۸) برای بررسی همبستگی سری‌های زمانی معرفی شد، انجام می‌شود. این آماره که به تحلیل الگوهای پراکنش و توزیع خطاها در فضا می‌پردازد، به آماره موران I[۴۶] معروف است. صورت ماتریسی آماره موران به صورت:

$$I = (n/S_۰)(\mathbf{e}^T\mathbf{W}\mathbf{e}/\mathbf{e}^T\mathbf{e})$$

است، که در آن $\mathbf{e}$ بردار مقادیر باقی‌مانده حاصل از رگرسیون OLS است و $S_۰ = \sum_i \sum_j w_{ij}$.

مقدار امید و واریانس این آماره تحت فرضیه نبود خودهمبستگی فضایی (فرضیه صفر) به صورت زیر است:

$$\mathrm{E}(I) = \frac{tr(\mathbf{MW})}{n-p} \qquad (۹.۱)$$

$$\mathrm{Var}(I) = \frac{tr(\mathbf{MWMW}^T) + tr(\mathbf{MWMW}) + [tr(\mathbf{MW})]^۲}{(n-p)(n-p+۲)} - [\mathrm{E}(I)]^۲ \qquad (۱۰.۱)$$

که در آن ماتریس $\mathbf{M}$ برابر است با

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-۱}\mathbf{X}^T.$$

[43] Edge contiguity matrix
[44] Vertex contiguity matrix
[45] Radius contiguity matrix
[46] Moran I statistic

این آماره نشان می‌دهد که پراکندگی مقادیر خطا با در نظر گرفتن متغیرهای مورد مطالعه از یک الگوی خوشه‌ای پیروی می‌کند یا نه. اگر باقی‌مانده‌ها یا مقادیر متغیرهای مربوط به آن‌ها به‌طور تصادفی در ناحیه تحت مطالعه توزیع شده باشند، نباید بین آن‌ها ارتباطی وجود داشته باشد. به‌طور کلی مقادیر مثبت یا منفی آماره موران نشان از خودهمبستگی مثبت یا منفی و همچنین مقادیر نزدیک به مقدار مورد انتظار (۹.۱)، بیانگر نبود خودهمبستگی یعنی الگوی فضایی تصادفی است. محدوده تغییرات این آماره برخلاف ضریب همبستگی، بازه $[-۱,۱]$ نیست.

همچنین آماره

$$Z = \frac{I - \mathrm{E}(I)}{\sqrt{\mathrm{Var}(I)}} \qquad (۱۱.۱)$$

برای نمونه‌های به اندازه کافی بزرگ دارای توزیع نرمال استاندارد است. اگر $|Z| < z_{\alpha/۲}$ آن‌گاه آزمون خودهمبستگی فضایی در سطح α معنی‌دار نبوده و بین جمله‌های خطا در مدل رگرسیونی، خودهمبستگی فضایی وجود ندارد.

۴.۴.۱ آزمون‌های مبتنی بر درستنمایی

با توجه به استفاده گسترده از روش درستنمایی ماکسیمم در برآورد مدل‌های فضایی، اکثر آزمون‌های فرضیه برای پارامترهای این مدل‌ها مبتنی بر ملاحظات مجانبی است. دو نمونه از این آزمون‌های مجانبی بر پایه درستنمایی ماکسیمم، آزمون نسبت درستنمایی[۴۷] (LR) و آزمون ضریب لاگرانژ[۴۸] (LM) است.

آزمون نسبت درستنمایی (LR)

آزمون نسبت درستنمایی، فرضیه‌های

$$\begin{cases} H_۰ : \lambda = ۰ \\ H_۱ : \lambda \neq ۰ \end{cases}$$

که در آن λ ضریب خودهمبستگی جمله خطا است را مورد آزمون قرار می‌دهد. آماره این آزمون بر پایه تفاضل تابع درستنمایی مدل رگرسیون فضایی با حضور λ و تابع درستنمایی آن تحت فرضیه صفر بنا شده است و به‌صورت زیر معرفی می‌شود:

$$\mathrm{LR} = n\Big(\ln \sigma_۰^۲ - \ln \sigma^۲\Big) + ۲ \ln |\mathbf{I} - \lambda \mathbf{W}|. \qquad (۱۲.۱)$$

این آماره دارای توزیع کای‌ـدو با یک درجه آزادی است که در آن n تعداد مشاهدات و $\sigma^۲$ و $\sigma_۰^۲$ به‌ترتیب برآورد درستنمایی ماکسیمم واریانس خطا در مدل فضایی و برآورد آن تحت فرضیه

[47] Likelihood ratio
[48] Lagrange multiplier

صفر است. بزرگی آماره LR نشان از اختلاف بین این دو مدل است و موجب رد فرضیه صفر می‌شود؛ یعنی خودهمبستگی فضایی در جمله‌های خطا موجود است.

آزمون ضریب لاگرانژ (LM)

آزمون ضریب لاگرانژ به دو شکل مختلف برای تشخیص حضور خودهمبستگی در جمله خطا و متغیر وابسته وجود دارد، که به ترتیب آزمون ضریب لاگرانژ خطای فضایی[۴۹] (LM_λ) و آزمون ضریب لاگرانژ تاخیر فضایی[۵۰] (LM_ρ) نامیده می‌شوند.

الف) آزمون ضریب لاگرانژ خطای فضایی

آماره آزمون ضریب لاگرانژ خطا به صورت زیر تعریف می‌شود:

$$LM_\lambda = \frac{۱}{T}\left(\frac{\mathbf{e}^T\mathbf{W}\mathbf{e}}{S^۲}\right)^۲. \qquad (۱۳.۱)$$

این آماره در مدل رگرسیون

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\upsilon} \qquad \boldsymbol{\upsilon} = \lambda\mathbf{W}\boldsymbol{\upsilon} + \boldsymbol{\varepsilon} \qquad (۱۴.۱)$$

که در آن $\boldsymbol{\varepsilon} \sim MVN(۰, \sigma^۲\mathbf{I})$ است، فرضیه‌های زیر را مورد آزمون قرار می‌دهد:

$$\begin{cases} H_۰ : \lambda = ۰ \\ H_۱ : \lambda \neq ۰ \end{cases}$$

یعنی آزمون می‌کند که آیا جمله‌های خطا در بردار $\boldsymbol{\upsilon}$ ناهمبسته هستند یا خیر. این مدل رگرسیونی (مدل خطای فضایی[۵۱] (SEM)) در فصل بعد به تفصیل بیان می‌شود.

در این آماره $S^۲$ برآورد درستنمایی ماکسیمم پارامتر $\sigma^۲$ است که به‌صورت زیر تعریف می‌شود:

$$S^۲ = \frac{\mathbf{e}^T\mathbf{e}}{n}$$

و همچنین

$$T = \mathrm{tr}(\mathbf{W}^T\mathbf{W} + \mathbf{W}^۲).$$

آماره آزمون (۱۳.۱) جدا از جمله مقیاس‌گذاری $\mathrm{tr}(\mathbf{W}^T\mathbf{W} + \mathbf{W}^۲)$ برابر توان دوم آماره موران است و تحت درستی فرضیه $H_۰$، دارای توزیع تقریبی کای‌ـدو با ۱ درجه آزادی است.

تحت فرضیه $H_۱$، لگاریتم تابع درستنمایی بردار $\mathbf{y}$ با توجه به (۱۴.۱) به‌صورت زیر است:

$$\ln L = -\frac{n}{۲}\ln(۲\pi\sigma^۲) + \ln|\mathbf{I}_n - \lambda\mathbf{W}| - \frac{۱}{۲\sigma^۲}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{I}_n - \lambda\mathbf{W})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

[49] Lagrange multiplier error test
[50] Lagrange multiplier lag test
[51] Spatial error model

برای به‌دست آوردن برآوردهای درستنمایی ماکسیمم پارامترها تحت فرضیه H_0، تابع لاگرانژ را به صورت زیر تعریف می‌کنیم:

$$L_R = \ln L - t\lambda.$$

اکنون مشتق تابع لاگرانژ را نسبت به تک تک پارامترها به‌دست می‌آوریم:

$$\begin{aligned}
\frac{\partial L_R}{\partial \boldsymbol{\beta}^T} &= \frac{1}{\sigma^2}\mathbf{X}^T(\mathbf{I}_n - \lambda\mathbf{W})^T(\mathbf{I}_n - \lambda\mathbf{W})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \\
\frac{\partial L_R}{\partial \sigma^2} &= \frac{1}{2\sigma^2}\{n + \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{I}_n - \lambda\mathbf{W})^T(\mathbf{I}_n - \lambda\mathbf{W})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\} = 0 \\
\frac{\partial L_R}{\partial \lambda} &= -t - \sum_{j=1}^{n}\delta_j(1 - \lambda\delta_j)^{-1} + \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\{(I_n - \lambda\mathbf{W})^T\mathbf{W} \\
&\quad + \mathbf{W}^T(\mathbf{I}_n - \lambda\mathbf{W})\}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \\
\frac{\partial L_R}{\partial t} &= -\lambda = 0 \qquad (۱۵.۱)
\end{aligned}$$

به‌طوری که δ_jها مقادیر ویژه ماتریس $\mathbf{W}$ هستند.

با حل معادلات (۱۵.۱)، برآورد ماکسیمم درستنمایی پارامتر t به‌صورت زیر به‌دست می‌آید:

$$\hat{t} = -\mathrm{tr}(\mathbf{W}) + \frac{1}{2\hat{\sigma}^2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{W} + \mathbf{W}^T)(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

به‌طوری که $\hat{\boldsymbol{\beta}}$ برآورد OLS پارامتر $\boldsymbol{\beta}$ است. هم‌چنین

$$S^2 = \hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n} = \frac{\mathbf{e}^T\mathbf{e}}{n}.$$

ار آن‌جا که $\mathrm{tr}(\mathbf{W}) = 0$، پس

$$\hat{t} = \frac{n\mathbf{e}^T(\mathbf{W} + \mathbf{W}^T)\mathbf{e}}{2\mathbf{e}^T\mathbf{e}}.$$

تحت فرضیه H_0، تابع اطلاع پارامتر λ به‌صورت زیر به‌دست می‌آید:

$$T = -\mathrm{E}_{H_0}\Big(\frac{\partial^2 \ln L}{\partial \lambda^2}\Big) = \mathrm{tr}(\mathbf{W}^2 + \mathbf{W}^T\mathbf{W}).$$

بنابراین آماره آزمون، بر اساس رفتار $\hat{t}\big[\mathrm{tr}(\mathbf{W}^2 + \mathbf{W}^T\mathbf{W})\big]^{-\frac{1}{2}}$ ساخته می‌شود.

برای اطلاعات بیش‌تر درباره اثبات بالا به بوریج (۱۹۸۰) مراجعه کنید.

ب) آزمون ضریب لاگرانژ تأخیر فضایی

این‌بار فرضیه مورد بررسی، وجود خودهمبستگی فضایی در متغیر وابسته مدل است. در صورت رد فرضیه صفر، مدل زیر، که به مدل تاخیر فضایی[۵۲] (SLM) معروف است و در فصل بعد به تفصیل معرفی خواهد شد، به داده‌ها برازش داده می شود:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \varepsilon. \qquad (۱۶.۱)$$

[52] Spatial lag model

آزمون فرضیه مورد نظر به‌صورت زیر تعریف می‌شود:

$$\begin{cases} H_۰ : \rho = ۰ \\ H_۱ : \rho \neq ۰ \end{cases} \qquad (۱۷.۱)$$

آماره آزمون ضریب لاگرانژ برای فرضیه‌های (۱۷.۱)، به‌صورت زیر است:

$$LM_\rho = \frac{۱}{J}\Big(\frac{\mathbf{e}^T\mathbf{W}\mathbf{y}}{S^۲}\Big)^۲ \qquad (۱۸.۱)$$

به‌طوری که

$$J = \frac{[(\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{M}(\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}}) + TS^۲]}{S^۲}.$$

برای محاسبه آماره آزمون (۱۸.۱) ابتدا مدل (۱۶.۱) که در آن $\varepsilon \sim MVN(۰, \sigma^۲\mathbf{I})$ است تحت فرضیه صفر به کمک روش OLS برآورد می‌شود، به‌طوری که $\hat{\beta}$ برآورد حداقل مربعات معمولی بردار پارامتر β است. آماره آزمون (۱۸.۱) نیز، تحت درستی فرضیه $H_۰$، دارای توزیع تقریبی کای‌ـدو با ۱ درجه آزادی است. اگر آزمون (۱۷.۱) معنی‌دار باشد، به این معناست که باید جمله تأخیری $\rho\mathbf{W}\mathbf{y}$ نیز در مدل رگرسیونی لحاظ شود. در غیر این‌صورت، یعنی اگر معنی‌دار نباشد، نیازی به درنظر گرفتن این جمله در مدل نیست. چگونگی به‌دست آمدن آماره آزمون (۱۸.۱) در بوریج (۱۹۸۰) آمده است.

۵.۴.۱ آزمون ناهمگنی در حضور خودهمبستگی

زمانی که جمله‌های خطا دارای وابستگی باشند، توابع پارامتری معمول برای آزمون ناهمگنی معتبر نیستند. چرا که توزیع مجانبی آماره این آزمون‌ها بر اساس پذیره استقلال توزیع خطاها است و در غیاب استقلال این نتایج حفظ نخواهند شد. بنابراین با اثبات حضور خودهمبستگی فضایی، نیازمند روشی هستیم که ناهمگنی را با وجود خودهمبستگی آزمون کند.

انسلین (۱۹۸۸a) برای آزمون وجود ناهمگنی در حضور خودهمبستگی فضایی آماره‌ای را معرفی کرده است. این آماره با ایجاد تغییراتی در ضابطه اصلی آماره بریچ‌ـپیگان[۵۳] (بریچ و پیگان، ۱۹۷۹) که برای آزمون ناهمگنی در شرایط استقلال به کار می‌رود، به شکل زیر معرفی شده است:

$$LM = (۱/۴\sigma^۴)\mathbf{f}^T\mathbf{Z}\mathbf{I}^{-۱}\mathbf{Z}^T\mathbf{f} \qquad (۱۹.۱)$$

که در آن $\mathbf{Z}$ یک ماتریس $n \times (p+۱)$ از بردارهای $\mathbf{z}$ برای هر مشاهده است، $f_i = (\sigma^{-۱}e_i)^۲ - ۱$ و e_iها درایه‌های بردار $\mathbf{e} = (۱-\lambda)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$، یعنی باقی‌مانده‌های درستنمایی ماکسیمم رگرسیون فضایی با خطاهای خودهمبسته هستند. $\sigma^۲$ نیز برآورد ML واریانس خطای رگرسیون فضایی است. برای اطلاع بیشتر در مورد این آماره به انسلین (۱۹۸۸a) مراجعه کنید.

[53] Breusch-Pagan

۶.۴.۱ ضرب هادامارد

تعاریف متعددی برای ضرب ماتریس‌ها وجود دارند. یکی از این تعاریف، ضرب هادامارد[۵۴] یا ضرب درایه‌ای است (میلین، ۲۰۰۷). در این روش دو ماتریس باید دارای ابعاد یکسان باشند و ماتریس حاصل ماتریسی با همان ابعاد است. ضرب هادامارد دو ماتریس $\mathbf{A}$ و $\mathbf{B}$ به‌صورت زیر تعریف می‌شود:

$$[\mathbf{A} \odot \mathbf{B}]_{ij} = [\mathbf{A}]_{ij}[\mathbf{B}]_{ij} \quad ۱ \leq i \leq m, ۱ \leq j \leq n.$$

روش هادامارد در الگوریتم‌های فشرده‌سازی (الگوریتم‌های رایج برای فشرده‌سازی داده‌های چندرسانه‌ای مثل الگوریتم JPEG) استفاده می‌شود.

[54] Hadamard

فصل ۲

مدل‌بندی ساختار وابستگی فضایی

۱.۲ مقدمه

به‌طور کلی در تحلیل مدل‌های رگرسیونی خطاهای مدل ناهمبسته در نظر گرفته می‌شوند. اما این پذیره در بسیاری از موقعیت‌های کاربردی برقرار نیست. به‌ویژه در تحلیل داده‌های فضایی که مشاهدات برحسب موقعیت به یکدیگر وابسته‌اند و مشاهدات نزدیک به هم همبستگی بیشتری دارند، این پذیره اشتباه است. بنابراین لازم است که این امر در مدل‌بندی داده‌های فضایی مورد توجه قرار گیرد.

یک راه مرسوم برای لحاظ کردن ساختار وابستگی فضایی، استفاده از مدل‌های اتورگرسیو فضایی است. در طول سال‌های اخیر، مدل‌های مختلفی بر این مبنا معرفی شده‌اند و پیوسته به‌روز می‌شوند؛ به‌عنوان مثال می‌توان به دو مدل متداول و بسیار محبوب داده‌های مشبکه‌ای یعنی اتورگرسیو شرطی[۱] (CAR) و اتورگرسیو هم‌زمان[۲] (SAR) اشاره کرد. هر دو مدل وابستگی فضایی را در ساختار ماتریس کواریانس، به عنوان یک تابع از ماتریس وزن $\mathbf{W}$ و اغلب یک پارامتر همبستگی فضایی ناشناخته ثابت، تولید می‌کنند. دسته‌ای دیگر از مدل‌های اتورگرسیو نیز وجود دارند که به دلیل کاربرد بسیار آن‌ها در مسائل اقتصادی (نظیر مطالعات بازرگانی، جمعیت‌شناسی و مالی) به مدل‌های اقتصادسنجی معروف شده‌اند. در این فصل به توضیح هر

[1] Conditional autoregressive model

[2] Simultaneously autoregressive model

یک از این مدل‌ها می‌پردازیم و به دلیل ناتوانی آن‌ها در مدل‌بندی و تفسیر ناهمگنی فضایی، مدل رگرسیون موزون جغرافیایی را معرفی می‌کنیم.

۲.۲ وارد کردن ساختار وابستگی از طریق ماتریس کواریانس

همان‌گونه که گفته شد، یک راه لحاظ کردن وابستگی فضایی در مدل رگرسیونی، وارد کردن آن از طریق ساختار وابستگی مقادیر خطاست. در مدل رگرسیونی پایه (۲.۱) جمله خطا یک متغیر تصادفی مستقل نرمال با میانگین صفر و واریانس ثابت معرفی شد، یعنی

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}). \qquad (۱.۲)$$

اما در مورد داده‌های دارای ساختار فضایی، این پذیره ناهمبستگی و استقلال قابل قبول نیست زیرا مشاهدات حداقل از طریق مکان قرارگیری به یکدیگر وابسته هستند. در این موارد بهتر است که مدل (۱.۲) به‌صورت $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ نوشته شود، که در آن $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ و $\boldsymbol{\Lambda}$ یک ماتریس واریانس متقارن مثبت است. در این ماتریس به دلیل ساختار فضایی موقعیت‌هایی که به یکدیگر نزدیکترند، دارای کواریانس بزرگ‌ترند.

بنابراین مدل رگرسیون فضایی به‌صورت عمومی زیر نوشته می‌شود:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(\circ, \boldsymbol{\Lambda}) \qquad (۲.۲)$$

به دلیل ساختار پیچیده معمول ماتریس $\boldsymbol{\Lambda}$، برای ساده‌سازی مدل‌بندی، پذیره‌هایی به مدل اضافه می‌شوند. در ادامه به معرفی این روش‌هاکه مورد توجه این پایان‌نامه است، می‌پردازیم.

ساختار وابستگی داده‌های مشبکه‌ای

فرض کنید $\{Y(A_i) : A_i \in (A_1, \cdots, A_n)\}$ یک میدان تصادفی نرمال است که $\{A_1, \cdots, A_n\}$ نواحی مشبکه‌ای هستند یعنی $A_1 \cup A_2 \cup \cdots \cup A_n = \mathrm{D}$ و $A_i \cap A_j = \varnothing \ \ \forall i \neq j$. در این بخش دو مدل متداول اتورگرسیو شرطی و اتورگرسیو هم‌زمان برای مدل‌بندی این ساختار معرفی می‌شوند.

مدل اتورگرسیو شرطی

یک راه مدل‌بندی $\{Y(A_i) : A_i \in (A_1, \cdots, A_n)\}$ با مشاهدات y_i، مدل اتورگرسیو شرطی است که به‌صورت

$$y_i \mid \{y_j, j \neq i\} \sim N(\mu_i + \sum_{j=1}^{n} c_{ij}(y_j - \mu_j), \tau^2) \qquad (۳.۲)$$

معرفی می‌شود. از این پس $y(A_i)$ به اختصار y_i نوشته می‌شود. در مدل‌های CAR، علاوه بر وابستگی متغیر پاسخ $\mathbf{y}$ به متغیرهای تبیینی $\mathbf{X}$، y_i ها نیز وابسته به یکدیگر در نظر گرفته می‌شوند. توزیع y_i به شرط سایر مقادیر بردار y نیز نرمال است و به صورت $y_j - \mu_j$، یعنی اختلاف بین مقدار واقعی y_j و مقدار مورد انتظار آن، بیان شده است. برای c_{ij} ها که مقادیر ثابت معلوم یا نامعلوم هستند، محدودیت‌هایی قرار دارند. از جمله $c_{ii} = 0$ و $c_{ij} = c_{ji}$. محدودیت اول به سادگی به این معناست که، توزیع شرطی y_i نمی‌تواند به خودش وابسته باشد و محدودیت دوم به لزوم متقارن بودن وزن‌ها تاکید می‌کند.

مدل (۳.۲) را می‌توان به شکل ماتریسی

$$\mathbf{y} \sim N(\boldsymbol{\mu}, (\mathbf{I} - \mathbf{C})\tau^2) \qquad (۴.۲)$$

نوشت، که در آن ماتریس مقادیر $\mathbf{C}$ برای بازتاب ساختار فضایی داده‌ها به صورت‌های مختلفی تعریف می‌شود. یک راه معمول ساخت $\mathbf{C}$ معرفی آن توسط یک پارامتر وابستگی فضایی (ρ_c) و ماتریس وزن $\mathbf{W}$ است، که ساختار همسایگی‌های فضایی را وارد مدل می‌کند. تعاریف مختلفی برای ماتریس وزن $\mathbf{W}$ ارائه شده‌اند؛ تعدادی از آن‌ها را در فصل یک معرفی کردیم. برای مطالعه بیشتر در خصوص نحوه انتساب وزن‌ها در این مدل، به کلایتون و برناردینی (۱۹۹۲) مراجعه کنید.

از مدل CAR به‌طور گسترده برای مدل‌بندی داده‌های مشبکه‌ای استفاده شده است. به‌عنوان مثال، در متون پزشکی برای تصویرسازی جغرافیایی بیماری‌ها توسط کلایتون و کالدر (۱۹۸۷) و مولی و ریچاردسان (۱۹۹۱) مورد استفاده قرار گرفته است. همچنین در متون اقتصادی می‌توان از انسلین و فلوراکس (۱۹۹۵) و کلیجان و پروچا (۱۹۹۹) نام برد. همچنین فودرینگ‌هام و همکاران (۲۰۰۳) با توجه به این که علاوه بر ویژگی‌های یک مسکن (مانند زیربنا و سن ساختمان)، ارزش بناهای مجاور نیز در تعیین قیمت مسکن موثر است، از مدل CAR برای تحلیل قیمت مسکن استفاده کردند.

مدل اتورگرسیو هم‌زمان

یک مدل اتورگرسیو هم‌زمان به صورت زیر تعریف می‌شود:

$$y_i \sim N(\mu_i + \sum_{j=1}^{n} b_{ij}(y_i - \mu_j), \lambda^2) \qquad (۵.۲)$$

در این مدل بر خلاف مدل CAR توزیع y_iها شرطی نیست؛ بلکه برای هر y_i به‌عنوان یک مجموعه از معادلات هم‌زمان تعریف شده است. به این معنا که مقدار واقعی y_i با مقدار مورد انتظار y_j همبستگی دارد. محدودیت $b_{ii} = 0$ برای مدل SAR نیز برقرار است اما شرط تقارن وجود ندارد. معادله (۵.۲) را می‌توان به‌صورت زیر بازنویسی کرد:

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \lambda^2(\mathbf{I} - \mathbf{B})^{-1}(\mathbf{I} - \mathbf{B}^T)^{-1}). \qquad (۶.۲)$$

در این مدل نیز همانند مدل CAR ماتریس $\mathbf{B}$ اغلب به صورت $\mathbf{B} = \rho_s \mathbf{W}$ بیان می‌شود و روش انتخاب وزن‌ها نیز همانند CAR است.

۳.۲ مدل‌های اقتصادسنجی

یکی از تحولات و پیشرفت‌های ایجادشده در به‌کارگیری روش‌های کمی در علوم رفتاری به ویژه اقتصاد، تکامل شاخه اقتصادسنجی به اقتصادسنجی فضایی است. تفاوت اقتصادسنجی فضایی با اقتصادسنجی مرسوم، در توانایی و کاربرد آن در مواجهه با داده‌هایی است که دارای اطلاعات مکانی هستند. اقتصادسنجی فضایی، اثرات فضایی را به مدل‌های رگرسیونی اضافه می‌کند. با توجه به این‌که اقتصادسنجی فضایی به تحلیل داده‌های فضایی می‌پردازد، نخستین کاربردهای آن در علم اقتصاد در علوم منطقه‌ای، اقتصاد شهری و املاك و مستغلات (از جمله بازار مسکن) و جغرافیای اقتصادی بروز می‌یابد، اما به تدریج در شاخه‌های دیگر علم اقتصاد نظیر مطالعات تقاضا، اقتصاد بین‌الملل، اقتصاد نیروی کار، اقتصاد بخش عمومی و اقتصاد کشاورزی و محیط زیستی نیز از روش اقتصادسنجی فضایی استفاده شده است. این شاخه از اقتصادسنجی توسط پالینک و کلاسن (۱۹۷۹) و انسلین (۱۹۸۰) گسترش پیدا کرده است.

کارهای متعددی توسط محققین مختلف درباره روش‌های مختلف برآورد و آزمون‌های تخصیص مدل‌های فضایی صورت گرفته‌اند. از جمله می‌توان به انسلین (۱۹۸۸a)، هاینینگ (۱۹۹۰)، کرسی (۱۹۹۳) و لی‌سیج و پیس (۲۰۰۹) اشاره کرد. همان‌طور که گفته شد، مدل‌های بسیاری در ادبیات اقتصادسنجی فضایی مطرح شده‌اند. اما علی‌رغم این گستردگی تنها تعدادی از این مدل‌ها در مطالعات تجربی مورد استفاده قرار می‌گیرند. دلیل این امر وجود چالش‌های مدل‌بندی و محاسباتی در تعمیم و برازش اقتصادسنجی سایر مدل‌های فضایی است. در ادامه به معرفی این مدل‌ها و نحوه برآورد پارامترهای آن‌ها می‌پردازیم.

۱.۳.۲ مدل خطای فضایی

برای معرفی این مدل از مدل عمومی رگرسیون فضایی (۲.۲) شروع می‌کنیم. برای مشاهده iام فرض می‌کنیم

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + v_i, \qquad i = 1, \cdots, n$$

که در آن جمله خطا برای ناحیه iام، یعنی v_i، خودهمبسته فضایی و دارای توزیع نرمال است. به‌طور دقیق‌تر

$$v_i = \lambda \sum_j w_{ij} v_j + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2), \qquad i = 1, \cdots, n.$$

همان‌طور که گفته شد، λ پارامتر وابستگی فضایی و w_{ij}ها وزن‌های فضایی با شرط $w_{ii} = 0$ هستند. در حقیقت در این مدل (همان‌طور که از اسم آن مشخص است) فرض شده است

که کل ساختار وابستگی‌های فضایی در جمله خطای مدل وجود دارد. مدل خطای فضایی را می‌توان به‌صورت ماتریسی زیر نوشت:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\upsilon}, \quad \boldsymbol{\upsilon} = \lambda \mathbf{W}\boldsymbol{\upsilon} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\circ, \sigma^{\mathsf{۲}}\mathbf{I}_n). \tag{۷.۲}$$

مدل (۷.۲) برای برآورد ضرایب، باید یکپارچه شده و به‌صورت یک مدل رگرسیون خطی تعمیم‌یافته نوشته شود. می‌دانیم

$$\begin{aligned} \boldsymbol{\upsilon} &= \lambda \mathbf{W}\boldsymbol{\upsilon} + \boldsymbol{\varepsilon} \\ &= (\mathbf{I}_n - \lambda \mathbf{W})^{-\mathsf{۱}}\boldsymbol{\varepsilon} = \mathbf{B}_\lambda^{-\mathsf{۱}}\boldsymbol{\varepsilon}. \end{aligned} \tag{۸.۲}$$

همچنین

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{\upsilon}) &= \mathrm{Cov}(\mathbf{B}_\lambda^{-\mathsf{۱}}\boldsymbol{\varepsilon}) = \mathbf{B}_\lambda^{-\mathsf{۱}}\mathrm{Cov}(\boldsymbol{\varepsilon})(\mathbf{B}_\lambda^{-\mathsf{۱}})^T \\ &= \mathbf{B}_\lambda^{-\mathsf{۱}}(\sigma^{\mathsf{۲}}\mathbf{I}_n)(\mathbf{B}_\lambda^{-\mathsf{۱}})^T = \sigma^{\mathsf{۲}}\mathbf{B}_\lambda^{-\mathsf{۱}}(\mathbf{B}_\lambda^T)^{-\mathsf{۱}} \\ &= \sigma^{\mathsf{۲}}(\mathbf{B}_\lambda^T\mathbf{B}_\lambda)^{-\mathsf{۱}} = \sigma^{\mathsf{۲}}\mathbf{V}_\lambda. \end{aligned} \tag{۹.۲}$$

بنابراین مدل (۷.۲) را می‌توان به شکل زیر بازنویسی کرد:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\upsilon}, \quad \boldsymbol{\upsilon} \sim N(\circ, \sigma^{\mathsf{۲}}\mathbf{V}_\lambda). \tag{۱۰.۲}$$

در صورت رد فرضیه صفر در آزمون موران (۱۱.۱)، نسبت درستنمایی (۱۲.۱) یا ضریب لاگرانژ خطا (۱۳.۱) از این مدل استفاده می‌شود. در ادامه دو روش برآورد پارامترهای این مدل فضایی را بیان خواهیم کرد.

روش GLS

با توجه به ساختار غیرقطری ماتریس واریانس جمله خطا در معادله (۱۰.۲) نمی‌توان از روش OLS برای برآورد ضرایب استفاده کرد. در مقابل می‌توان از روش GLS استفاده کرد. همان‌طور که در فصل ۱ بیان شد، این روش در صورت وجود ماتریس متقارن $\mathbf{K}$ به‌طوری که

$$\mathbf{KK} = \mathbf{K}^T\mathbf{K} = \mathbf{V} = \mathrm{Var}(\boldsymbol{\upsilon})/\sigma^{\mathsf{۲}}$$

با اعمال تغییراتی در داده‌ها شرایط برآورد OLS را فراهم می‌آورد. با توجه به (۹.۲) در مدل SEM این شرط برقرار است، زیرا

$$\mathrm{Var}(\boldsymbol{\upsilon}) = \sigma^{\mathsf{۲}}\mathbf{V} = \sigma^{\mathsf{۲}}(\mathbf{B}^{-\mathsf{۱}})^T\mathbf{B}^{-\mathsf{۱}}.$$

با تغییر متغیرهای

$$\tilde{\mathbf{Y}} = \mathbf{B}\mathbf{y}, \quad \tilde{\mathbf{X}} = \mathbf{B}\mathbf{X}, \quad \tilde{\upsilon} = \mathbf{B}\boldsymbol{\upsilon}$$

خواهیم داشت

$$\mathrm{E}(\tilde{\boldsymbol{\upsilon}}) = ۰$$

و

$$\mathrm{Var}(\tilde{\boldsymbol{\upsilon}}) = \mathrm{Var}(\mathbf{B}\boldsymbol{\upsilon}) = \mathbf{B}\mathrm{Var}(\boldsymbol{\upsilon})\mathbf{B}^T = \mathbf{B}\sigma^۲(\mathbf{B}^{-۱})^T\mathbf{B}^{-۱}\mathbf{B}^T = \sigma^۲\mathbf{B}\mathbf{B}^{-۱}(\mathbf{B}^T)^{-۱}\mathbf{B}^T = \sigma^۲\mathbf{I}.$$

بنابراین

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-۱}\tilde{\mathbf{X}}^T\tilde{\mathbf{Y}}$$

برآوردگر OLS متغیرهای جدید است. در نتیجه، برآوردگر GLS پارامترهای مدل SEM به صورت زیر خواهد بود:

(۱۱.۲)

$$\hat{\boldsymbol{\beta}} = ((\mathbf{BX})^T(\mathbf{BX}))^{-۱}(\mathbf{BX})^T(\mathbf{By}) = (\mathbf{X}^T\mathbf{B}^T\mathbf{BX})^{-۱}\mathbf{X}^T\mathbf{B}^T\mathbf{By} = (\mathbf{X}^T\mathbf{VX})^{-۱}\mathbf{X}^T\mathbf{Vy}.$$

روش درستنمایی ماکسیمم

از آن‌جا که بردار تصادفی $\mathbf{y}$ ترکیبی خطی از بردار تصادفی $\boldsymbol{\upsilon}$ می‌باشد (رابطه (۱۰.۲) را ببینید)، بنابراین $\mathbf{y}$ نیز دارای توزیع نرمال n متغیره به‌شکل زیر است:

$$\mathbf{y} \sim \mathrm{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^۲\mathbf{V}_\lambda).$$

بنابراین تابع درستنمایی برای مشاهداتی از این بردار تصادفی به‌صورت

$$\begin{aligned} L(\boldsymbol{\theta}) = {} & (۲\pi)^{-\frac{n}{۲}}|\sigma^۲\mathbf{V}_\lambda|^{-\frac{۱}{۲}} \\ & \times \exp\Big\{-\frac{۱}{۲\sigma^۲}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}_\lambda^{-۱}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\Big\} \end{aligned}$$

است، که در آن $\boldsymbol{\theta} = [\lambda, \sigma^۲, \boldsymbol{\beta}^T]$. از ویژگی‌های دترمینان و هم‌چنین تقارن B نتیجه می‌شود

$$|\sigma^۲\mathbf{V}_\lambda| = |\sigma^۲(\mathbf{B}_\lambda^T\mathbf{B}_\lambda)^{-۱}| = (\sigma^۲)^n|\mathbf{B}_\lambda|^{-۲}.$$

هم‌چنین از (۸.۲) و (۱۰.۲) نتیجه می‌شود

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{B}_\lambda^{-۱}\boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &= \mathbf{B}_\lambda(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \mathbf{B}_\lambda\mathbf{y} - \mathbf{B}_\lambda\mathbf{X}\boldsymbol{\beta}. \end{aligned} \qquad (۱۲.۲)$$

بنابراین تابع لگاریتم درستنمایی مدل به‌شکل زیر به‌دست می‌آید:

$$\ln L = -\frac{n}{۲}\ln(۲\pi) - \frac{n}{۲}\ln(\sigma^۲) + \ln|B_\lambda| - \frac{۱}{۲\sigma^۲}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}. \qquad (۱۳.۲)$$

ماکسیمم کردن تابع (۱۳.۲) نسبت به β، معادل مینیمم کردن جمله آخر آن نسبت به β است. یعنی برآوردگر OLS پارامتر β در مدل (۱۲.۲) همان برآوردگر درستنمایی ماکسیمم (ML) β است. یعنی

$$\hat{\boldsymbol{\beta}}_{ML} = \Big((\mathbf{B}_\lambda \mathbf{X})^T (\mathbf{B}_\lambda \mathbf{X})\Big)^{-1} (\mathbf{B}_\lambda \mathbf{X})^T (\mathbf{B}_\lambda \mathbf{y}). \qquad (۱۴.۲)$$

همچنین برآوردگر درستنمایی ماکسیمم برای σ^2 با حل معادله زیر محاسبه می‌شود:

$$\begin{aligned} \frac{\partial \ln L}{\partial \sigma^2} &= \frac{-n}{2\sigma^2} + \frac{2\mathbf{e}^T\mathbf{e}}{4\sigma^4} \\ &= \frac{-n\sigma^2 + \mathbf{e}^T\mathbf{e}}{2\sigma^4} = 0 \end{aligned} \qquad (۱۵.۲)$$

که در آن $\mathbf{e} = \mathbf{B}_\lambda \mathbf{y} - \mathbf{B}_\lambda \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}$. نتیجه حل معادله (۱۵.۲) به برآوردگر زیر منتهی می‌شود:

$$\hat{\sigma}^2_{ML} = \frac{\mathbf{e}^T\mathbf{e}}{n}.$$

۲.۳.۲ مدل تاخیر فضایی

یک مدل خطی دیگر بر پایه خودهمبستگی فضایی، مدل تاخیر فضایی است. در این مدل فرض شده است که روابط خودهمبستگی در متغیرهای وابسته وجود دارند. اگر مشابه مدل قبل، فرض کنیم که این روابط فضایی از طریق ماتریس وزن $\mathbf{W} = (w_{ij} : i,j = 1,\cdots,n)$ با شرط $w_{ii} = 0$ لحاظ می‌شوند، آن‌گاه یک راه ساده برای نوشتن مدل به شکل زیر است:

$$y_i = \beta_0 + \rho \sum_{h \neq i} w_{ih} y_h + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \qquad i = 1,\cdots,n \qquad (۱۶.۲)$$

که در آن $\varepsilon_i \sim N(0,\sigma^2)$ و جمله خودهمبستگی، $\rho \sum_{h\neq i} w_{ih} y_h$، وابستگی ممکن y_i با مقادیر سایر نقاط، y_hها، را نشان می‌دهد. به‌عنوان یک مثال ساده برای این مدل می‌توان از قیمت مسکن در یک شهر نام برد. اگر y_i متوسط قیمت مسکن با ویژگی‌های $(x_{ij} : j = 1,\cdots,p)$ در ناحیه i باشد، آن‌گاه این قیمت تحت تاثیر قیمت مسکن‌های نواحی مجاور است. بنابراین خودهمبستگی فضایی در متغیرهای وابسته وجود دارد نه در جمله خطا.

مدل (۱۶.۲) را می‌توان به‌صورت ماتریسی زیر نوشت:

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n). \qquad (۱۷.۲)$$

معادله (۱۷.۲) را می‌توان به‌صورت زیر خلاصه کرد:

$$\begin{aligned} \mathbf{y} - \rho\mathbf{W}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ (\mathbf{I}_n - \rho\mathbf{W})\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \mathbf{B}_\rho \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \mathbf{y} &= \mathbf{B}_\rho^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{B}_\rho^{-1}\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n). \end{aligned} \qquad (۱۸.۲)$$

با تغییر متغیرهای $\mathbf{X}_\rho = \mathbf{B}_\rho^{-1}\mathbf{X}$ و $\boldsymbol{\upsilon} = \mathbf{B}_\rho^{-1}\varepsilon$، مدل (۱۸.۲) به شکل زیر نوشته می‌شود:

$$\mathbf{y} = \mathbf{X}_\rho\boldsymbol{\beta} + \boldsymbol{\upsilon}, \quad \boldsymbol{\upsilon} \sim N(0, \sigma^2\mathbf{V}_\rho) \tag{۱۹.۲}$$

که در آن مشابه مدل SEM، $\mathbf{V}_\rho = (\mathbf{B}_\rho^T\mathbf{B}_\rho)^{-1}$. در صورت رد فرضیه صفر در آزمون ضریب لاگرانژ تأخیر فضایی (۱۸.۱)، از این مدل برای مدل‌بندی ساختار فضایی استفاده می‌کنیم.

برآورد GLS

به منظور برآورد پارامترهای این مدل، معادله (۱۹.۲) را در نظر می‌گیریم. تمامی مراحل عیناً همانند مدل خطای فضایی دنبال می‌شوند و در نهایت برآورد GLS مدل تاخیر فضایی به صورت زیر است:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{B}\mathbf{y}. \tag{۲۰.۲}$$

برآورد ML

مشابه مدل SEM، تابع درستنمایی ماکسیمم بردار مشاهدات $\mathbf{y}$ به صورت زیر تعریف می‌شود:

$$L(\theta) = (2\pi)^{-\frac{n}{2}}|\sigma^2\mathbf{V}_\rho|^{-\frac{1}{2}} \times \exp\left\{\frac{-1}{2\sigma^2}(\mathbf{y} - \mathbf{X}_\rho\boldsymbol{\beta})^T\mathbf{V}_\rho^{-1}(\mathbf{y} - \mathbf{X}_\rho\boldsymbol{\beta})\right\}$$

که در این مدل $\boldsymbol{\theta} = [\rho, \sigma^2, \boldsymbol{\beta}^T]$. از (۱۸.۲) نتیجه می‌شود

$$\varepsilon = \mathbf{B}_\rho(\mathbf{y} - \mathbf{X}_\rho\boldsymbol{\beta}) = \mathbf{B}_\rho\mathbf{y} - \mathbf{B}_\rho\mathbf{B}_\rho^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{B}_\rho\mathbf{y} - \mathbf{X}\boldsymbol{\beta}. \tag{۲۱.۲}$$

بنابراین تابع لگاریتم درستنمایی مدل به‌صورت زیر نتیجه می‌شود:

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) + \ln|\mathbf{B}_\rho| - \frac{\varepsilon^T\varepsilon}{2\sigma^2}. \tag{۲۲.۲}$$

مشابه قبل، ماکسیمم کردن تابع (۲۲.۲) نسبت به $\boldsymbol{\beta}$ مستلزم مینیمم کردن جمله آخر آن نسبت به $\boldsymbol{\beta}$ است، یعنی طبق (۲۱.۲) برآوردگر OLS رگرسیون $\mathbf{B}_\rho\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$، برآوردگر درستنمایی ماکسیمم $\boldsymbol{\beta}$ است:

$$\hat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{B}_\rho\mathbf{y}. \tag{۲۳.۲}$$

هم‌چنین برآوردگر درستنمایی ماکسیمم σ^2 نیز مشابه مدل SEM با حل معادله (۱۵.۲)، به‌صورت زیر است:

$$\hat{\sigma}^2_{ML} = \frac{\mathbf{e}^T\mathbf{e}}{n}$$

با این تفاوت که $\mathbf{e} = \mathbf{B}_\rho\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}$.

برای محاسبه پارامترهای λ یا ρ، برآورد درستنمایی ماکسیمم پارامترهای $\boldsymbol{\beta}$ و $\sigma^۲$ را در تابع لگاریتم درستنمایی قرار داده و تابع لگاریتم درستنمایی متمرکز حاصل به شکل تابع غیرخطی از λ یا ρ را با استفاده از روش‌های عددی ماکسیمم می‌کنیم.

با اعمال تغییراتی روی دو مدل معرفی‌شده SEM و SLM، مدل‌های متنوعی تولید شده‌اند. در ادامه دو مدل پرکاربرد که در این پایان‌نامه نیز مورد استفاده قرار گرفته‌اند، را معرفی می‌کنیم.

۳.۳.۲ مدل ترکیبی فضایی

هنگام طراحی مدل تاخیر فضایی این سوال مطرح می‌شود که چرا تمام اثرات ناشناخته، ε_i، باید مستقل از فضا در نظر گرفته شوند. زیرا ممکن است که خودهمبستگی فضایی در جمله خطا و متغیر وابسته هر دو با هم وجود داشته باشد یا به عبارت دیگر شرایط SEM و SLM هم‌زمان برقرار باشند. بنابراین می‌توان ترکیب این دو مدل را به صورت ماتریسی زیر نوشت:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\upsilon}, \qquad \boldsymbol{\upsilon} = \lambda\mathbf{M}\boldsymbol{\upsilon} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(۰, \sigma^۲\mathbf{I}_n) \tag{۲۴.۲}$$

که در آن $\mathbf{M}$ نیز مشابه $\mathbf{W}$ ماتریس وزن‌های فضایی و λ مشابه ρ پارامتر وابستگی فضایی است. این مدل توسط کلجیان و پروچا (۲۰۱۰) معرفی شد و در متون اقتصادسنجی با نام مدل ترکیبی فضایی[۳] (SAC) یا مدل اتورگرسیو فضایی با خطاهای اتورگرسیو[۴] (SARAR) شناخته می‌شود. در این مدل، اولویت اصلی ایجاد یک مدل نسبی از ترکیب دو مدل تاخیر فضایی و خطای فضایی به‌عنوان دو مدل با ساختار مشابه است. از این رو تمرکز بر حالت خاص $\mathbf{W} = \mathbf{M}$ قرار می‌گیرد و مدل (۲۴.۲) به صورت

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\upsilon}, \qquad \boldsymbol{\upsilon} = \lambda\mathbf{W}\boldsymbol{\upsilon} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(۰, \sigma^۲\mathbf{I}_n)$$

نوشته می‌شود و به شکل زیر یک‌پارچه می‌شود:

$$\mathbf{y} = (\mathbf{B}_\rho)^{-۱}\mathbf{X}\boldsymbol{\beta} + (\mathbf{B}_\rho)^{-۱}(\mathbf{B}_\lambda)^{-۱}\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(۰, \sigma^۲\mathbf{I}_n). \tag{۲۵.۲}$$

در حالت خاص اگر $\rho = ۰$ $(\lambda = ۰)$، مدل (۲۵.۲) معادل SEM (SLM) می‌شود. این مدل در شرایطی که هر دو آزمون ضریب لاگرانژ خطا و آزمون ضریب لاگرانژ تأخیر فضایی معنی‌دار شوند، مورد استفاده قرار می‌گیرد.

استنباط مدل

برای برآورد ضرایب این مدل رگرسیونی از معادله (۲۵.۲) استفاده می‌کنیم و با تغییر متغیر

$$\mathbf{X}_\rho = \mathbf{B}_\rho^{-۱}\mathbf{X}$$

[3] Spatial autoregressive combined model

[4] Spatial autoregressive model with autoregressive disturbances

آن را به‌صورت

$$\mathbf{y} = \mathbf{X}_\rho \boldsymbol{\beta} + (\mathbf{B}_\lambda \mathbf{B}_\rho)^{-1}\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n) \tag{۲۶.۲}$$

که یک مدل خطای فضایی با جمله خطا $\mathbf{u} \sim N(0, \sigma^2 \mathbf{B}_\rho^{-1}\mathbf{V}_\lambda(\mathbf{B}_\rho^T)^{-1})$ است، بازنویسی می‌کنیم . بنابراین برآورد ضرایب رگرسیونی آن طبق (۱۱.۲) و (۱۴.۲) به‌صورت زیر محاسبه می‌شود:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= ((\mathbf{B}_\lambda \mathbf{B}_\rho \mathbf{X}_\rho)^T (\mathbf{B}_\lambda \mathbf{B}_\rho \mathbf{X}_\rho))^{-1} (\mathbf{B}_\lambda \mathbf{B}_\rho \mathbf{X}_\rho)^T (\mathbf{B}_\lambda \mathbf{B}_\rho \mathbf{y}) \\ &= ((\mathbf{B}_\lambda \mathbf{X})^T (\mathbf{B}_\lambda \mathbf{X}))^{-1} (\mathbf{B}_\lambda \mathbf{X})^T (\mathbf{B}_\lambda \mathbf{B}_\rho \mathbf{y}) \end{aligned} \tag{۲۷.۲}$$

همچنین مطابق با مدل SEM برآورد σ^2 به‌صورت زیر است:

$$\hat{\sigma}^2 = \frac{\mathbf{e}^T \mathbf{e}}{n}$$

که در آن $\mathbf{e} = \mathbf{B}_\lambda \mathbf{B}_\rho \mathbf{y} - \mathbf{B}_\lambda \mathbf{X} \hat{\boldsymbol{\beta}}$، محاسبه می‌شود.

۴.۳.۲ مدل دوربین فضایی

مفید و کاربردی‌ترین مدل اتورگرسیو مدل دوربین فضایی[۵] (SDM) است که توسط انسلین (۱۹۸۸a) معرفی شده است. این مدل از ایجاد تغییر در مدل تاخیر فضایی حاصل می‌شود و در مقایسه با آن از عملکرد بهتری برخوردار است. برای معرفی این مدل از مثال تعیین قیمت مسکن مورد استفاده در این پایان‌نامه شروع می‌کنیم. در این مورد اگر قیمت مسکن در ناحیه i، y_i، تحت تاثیر قیمت مسکن‌های همسایه باشد، آن‌گاه غیرمنطقی نیست که تحت تاثیر ویژگی‌های آن‌ها نیز باشد. بنابراین، این اثرات فضایی به‌عنوان شرایط اضافی به مدل افزوده می‌شود:

$$y_i = \beta_0 + \rho \sum_{h \neq i} w_{ih} y_h + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{h=1}^{n} m_{ih} (\sum_{j=1}^{p} \alpha_j x_{hj}) + \varepsilon_i, \qquad i = 1, \cdots, n. \tag{۲۸.۲}$$

برای نوشتن صورت ماتریسی این مدل، لازم است تا ماتریس $\mathbf{X}$ و بردار $\boldsymbol{\beta}$ را به شکل زیر افراز کنیم:

$$\mathbf{X} = \begin{bmatrix} 1_n & \mathbf{X}_v \end{bmatrix} \quad , \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_v \end{bmatrix}. \tag{۲۹.۲}$$

بنابراین می‌توان معادله (۲۸.۲) را به شکل

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \beta_0 1_n + \mathbf{X}_v \boldsymbol{\beta}_v + \mathbf{M}\mathbf{X}_v \boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$$

[۵] Spatial durbin model

یا

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{M}\mathbf{X}_v\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\circ, \sigma^{\mathsf{۲}}\mathbf{I}_n). \tag{۳۰.۲}$$

نوشت؛ که در آن $\boldsymbol{\alpha}$ یک بردار ستونی از ضرایب رگرسیونی با بعد p است. در این مدل نیز می‌توان وزن‌های فضایی را یکسان در نظر گرفت. یعنی

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_v\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\circ, \sigma^{\mathsf{۲}}\mathbf{I}_n). \tag{۳۱.۲}$$

استنباط مدل

برآورد ضرایب این مدل رگرسیونی کاملا مشابه مدل تأخیر فضایی است و تنها با تغییر متغیر انجام می‌شود. به این منظور، ماتریس متغیرهای تبیینی را به‌صورت

$$\mathbf{Z} = [\mathbf{X} \quad \mathbf{W}\mathbf{X}_v]$$

که ماتریسی با بعد $n \times (\mathsf{۲}p + \mathsf{۱})$ است و بردار ضرایب رگرسیونی را به‌شکل

$$\boldsymbol{\delta} = [\boldsymbol{\beta} \quad \boldsymbol{\alpha}]^T$$

که یک بردار با بعد $(\mathsf{۲}p + \mathsf{۱}) \times \mathsf{۱}$ است، تعریف می‌کنیم. با این تعاریف معادله (۳۱.۲) به شکل زیر بازنویسی می‌شود:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\circ, \sigma^{\mathsf{۲}}\mathbf{I}_n). \tag{۳۲.۲}$$

برآورد ضرایب رگرسیونی مدل (۳۲.۲)، که یک مدل تاخیر فضایی است، مطابق (۲۳.۲) و (۲۰.۲) به‌شکل

$$\hat{\boldsymbol{\delta}} = (\mathbf{Z}^T\mathbf{Z})^{-\mathsf{۱}}\mathbf{Z}^T\mathbf{B}_\rho\mathbf{y} \tag{۳۳.۲}$$

همچنین برآورد $\sigma^{\mathsf{۲}}$، مشابه SLM به‌صورت

$$\hat{\sigma}^{\mathsf{۲}} = \frac{\mathbf{e}^T\mathbf{e}}{n}$$

که در آن $\mathbf{e} = \mathbf{B}_\rho\mathbf{y} - \mathbf{Z}\boldsymbol{\delta}$ است، محاسبه می‌شوند.

تمامی مدل‌هایی که تا کنون بیان شدند، مدل‌های فراموضعی[۶] فضایی هستند.

تعریف ۱.۳.۲. مدل‌های فراموضعی: مدل‌هایی هستند که ضرایب برآوردشده در مدل برای پیشگویی متغیر پاسخ $\hat{y}$ به ازای هر مقدار $x_\circ$ در کل ناحیه تحت مطالعه ثابت است.

[۶] Global model

در عمل و به خصوص در مورد داده‌های فضایی ممکن است متغیرهای تبیینی در نواحی مختلف دارای اثرات متفاوت باشند؛ به عبارتی داده‌ها ناهمگن فضایی باشند. در مثال تعیین قیمت مسکن ذکرشده، این حقیقت وجود دارد که با وجود وابستگی قیمت مسکن ناحیه i به نواحی مجاور، اما متغیرهای موثر بر قیمت آن ممکن است دارای اثرات متفاوت از نواحی مجاور باشند. برای مثال، تعداد پارکینگ‌های یک منزل در تعیین قیمت آن موثر هستند. اما این متغیر در نواحی جنوبی یک شهر که افراد با درامد کمتر در آن سکونت دارند و اغلب فاقد وسیله نقلیه هستند، از اهمیت کمتری نسبت به نواحی دیگر برخوردار است. لذا لازم است که برای عملکرد بهتر، این ناهمگنی در نظر گرفته شود.

تعریف ۲.۳.۲. مدل‌های موضعی[۷]: مدل‌هایی هستند که ضرایب برآوردشده در مدل برای پیشگویی متغیر پاسخ به ازای هر مقدار $x_\circ$ ثابت نبوده و وابسته به آن هستند. یعنی در مدل‌های موضعی، برای برآورد متغیر پاسخ در نقطه $x_\circ$ ابتدا یک همسایگی از آن نقطه انتخاب شده سپس برازش مدل فقط بر روی مشاهدات داخل همسایگی صورت می‌گیرد نه بر روی کل مشاهدات. بنابراین واضح است که با تغییر $x_\circ$ همسایگی نیز تغییر نموده و چون مشاهدات داخل همسایگی نیز تغییر می‌کنند، مدل برازش‌شده فعلی با مدل قبلی متفاوت خواهد بود.

یکی از کاربردی‌ترین مدل‌های موضعی، رگرسیون موزون جغرافیایی است که در ادامه به اختصار آن را معرفی می‌کنیم.

۴.۲ رگرسیون موزون جغرافیایی

در بخش‌های پیشین بیان کردیم که مدل‌های رگرسیونی فراموضعی فضایی در مواجهه با ناهمگنی فضایی مناسب نیستند. ناهمگنی فضایی بیانگر این حقیقت است که در هر منطقه رابطه‌ای متفاوت بین متغیر پاسخ و متغیرهای تبیینی وجود دارد. از این رو اگر این ناهمگنی در نظر گرفته نشود و رابطه‌ای یکسان برای تمامی مناطق برآورد شود، تخمینی نادقیق از روابط منطقه‌ای نتیجه خواهدشد. از میان روش‌های ناپارامتری، رگرسیون موزون موضعی[۸] (LWR) (سلولند و دولین، ۱۹۸۸) یا رگرسیون موزون جغرافیایی که توسط فودرینگ‌هام و همکاران (۲۰۰۳) به‌طور کامل معرفی شد؛ این امکان را فراهم می‌آورند که پارامترها را (به جای فراموضعی) ناحیه‌ای برآورد کنیم. ایده اساسی برآورد موضعی این است که توابع خطی ساده ممکن است برای مشاهدات نزدیک به یک نقطه مناسب باشند، اما احتمالا زمانی که مشاهدات دوردست بیشتر گنجانده شوند، نامناسب خواهند بود. بنابراین محدود کردن برآورد به یک ناحیه از مشاهدات، مشکل واریانس نامتجانس و خودهمبستگی را که در داده‌های فضایی وجود دارند، رفع می‌کند.

[7] Local models

[8] Locally weighted regression

در روش رگرسیون موزون موضعی پارامترهای مدل را می‌توان در هر نقطه‌ای از فضای مورد مطالعه برآورد کرد. یعنی برای هر نقطه در فضا یک معادله رگرسیونی جداگانه تعریف می‌شود. در این روش برای برآورد پارامترهای مدل در هر نقطه از مشاهدات اطراف آن نقطه استفاده می‌شود، اما به مشاهدات نزدیک وزن بیشتر و به مشاهدات دور وزن کمتری داده می‌شود. این امر نیز با ساختار داده‌های فضایی، که در موقعیت‌های نزدیک به هم دارای همبستگی بیشتر هستند، متناسب است. بنابراین، تابع خطی ساده را می‌توان برای برآورد موضعی پارامترها به روش زیر نوشت:

$$\mathbf{y} = (\boldsymbol{\beta} \odot \mathbf{X})\mathbf{1} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim MVN(0, \sigma_{\varepsilon}^{2}\mathbf{I}) \tag{۳۴.۲}$$

که در آن $\mathbf{y}$ بردار ستونی با بعد n و $\boldsymbol{\beta}$ یک ماتریس $n \times (p+1)$ با سطر iام

$$\boldsymbol{\beta}_i = (\beta_{i0}, \beta_{i1}, \cdots, \beta_{ip})^T$$

است. عملگر $\odot$ ضرب هادامارد و ماتریس $\mathbf{X}$ دارای بعد یکسان با $\boldsymbol{\beta}$ و سطر iام

$$\mathbf{x}_i = (x_{i0}, x_{i1}, \cdots, x_{ip})$$

است. بردار ستونی ε، nبعدی و $\mathbf{1}$ یک بردار ستونی $(p+1)$بعدی از ۱ها است. مدل (۳۴.۲) را می‌توان برای هر مشاهده، به‌صورت

$$y_i = \beta_{i0} + \beta_{i1}x_{i1} + \cdots + \beta_{ip}x_{ip} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^{2}) \quad i = 1, 2, \cdots, n \tag{۳۵.۲}$$

نوشت.

با این‌که در صورت کلی مدل GWR واریانس جمله خطا ثابت فرض شده است $(\mathrm{Var}(\varepsilon) = \sigma^{2})$ و ساختار آن عملا خودهمبستگی و ناهمگنی را نشان نمی‌دهد، اما باید توجه داشت که اکثر این اثرات با برآورد ضرایب مختلف برای هر مشاهده با استفاده از همسایه‌های آن، از بین رفته است.

۱.۴.۲ استنباط مدل

به منظور برآورد یک مجموعه از پارامترهای نقطه i، می‌توان تابع هدف زیر را به روش کمترین توان‌های دوم، در زیرمجموعه‌ای از نقاط که نزدیک i هستند، کمینه کرد:

$$Q(\boldsymbol{\beta}_i) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_i)^T\mathbf{W}_i(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_i) = \boldsymbol{\varepsilon}^T\mathbf{W}_i\boldsymbol{\varepsilon} \quad i = 1, 2, \cdots, n \tag{۳۶.۲}$$

که در آن $\mathbf{W}_i$ یک ماتریس قطری از بعد n است، که درایه‌های غیرقطری آن صفر و درایه‌های روی قطر اصلی یعنی $(w_{i1}, w_{i2}, \cdots, w_{in})$ وزن هریک از n مشاهده برای نقطه i است. یعنی

$$\mathbf{W}_i = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix}.$$

این وزن‌ها در مدل GWR فضایی، متناسب با فواصل جغرافیایی بین مشاهدات هستند. به این ترتیب برآوردگر کمترین توان‌های دوم موزون موضعی[9] (LWLS)، با تکرار WLS برای هر مشاهده i به شکل زیر به‌دست می‌آید:

$$\hat{\boldsymbol{\beta}}_i^{WLS} = (\mathbf{X}^T\mathbf{W}_i\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_i\mathbf{y}, \qquad i = 1, \cdots, n. \tag{۳۷.۲}$$

بنابراین ما n ماتریس قطری از وزن‌های فضایی و n مجموعه از برآوردگرهای پارامترهای موضعی داریم. می‌توان برآوردگر موزون موضعی را به‌صورت

$$\hat{\boldsymbol{\beta}}_i^{WLS} = \mathbf{C}\mathbf{y} \quad i = 1, 2, \cdots, n$$

نوشت که درآن

$$\mathbf{C} = (\mathbf{X}^T\mathbf{W}_i\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_i.$$

امید و واریانس $\boldsymbol{\beta}_i$ برابر هستند با

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\beta}}_i) &= \mathrm{E}(\mathbf{C}\mathbf{y}) = (\mathbf{X}^T\mathbf{W}_i\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_i\mathrm{E}(\mathbf{y}) \\ &= (\mathbf{X}^T\mathbf{W}_i\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_i\mathbf{X}\boldsymbol{\beta}_i = \boldsymbol{\beta}_i \end{aligned} \tag{۳۸.۲}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_i) = \mathrm{Var}(\mathbf{C}\mathbf{y}) = \mathbf{C}\mathrm{Var}(\mathbf{y})\mathbf{C}^T = \sigma^2\mathbf{C}\mathbf{C}^T \tag{۳۹.۲}$$

که در آن σ^2 واریانس جمله خطا در رگرسیون موضعی است. علاوه بر ارزیابی برآوردهای پارامترهای موضعی، بررسی خطاهای استاندارد موضعی نیز حائز اهمیت است. این کار به منظور محاسبه تغییرات داده‌های مورد استفاده برای تخمین برآوردها باید انجام شود. برای مثال، در بعضی موارد ممکن است که برآورد پارامترهای موضعی تابعی از تعداد مشاهدات نسبتا کم باشد؛ یا برخی مشاهدات به دلیل دور بودن از نقطه رگرسیونی در برآورد پارامترهای موضعی دارای وزن کوچک باشند. بنابراین محاسبه تغییرات خطای استاندارد موضعی لازم است.

[9] Locally weighted least squares

همچنین طبق (۳۷.۲) برای برآورد ضرایب رگرسیونی مشاهده i از ماتریس وزن W_i استفاده می‌شود. با این کار مشاهدات همسایه i که بیشترین وابستگی را با آن دارند، در برآورد ضرایب رگرسیونی آن نیز بیشترین تاثیر را خواهند گذاشت. لذا واضح است که با تغییر نقاط رگرسیونی، $i = ۱, \cdots, n$، ماتریس‌های وزن و تحت تاثیر آن ضرایب رگرسیونی موضعی β_i و باقی‌مانده‌های استانداردشده رگرسیون موضعی تغییر خواهند کرد. یعنی تغییرات باقی‌مانده‌های استانداردشده با تغییر نقاط رگرسیونی متفاوت می‌شود. لذا برآورد واریانس جمله خطا در رگرسیون موضعی برای هر مشاهده $i = ۱, \cdots, n$ به‌صورت

$$\hat{\sigma}^۲ = \frac{\sum_i (y_i - \hat{y}_i)^۲}{(n - ۲tr(\mathbf{S}) + tr(\mathbf{S}^T\mathbf{S}))} \qquad (۴۰.۲)$$

محاسبه می‌شود؛ که در آن $\mathbf{S}$ ماتریس طرح $\hat{\mathbf{y}}$ روی $\mathbf{y}$

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$$

است و هر سطر آن به‌صورت

$$\mathbf{s}_i = \mathbf{X}_i(\mathbf{X}^T\mathbf{W}_i\mathbf{X})^{-۱}\mathbf{X}^T\mathbf{W}_i \qquad (۴۱.۲)$$

تعریف می‌شود. در تمام مطالب گفته‌شده در مورد رگرسیون موزون جغرافیایی، ابتدا لازم است که ماتریس وزن فضایی $\mathbf{W}_i$ مشخص شود.

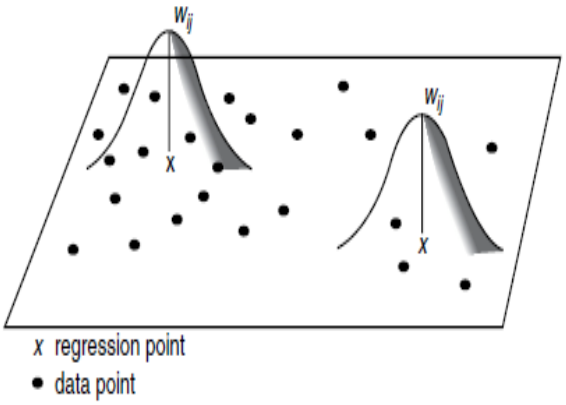
۲.۴.۲ روش‌های انتساب وزن

توابع وزن‌دهی هسته باید در عبارت (۳۷.۲) تعریف شوند تا $\mathbf{W}_i$ مشخص گردد. این توابع عموما نمایی، گاوسی، توان‌دو[۱۰]، توان‌سه[۱۱] هستند که مشابه یکدیگر در نقاط نزدیک وزن بزرگ‌تر و در نقاط دور وزن کوچک‌تر می‌گیرند؛ اما با توجه به نوع پهنای باند[۱۲] متفاوتند. انتخاب پهنای باند بسیار مهم و حیاتی است. زیرا همسایگی مشاهده i توسط مقدار پهنای باند b مشخص می‌شود. این مقدار، تعداد مشاهداتی را که در برآورد برای نقطه i وزن می‌گیرند و نیز چگونگی کاهش وزن با افزایش فاصله را تعیین می‌کند. پهنای باند می‌تواند یک فاصله ثابت[۱۳] یا سازوار[۱۴] در نظر گرفته شود.

رگرسیون موزون فضایی با هسته فضایی ثابت

در روش رگرسیون موزون با هسته فضایی ثابت، دامنه‌ای در اطراف هر نقطه مرجع در نظر گرفته می‌شود و این فاصله به ازای هر نقطه ثابت است.

[10] Bi-square
[11] Tri-cube
[12] Bandwidth
[13] Fixed
[14] Adaptive



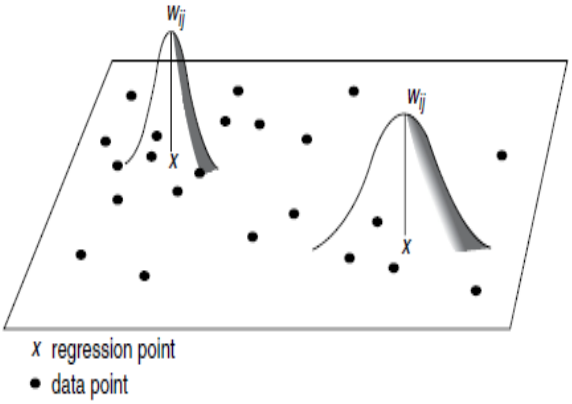
شکل ۱.۲: رگرسیون موزون فضایی با هسته فضایی ثابت

مدل‌های رگرسیون موزون فضایی با هسته ثابت[۱۵] (شکل ۱.۲ را ببینید)، زمانی که توزیع مشاهدات حول نقاط مرجع در گستره فضای مورد مطالعه یکسان نباشد، دچار مشکل می‌شود. در این صورت اگر تابعی یکسان برای موزون کردن مشاهدات اطراف نقاط مرجع به کار گرفته شود، این احتمال وجود دارد که برای نقطه‌ای از اطلاعاتی کمتر استفاده شود و بنابراین مدلی نادقیق برآورد شود و بالعکس برای نقطه‌ای دیگر که تجمع مشاهدات حول آن بیشتر است از اطلاعات زیاد و نامرتبط استفاده شود و مجددا از دقت مدل کاسته شود. برای مقابله با این احتمال، رگرسیون موزون فضایی با هسته فضایی سازوار[۱۶] معرفی شده است (فودرینگ‌هام و همکاران، ۲۰۰۳).

رگرسیون موزون فضایی با هسته فضایی سازوار

در روش رگرسیون موزون فضایی با هسته‌ی سازوار، پهنای باند تابع وزن‌دهی با پراکندگی مشاهدات حول نقطه مرجع تطبیق داده می‌شود. به‌طوری که زمانی که مشاهدات پراکنده هستند، پهنای باند بزرگ‌تری در نظر گرفته می‌شود و زمانی که مشاهدات حول نقطه مرجع متراکم باشند، پهنای باند کوچک‌تری انتخاب می‌شود (شکل ۲.۲ را ببینید). به عبارتی در این روش به جای فاصله ثابت از تعداد ثابت برای تعیین پهنای باند استفاده می‌شود. این روش تعیین همسایگی k ـ نزدیک‌ترین همسایگی[۱۷] نیز نامیده می‌شود.

[15] Fixed kernel
[16] Adaptive kernel
[17] k-nearest neighbour



شکل ۲.۲: رگرسیون موزون فضایی با هسته فضایی سازوار

در ادامه نمونه‌هایی از توابع هسته ثابت و سازوار را معرفی می‌کنیم.

توابع هسته

ساده‌ترین روش، استفاده از یک طرح وزن‌دهی دوتایی به‌صورت زیر است (فودرینگ‌هام و همکاران، ۲۰۰۳):

$$w_{ij} = \begin{cases} ۱ & \text{اگر } d_{ij} \leq b \\ ۰ & \text{سایر جاها} \end{cases}$$

که در آن d_{ij} فاصله بین مشاهدات i و j و b پهنای باند است. این تابع هسته در مواردی که نقاط کالیبره[۱۸] (نقاطی که در مرزهای پهنای باند قرار می‌گیرند) در یک منطقه با تراکم کم (تُنُک) واقع شده است در مقایسه با یک منطقه متراکم، می‌تواند دست‌آورد بهتری داشته باشد.

یک تابع هسته دیگر به‌صورت

$$w_{ij} = \begin{cases} ۱ & \text{اگر } y_j \in \mathbf{Y}_i(N) \\ ۰ & \text{سایر جاها} \end{cases}$$

معرفی می‌شود (ویلر و پائز، ۲۰۱۰). در این تابع هسته $\mathbf{Y}_i(N)$ مجموعه‌ای از N نزدیکترین مشاهدات همسایه نقطه i است. N مقداری است که باید برآورد شود. در این روش بر خلاف روش قبلی، تابع هسته برای هر نقطه از تعداد یکسانی از مشاهدات استفاده می‌کند، اما این تعداد از مشاهدات ممکن است که برای هر نقطه i در مساحت‌های متفاوتی قرار گرفته باشند. با وجود این سادگی، این نوع تابع هسته به‌طور وسیعی استفاده نشده است.

[18] Calibration

در بیشتر موارد در مدل GWR استفاده از توابع هسته پیوسته ترجیح داده می‌شود، به این دلیل که این توابع وزن‌هایی تولید می‌کنند که به‌طور یکنواخت با افزایش فاصله کاهش می‌یابند. به‌عنوان نمونه تابع هسته گاوسی (فودرینگ‌هام و همکاران، ۲۰۰۳) را با ضابطه

$$w_{ij} = \exp\left\{ -\circ/\Delta\left(\frac{d_{ij}}{b}\right)^{\Upsilon}\right\}$$

می‌توان معرفی کرد که در این تابع وزن مشاهده j در ارتباط با مشاهده i، براساس فاصله این دو مشاهده از هم، d_{ij}، تغییر می‌کند و b پهنای باند است که حد فاصله و نزول خودهمبستگی فضایی را کنترل می‌کند. همان‌طور که در بخش‌های قبلی گفته شد، پهنای باند می‌تواند به دو روش ثابت و سازوار مورد استفاده قرار گیرد. در این پایان‌نامه از تابع هسته گاوسی با پهنای باند سازوار استفاده می‌کنیم.

در این حالت تابع هسته گاوسی به شکل زیر معرفی می‌شود (ویلر و پائز، ۲۰۱۰):

(۴۲.۲)

$$w_{ij} = \begin{cases} \exp\left(-\circ/\Delta\left(\frac{d_{ij}}{d_{iN}}\right)^{\Upsilon}\right) & \text{اگر } j \text{ یکی از } N \text{ نزدیک‌ترین همسایگی‌های } i \text{ باشد} \\ \circ & \text{سایر جاها} \end{cases}$$

که در آن d_{iN} فاصله نقطه‌ی i تا Nامین (آخرین) نزدیک‌ترین همسایگی است. مقدار N به منظور محاسبه وزن‌ها برای برآورد ضرایب رگرسیونی موضعی باید تعیین شود.

تابع هسته دیگر، تابع نمایی ساده به صورت

$$w_{ij} = \exp\left(-\frac{d_{ij}}{b}\right) \qquad \text{اگر} \quad d_{ij} < b$$

است که در آن مقیاس‌گذاری و توان نسبت به مشابه گاوسی خود، حذف شده است (ویلر و پائز، ۲۰۱۰).

همچنین تابع هسته توان‌دو به صورت

$$w_{ij} = \begin{cases} \left(\Upsilon - \frac{d_{ij}^{\Upsilon}}{b^{\Upsilon}}\right)^{\Upsilon} & \text{اگر} \quad d_{ij} < b \\ \circ & \text{سایر جاها} \end{cases}$$

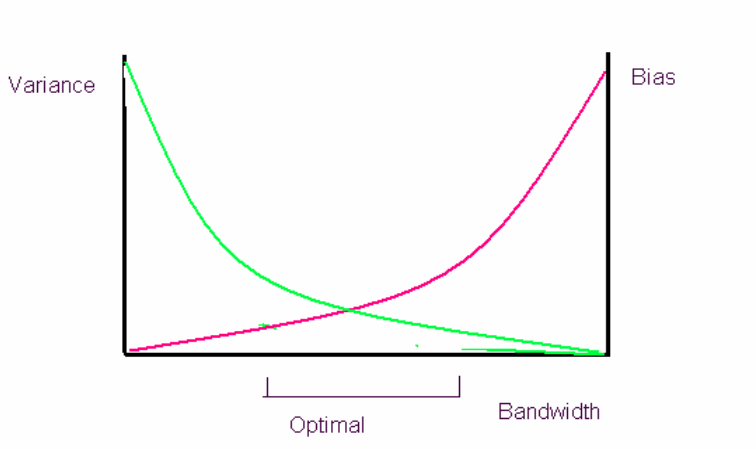
است (فودرینگ‌هام و همکاران، ۲۰۰۳). شکل سازوار این تابع هسته، مشابه تابع هسته گاوسی سازوار تعریف می‌شود (ویلر و پائز، ۲۰۱۰). یعنی

$$w_{ij} = \begin{cases} \left(\Upsilon - \left(\frac{d_{ij}}{d_{iN}}\right)^{\Upsilon}\right)^{\Upsilon} & \text{اگر } j \text{ یکی از } N \text{ نزدیک‌ترین همسایگی‌های } i \text{ باشد} \\ \circ & \text{سایر جاها} \end{cases}$$

همان‌طور که در تمامی توابع هسته گفته شد، مقدار پهنای باند در محاسبه وزن‌ها بسیار اثرگذار است. به‌همین منظور تعیین فاصله آن در حالت هسته ثابت و تعیین مقدار N در حالت هسته سازوار بسیار حیاتی است. در بخش بعد به معرفی چند معیار برای تعیین این مقادیر می‌پردازیم.

۳.۴.۲ انتخاب بهترین همسایگی

همان‌طور که گفته شد، انتخاب پهنای باند و در نتیجه آن انتخاب همسایگی مناسب، در برآورد ضرایب بسیار موثر است. بنابراین سوال مهم این است که چگونه می‌توان یک همسایگی مناسب برای i تعریف کرد. در انتخاب b همواره یک مبادله میان واریانس و اریبی رخ می‌دهد (شکل ۳.۲). زیرا در برآورد WLS پارامتر نقطه i، $\hat{\boldsymbol{\beta}}_i^{WLS}$، اگر پهنای باند را کاهش دهیم و فقط به مشاهدات خیلی نزدیک بسنده کنیم، با این کار به دلیل کاهش اندازه نمونه، خطای استاندارد $\hat{\boldsymbol{\beta}}_i^{WLS}$ افزایش می‌یابد. همچنین با افزایش b خطای استاندارد کاهش می‌یابد و اریبی افزایش می‌یابد.



شکل ۳.۲: مبادله واریانس–اریبی

برای حل مشکل انتخاب پهنای باند بهینه، b^{opt}، اعتبارسنجی متقابل[19] CV کاربرد گسترده‌ای دارد (سیلورمن، ۱۹۸۴). به‌عبارتی اعتبارسنجی متقابل به‌دنبال پهنای باندی برای تابع هسته است که باقی‌مانده‌های برآورد y_iها توسط یک زیرمجموعه از داده‌ها را مینیمم کند. این رهیافت برای رگرسیون موضعی توسط سلولند (۱۹۷۹) و برای برآورد تابع هسته توسط بومن (۱۹۸۴)، پیشنهاد شد.

برونس‌دان و همکاران (۱۹۹۶) رهیافت CV را به‌صورت

$$\mathrm{CV} = \sum_{i=1}^{n} (y_i - \hat{y}_{\neq i}(b))^2$$

برای رگرسیون موزون جغرافیایی معرفی کردند که در آن $\hat{y}_{\neq i}(b)$ مقدار برآوردشده Y_i پس از حذف مشاهده i از فرآیند است و مجموع توان‌های دوم باقی‌مانده پس از حذف مشاهده i کمینه می‌شود. به‌عنوان مثال، فودرینگ‌هام و همکاران (۲۰۰۳) در تحلیل داده‌های قیمت

[19] Cross-validation

مسکن شهر لندن برای انتخاب بهترین همسایگی در رگرسیون موزون جغرافیایی با هسته فضایی سازوار، از این معیار استفاده کردند.

یک رهیافت متفاوت و کلی‌تر که بر پایه مبادله بین درجه آزادی و نیکی برازش است، معیار اطلاع آکاییک[۲۰] (AIC) است (آکاییک، ۱۹۷۶). در این روش مقدار بهینه پهنای باند از طریق کمینه کردن مقدار AIC به‌دست می‌آید. این معیار در متون مختلف با صورت‌های مختلف ارائه شده است. فودرینگ‌هام و همکاران (۲۰۰۳) یک نسخه تصحیح‌شده از معیار اطلاع آکاییک را برای GWR، پیرو روش معرفی‌شده توسط هورویچ و همکاران (۱۹۹۸)، به شکل زیر معرفی کردند:

$$\mathrm{AIC}_c = ۲n\ln(\hat{\sigma}_\varepsilon) + n\ln(۲\pi) + n\left\{\frac{n + tr(\mathbf{S})}{n - ۲ - tr(\mathbf{S})}\right\}$$

بنابراین مقدار پهنای باند بهینه به‌صورت

$$b^{opt} = \min_b \mathrm{AIC}_c = \min_b \left\{ ۲n\ln(\hat{\sigma}_\varepsilon) + n\ln(۲\pi) + n\left\{\frac{n + tr(\mathbf{S})}{n - ۲ - tr(\mathbf{S})}\right\}\right\} \qquad (۴۳.۲)$$

است، که در آن $\hat{\sigma}_\varepsilon$ برآورد انحراف معیار استاندارد خطا در معادله (۳۴.۲) است. توجه داشته باشید که برآورد σ_ε به روش (۴۰.۲) محاسبه نمی‌شود؛ بلکه بر اساس درستنمایی ماکسیمم به‌صورت

$$\hat{\sigma}_\varepsilon^۲ = \frac{\mathrm{RSS}}{n}$$

برآورد می‌شود؛ که در آن RSS مجموع توان‌های دوم باقی‌مانده در رگرسیون موزون جغرافیایی است. (فودرینگ‌هام و همکاران، ۲۰۰۳).

در مدل GWR درجه آزادی برابر $n - (۲tr(\mathbf{S}) - tr(\mathbf{S}^T\mathbf{S}))$ است، که در آن $tr(\mathbf{S})$ مجموع درایه‌های روی قطر اصلی ماتریس طرح (۴۱.۲) است. $tr(\mathbf{S})$ در رگرسیون OLS برابر تعداد پارامترهای مدل p است. اما در مدل GWR تعداد پارامترهای موثر[۲۱] مدل برابر

$$p \leq ۲tr(\mathbf{S}) - tr(\mathbf{S}^T\mathbf{S}) \leq n$$

است که در آن p تعداد پارامترها در مدل فراموضعی متناظر است. چون عموما $tr(\mathbf{S})$ و $tr(\mathbf{S}^T\mathbf{S})$ مشابه هستند، تعداد پارامترهای موثر مدل تقریبا برابر $tr(\mathbf{S})$ است؛ بنابراین مشخص است که اگر $۲tr(\mathbf{S}) - tr(\mathbf{S}^T\mathbf{S}) = p$ آن‌گاه پهنای باند به بی‌نهایت میل می‌کند و برای برآورد هر نقطه مرجع i از تمام مشاهدات نمونه استفاده می‌شود. اما اگر $۲tr(\mathbf{S}) - tr(\mathbf{S}^T\mathbf{S}) = n$ آن‌گاه پهنای باند به صفر میل می‌کند.

رگرسیون موزون جغرافیایی، مشکل ناهمگنی فضایی را با برازش مدل خطی به تک تک مشاهدات و تغییرات هموار ضرایب رگرسیونی روی فضای مشاهدات نمایش می‌دهد. در واقع ضرایب GWR وابسته به فضا تغییر می‌کنند، یعنی به ازای هر مشاهده $p+۱$ ضریب به مدل اضافه

[20] Akaike information criterion

[21] Effective number of parameters

می‌شود. این امر بررسی آن‌ها را در مواردی که تعداد مشاهدات و ضرایب مورد بررسی زیاد است، دشوار می‌کند و نمی‌تواند تفاوت معنی‌دار آماری ارائه دهد. موضوع اصلی در تحلیل‌های فضایی ارائه روش‌های جدید موضعی برای تحلیل فضایی یا ارائه مدل‌های پیچیده‌تر برای ساختار داده‌ها نیست، بلکه ارائه ابزارها برای کشف جامعه و مکانیسم رفتاری که الگوهای فضایی نشان داده‌اند، مهم است. به گفته انسلین (۲۰۱۰): در مدل‌های ناهمگنی فضایی، ضرایب مختلف فضایی گواه حضور ناهمگنی است؛ اما آن را توضیح نمی‌دهند. این در حالی است که همیشه هدف ایده‌آل تهیه یک ساختار از ناهمگنی و (یا) خودهمبستگی است.

بر این اساس، در فصل سوم مدلی را که همچون GWR خودهمبستگی و ناهمگنی فضایی را توام در نظر می‌گیرد، اما ایراد وارد بر GWR یعنی عدم توضیح‌پذیری به آن وارد نیست، معرفی می‌کنیم.

فصل ۳

مدل رگرسیون موزون موضعی تکراری

۱.۳ مقدمه

مشکل ناهمگنی فضایی به شکل پارامترهای مختلف فضایی در تحلیل‌های فضایی و به ویژه اقتصادسنجی فضایی ناشناخته مانده است؛ چرا که عموما هدف آن‌ها کنترل اثرات سرریز و خودهمبستگی فضایی است. به گفته پوستیج‌لیون و همکاران (۲۰۱۳) مشکل ناهمگنی فضایی اغلب در تحلیل‌های اقتصادی داده‌های فضایی کنار گذاشته و فراموش می‌شود و این عدم توجه می‌تواند در برآورد ضرایب مدل اثرات مخرب آشکاری داشته باشد.

گروهی از محققان نظیر انسلین (۱۹۸۸b)، پد و همکاران (۲۰۱۴) و برخی دیگر قصد داشتند که حضور ناهمگنی فضایی را با ساخت آزمون‌هایی که معمولا بر پایه آماره ضریب لاگرانژ هستند، آشکار کنند. متاسفانه هنگامی که حضور ناهمگنی فضایی شناسایی شد، هیچ آزمونی قادر به نشان دادن مدل‌بندی درست داده‌های فضایی نیست.

اخیرا ابراگیموف و مولر (۲۰۱۰) ویژگی‌های آماره t را در نمونه‌های کوچک و بزرگ و همچنین در برخورد با داده‌های فضایی، با این فرض که داده‌ها به q گروه منطقی افراز شده باشند، بررسی کردند. سوال کلیدی این رهیافت در تعداد و ترکیب گروه‌ها است. در این مورد آن‌ها تاکید کردند که نمی‌توان تصمیم‌گیری درباره انتخاب گروه مناسب را به داده‌ها محول کرد و به اطلاعات پیشین در خصوص ساختار همبستگی آن‌ها نیاز است. این در حالی است که در بسیاری از موارد عملی هیچ‌گونه اطلاعی درباره ساختار همبستگی و ناهمگنی داده‌ها

که باعث شود گروهی به دیگری ترجیح داده‌شود، وجود ندارد.

در این فصل یک مدل ایده‌آل برای ساختار داده‌های فضایی معرفی می‌شود که ناهمگنی و وابستگی فضایی را با هم در بر دارد. مدل پیشنهادی ما شامل یک روش قدرتمند تکراری بر پایه ترکیب روش رگرسیون موزون جغرافیایی و الگوریتم هموارسازی وزن‌های سازوار[۱] (AWS) است. با ترکیب این دو روش، نواحی همگن گروه‌بندی و مشخص می‌شوند. در متون اقتصادی به این نواحی همگن، رژیم‌های فضایی می‌گویند. همچنین می‌توان با لحاظ کردن این رژیم‌های فضایی در یک مدل فراموضعی، ضرایب رگرسیونی را به‌طور مجزا برای هر رژیم برآورد و نارسایی مدل GWR در توضیح و نمایش ساختار مدل را رفع کرد. بیل و همکاران (۲۰۱۷) این روش ترکیبی برای تعیین رژیم‌های همگن را رگرسیون موزون موضعی تکراری[۲] (IGWR) و مدل‌های حاصل بر مبنای استفاده از رژیم‌های تعیین‌شده توسط IGWR را مدل‌های با رژیم فضایی درونی[۳] (ESRMs) نامیدند.

در ادامه این فصل رهیافت پیشنهادی خود را در دو مرحله کلی شامل:

مرحله اول: الگوریتم کنترل ناهمگنی فضایی (IGWR)

مرحله دوم: مدل‌های اقتصادسنجی فضایی با رژیم فضایی درونی (ESRMs)

را شرح خواهیم داد.

۲.۳ مرحله اول: کنترل ناهمگنی فضایی

در مرحله اول رهیافت، یعنی IGWR، نواحی همگن توسط یک الگوریتم تکراری مشخص می‌شوند. ایده این روش از الگوریتم AWS که اولین بار توسط پولزهل و اسپوکوینی (۲۰۰۰ و ۲۰۰۶) در متون شبیه‌سازی مطرح شد، سرچشمه می‌گیرد. هدف الگوریتم AWS تعریف بزرگترین ناحیه همسایگی ممکن برای هر نقطه i در فضای مشاهدات است. این همسایگی باید به گونه‌ای باشد که مدل پارامتری هر ناحیه کاملا منطبق بر داده‌های آن باشد؛ یعنی به وسیله یک مدل با مجموعه‌ای از پارامترهای ثابت به خوبی تقریب زده شوند. طرز کار این الگوریتم مبتنی بر گسترش متوالی ناحیه همسایگی حول نقطه i و تعریف مدل‌های موضعی بر اساس تخصیص وزن، وابسته به نتایج گام قبل، به هر واحد فضایی است. بدین منظور، الگوریتم از یک همسایگی خیلی کوچک برای هر نقطه شروع و در طی فرآیند تکرار تمام همسایگی‌ها با ورود نقاط جدیدی که موجب نقض پارامترهای موضعی نشوند، گسترش می‌یابند.

مقصود رهیافت IGWR بیان این حقیقت است که می‌توان قدرت روش‌های ناپارامتری و سودمندی الگوریتم AWS را ترکیب و به این طریق نواحی همگن را با برازش مدل خاص به

[1] Adaptive weights smoothing

[2] Iterative geographically weighted regression

[3] Endogenous spatial regime models

هر کدام مشخص کرد. به‌طور دقیق در این رهیافت، روش رگرسیون موزون جغرافیایی به‌طور مداوم تکرار و در هر تکرار وزن‌های GWR یعنی درایه‌های قطر اصلی ماتریس W_i به‌روزرسانی و مجددا GWR برازش داده می‌شود. این به‌روزرسانی و اختصاص وزن‌های جدید بر اساس الگوریتم AWS انجام می‌شود؛ یعنی برای هر نقطه i، وزن‌های w_{ij} در جهت گسترش ناحیه همسایگی همگن آن به‌روز می‌شوند. بنابراین به منظور کنترل همگنی در هر مرحله، انجام آزمون به روی ضرایب برآوردشده در (۳۷.۲) الزامی است.

به این ترتیب الگوریتم IGWR دارای مراحل زیر است:

- اختصاص وزن‌های اولیه و برازش مدل GWR
- تکرار مراحل تا برقراری شرط توقف فرآیند
- انجام آزمون همگنی ضرایب
- الگوریتم به‌روزرسانی وزن‌ها

در ادامه هر کدام از مراحل و هدف از انجام آن‌ها به ترتیب تشریح می‌شوند.

۱.۲.۳ اختصاص وزن‌های اولیه و برازش مدل GWR

مرحله اول رهیافت IGWR با برازش مدل GWR آغاز می‌شود. برای این کار، باید ماتریس وزن‌های W_i معرفی شوند. لذا آن‌گونه که در فصل دوم گفته شد، از یک تابع هسته به شکل $w_{ij}^{\circ} = K(d_{ij}; b)$ استفاده می‌کنیم که تابعی از فاصله بین دو مشاهده i, j و مقدار پهنای باند است. همان‌طور که گفته شد $K(\cdot)$ یک تابع هسته دلخواه و w_{ij}° نشان‌دهنده وزن‌های اولیه (تکرار صفر) است. مقدار پهنای باند مناسب به روش AIC گفته‌شده در فصل دوم، محاسبه و در تابع هسته انتخاب‌شده قرار می‌گیرد.

پس از آن‌که وزن‌های اولیه معرفی شدند، ضرایب رگرسیونی و همچنین واریانس جمله خطا، با استفاده از (۳۷.۲) و (۴۰.۲) برآورد می‌شوند و آن‌ها را با نماد $\hat{\boldsymbol{\beta}}_i^{\circ}$ و $\hat{\sigma}_{\varepsilon_i}^{2\circ}$ که نشان‌دهنده برآورد ضرایب GWR و واریانس جمله خطا در تکرار صفر هستند، نشان می‌دهیم.

قابل ذکر است که آندرانو و همکاران (۲۰۱۶) در روش خود پهنای باند بهینه اولیه را اعمال نکردند. در صورتی که تحمیل پهنای باند بهینه به روش AIC برای وزن‌های اولیه، در صورتی که پهنای باند سازوار مورد استفاده باشد، برای همگرا شدن شرط توقف که در بخش بعد به آن می‌پردازیم، لازم است.

۲.۲.۳ شرط توقف الگوریتم تکرار

پس از معرفی وزن‌های اولیه و انجام رگرسیون موزون جغرافیایی، یک ماتریس $n \times (p+1)$ از ضرایب رگرسیونی به شکل

$$\beta = \begin{bmatrix} \beta_{1\circ} & \beta_{11} & \cdots & \beta_{1p} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{i\circ} & \beta_{i1} & \cdots & \beta_{ip} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{n\circ} & \beta_{n1} & \cdots & \beta_{np} \end{bmatrix}$$

داریم. یعنی هر مشاهده i یک بردار $(p+1) \times 1$ از ضرایب رگرسیونی دارد. برای مشخص شدن نواحی همگن یا به عبارتی مشاهداتی که رفتار و ضرایب مدل در آن‌ها مشابه است، باید این بردار از ضرایب دو به دو با یکدیگر مقایسه شوند. در صورت وجود شباهت با به‌روزرسانی وزن‌ها شرط قرارگیری مشاهدات آن‌ها در یک رژیم فراهم می‌شود. پس از به‌روزرسانی وزن‌ها، برازش مدل GWR با ماتریس‌های W_i از وزن‌های جدید تکرار می‌شود. اما این تکرار و به‌روزرسانی وزن‌ها تا کجا ادامه می‌یابد؟

تا زمانی که شرط (۱.۳)، به ازای مقادیر کوچک ω (برای مثال ۰/۰۰۰۱) برقرار شود، فرآیند تکرار ادامه می‌یابد و وزن‌ها در هر تکرار به‌روز می‌شوند:

$$\max|w_{ij}^{(\ell-1)} - w_{ij}^{\ell}| < \omega \qquad \forall ij, i \neq j. \tag{۱.۳}$$

همان‌طور که گفته شد در هر تکرار، برای مثال تکرار ℓام، ضرایب مدل GWR برآوردشده برای نقاط مختلف فضایی دو به دو با یکدیگر مقایسه می‌شوند. یعنی $\boldsymbol{\beta}_i^{\ell}$ با $\boldsymbol{\beta}_j^{\ell}$ به ازای همه $i, j : i \neq j$ مقایسه می‌شوند تا مشخص شود آیا می‌توان مشاهدات آن‌ها را در یک رژیم فضایی قرار داد. مسلما با افزایش اختلاف $\boldsymbol{\beta}_i^{\ell}$ و $\boldsymbol{\beta}_j^{\ell}$ احتمال قرار گیری i, j در یک رژیم کاهش می‌یابد. لذا نیازمند یک آماره آزمون بر مبنای این اختلاف یا فاصله ضرایب هستیم.

۳.۲.۳ آماره آزمون همگنی ضرایب

برای مقایسه ضرایب مدل GWR در هر تکرار، نیازمند یک آماره آزمون هستیم. آماره

$$\chi_{ij}^{\ell} = (\hat{\boldsymbol{\beta}}_i^{\ell} - \hat{\boldsymbol{\beta}}_j^{\ell})^T (\Sigma^{\ell})^{-1} (\hat{\boldsymbol{\beta}}_i^{\ell} - \hat{\boldsymbol{\beta}}_j^{\ell}) \tag{۲.۳}$$

به این منظور معرفی شده است (بیل و همکاران، ۲۰۱۷)؛ که در آن $\hat{\boldsymbol{\beta}}_i^{\ell}$ و $\hat{\boldsymbol{\beta}}_j^{\ell}$ برآورد ضرایب مدل GWR برای دو مشاهده i و j در تکرار ℓام هستند. همچنین Σ^{ℓ}، ماتریس واریانس ادغام‌شده[۴]

[4] Pooled variance matrix

نامیده می‌شود و از میانگین دو ماتریس واریانس ضرایب β_i^ℓ و β_j^ℓ در تکرار ℓام، به صورت زیر به دست می‌آید:

$$\Sigma^\ell = \frac{tr(\mathbf{W}_i^\ell)\Sigma_i^\ell + tr(\mathbf{W}_j^\ell)\Sigma_j^\ell}{tr(\mathbf{W}_i^\ell) + tr(\mathbf{W}_j^\ell)}$$

که در آن ماتریس واریانس $\hat{\boldsymbol{\beta}}_i^\ell$ طبق (۳۹.۲) برای هر تکرار به صورت

$$\Sigma_i^\ell = [(\mathbf{X}^T\mathbf{W}_i^\ell\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_i^\ell][(\mathbf{X}^T\mathbf{W}_i^\ell\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_i^\ell]^T\hat{\sigma}_{\varepsilon_i}^2$$

محاسبه می‌شود.

آماره χ_{ij} یک تابع درجه دوم بر حسب فاصله است. مشخص است مقادیر بزرگ آن نشان‌دهنده اختلاف بین مقادیر $\hat{\boldsymbol{\beta}}_i$ و $\hat{\boldsymbol{\beta}}_j$ است. این اختلاف بین ضرایب رگرسیونی دو موقعیت فضایی i و j، گواه وجود ناهمگنی فضایی است. بنابراین دو موقعیت i و j نمی‌توانند در یک رژیم فضایی قرار بگیرند و بالعکس. یعنی مقادیر کوچک آماره χ_{ij} نشان‌دهنده اختلاف اندک میان ضرایب رگرسیون دو موقعیت فضایی است و بیان‌گر این حقیقت که دو موقعیت i و j می‌توانند در یک رژیم فضایی قرار گرفته و ضرایب یکسان داشته باشند.

با توجه به آن‌چه گفته شد، برای بررسی دقیق برآورد ضرایب و رژیم‌بندی موقعیت‌ها نیازمند آزمون آماری بر مبنای آماره χ_{ij} هستیم، که در بخش بعد معرفی می‌شود.

۴.۲.۳ آزمون همگنی ضرایب

همان‌طور که در فصل دوم گفته شد، هر بردار $\hat{\boldsymbol{\beta}}_i$ یک برآوردگر ضرایب رگرسیونی به روش WLS برای هر مشاهده $i = 1, \cdots, n$، است. این برآوردگر با استفاده از یک زیرنمونه از مشاهدات که توسط پهنای باند و تابع هسته انتخابی مشخص شده‌اند، برآورد را محاسبه می‌کند. اگر تعداد مشاهدات در همه زیرنمونه‌های معرفی‌شده به اندازه کافی بزرگ باشد، آن‌گاه طبق (۳۵.۲) و (۳۷.۲) نتیجه زیر برقرار است (ابراگیموف و مولر، ۲۰۱۰):

$$\hat{\boldsymbol{\beta}}_i^\ell \sim N(\boldsymbol{\beta}_i^\ell, \Sigma_i^\ell).$$

همچنین $\hat{\boldsymbol{\beta}}_i$ و $\hat{\boldsymbol{\beta}}_j$ در هر تکرار مستقل از یکدیگرند (برای اطلاعات بیشتر به ابراگیموف و مولر (۲۰۱۰) مراجعه شود).

بنابراین آماره (۲.۳) برای هر تکرار دارای توزیع مجانبی χ^2 با $p+1$ (برابر با بعد بردار $\hat{\boldsymbol{\beta}}_i$) درجه آزادی است. بدین ترتیب آزمون با فرضیه‌های زیر در هر تکرار انجام می‌گیرد:

$$\begin{aligned} H_0 &: \hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_j \\ H_1 &: \hat{\boldsymbol{\beta}}_i \neq \hat{\boldsymbol{\beta}}_j \end{aligned} \qquad (۳.۳)$$

در این آزمون اگر آماره آزمون مقادیر بزرگ اختیار کند، یعنی $\chi^{\ell}_{ij} > \chi^{۲}_{p+۱}(۱-\alpha)$، آن‌گاه فرضیه صفر در سطح α رد می‌شود. یعنی ضرایب رگرسیونی دو موقعیت i و j یکسان نبوده و این دو موقعیت فضایی به دو رژیم مختلف فضایی تعلق دارند.

آماره χ_{ij} در هر تکرار به‌عنوان معیار برای مجازات عمل می‌کند. به این ترتیب که اگر مقدار آن بزرگ باشد، لازم است که در تکرار بعد به وزن w_{ij} مقدار کوچک‌تری تخصیص داده شود تا دو مشاهده i و j که ناهمگن هستند در دو رژیم مختلف قرار بگیرند. همچنین اگر مقدار χ_{ij} کوچک باشد، آن‌گاه باید وزن آن‌ها در تکرار بعد افزایش یابد. البته برای اطمینان از همگرا بودن وزن‌های نهایی تخصیص‌یافته به مقادیر معقول صفر و یک، باید روش به‌روزرسانی وزن‌ها به آرامی انجام شود. در ادامه بیان می‌کنیم که این به‌روزرسانی (میزان افزایش یا کاهش) وزن‌ها در تکرار بعدی به چه شکل اجرا می‌شود.

۵.۲.۳ الگوریتم به‌روزرسانی وزن‌ها

همان‌طور که در بخش‌های پیش گفته شد، اولین گام در انجام رهیافت IGWR، به‌منظور مشخص کردن ناهمگنی فضایی و در نتیجه آن رژیم‌های فضایی، معرفی وزن‌های اولیه رگرسیون موزون جغرافیایی با استفاده از یک تابع وزن‌دهی و پهنای باند بهینه است. عموماً برای انجام این مهم در مدل GWR (گام اول IGWR)، توابع هسته و معیار AIC به‌کار گرفته می‌شوند. اما در گام‌های بعدی این رهیافت (یعنی تکرار اول به بعد)، وزن‌ها به این طریق محاسبه نمی‌شوند، بلکه مقدار آن‌ها یا به عبارتی به‌روزرسانی آن‌ها نسبت به مرحله قبل با توجه به نتیجه آماره آزمون (۲.۳) صورت می‌گیرد. ایده این روش وزن‌دهی بر پایه الگوریتم هموارسازی وزن‌های سازوار است. همچنین برای اطمینان از همگرا بودن وزن‌های تخصیصی نهایی به مقادیر منطقی ۰ و ۱، باید فرآیند همگرایی وزن‌ها به آهستگی انجام شود. در نتیجه تابع وزن به‌روزرسان w^{ℓ}_{ij} برای هر تکرار ℓ را به‌صورت ضرب تابع هسته $K(d_{ij};b)$ در یک هسته که تابعی از آماره آزمون χ_{ij} در تکرار ℓام است، معرفی می‌کنیم.

بنابراین قرار می‌دهیم

$$w^{\ell}_{ij} = K(d^{\ell}_{ij};b)K(\chi_{ij};\tau), \qquad l = ۱,۲,\cdots \tag{۴.۳}$$

در این تابع $d^{\ell}_{ij} = d_{ij}/\ell$ می‌باشد. این تعریف از فاصله برای هر تکرار به این منظور صورت می‌گیرد تا تضمین کند با جلو بردن فرآیند به‌روزرسانی وزن و تکرارهای مجدد، از دقت برآورد کاسته نشود. در مورد d^{ℓ}_{ij} همچنین می‌توان گفت که با افزایش تکرارها (ℓ)، وزن فاصله d_{ij} کاهش می‌یابد. افزایش وزن‌ها به χ^{ℓ}_{ij} وابسته است.

تابع هسته دوم در (۴.۳) به شکل زیر معرفی می‌شود:

$$K(\chi^{\ell}_{ij};\tau) = \exp\left(-۰/۵(\chi^{\ell}_{ij}\tau)^{۲}\right)$$

که در آن τ پارامتر توزین کردن آماره آزمون χ^{ℓ}_{ij} است. مانند پهنای باند مقادیر نسبتاً بزرگ آن می‌تواند به عملکرد ناپایدار روش منجر شود؛ در حالی که مقادیر کوچک آن حساسیت به تغییر ساختار را کاهش می‌دهد.

پولزهل و اسپوکوینی (۲۰۰۶)، با توجه به این نکته که $\chi^{\ell}_{ij} \sim \chi^{۲}_{(p+۱)}$ ، مقدار τ را به‌صورت زیر معرفی کردند:

$$\tau = \frac{۱}{Q_{\alpha}(\chi^{۲}_{(p+۱)})}$$

که در آن $Q_{\alpha}(\chi_{(p+۱)})$ چندک αام توزیع کی دو با $(p+۱)$ درجه آزادی است. آن‌ها در نهایت وزن‌های معرفی‌شده در (۴.۳) را یکبار دیگر به‌روزرسانی کردند. این به‌روزرسانی با میانگین‌گیری بین وزن‌های حاصل w_{ij} در تکرار ℓام و مقادیر میانگین محاسبه‌شده آن‌ها در تکرار قبلی $\bar{w}^{(\ell-۱)}_{ij}$، به‌صورت زیر حاصل می‌شود:

$$\bar{w}^{\ell}_{ij} = (۱-\eta)\bar{w}^{(\ell-۱)}_{ij} + \eta w^{\ell}_{ij}. \qquad (۵.۳)$$

در این تابع وزن، η پارامتر کنترل است و می‌تواند مقادیر بازه $(۰,۱)$ را اختیار کند. برای مثال اگر $\eta = ۰/۵$ تنظیم شود، آن‌گاه دقیقا میانگین حسابی وزن‌ها در تکرار ℓام و تکرار $(\ell-۱)$ام محاسبه می‌شود. پولزهل و اسپوکوینی η را پارامتر حافظه[۵] نیز نامیده‌اند. با این پارامتر، چگونگی و میزان وزن‌ها در تکرار قبل در تعیین وزن‌ها در تکرار جدید موثر خواهد شد و میزان این تاثیر نیز تحت کنترل خواهد بود. به این ترتیب وزن‌های جدید هموار هستند.

در طی فرآیند همگرایی و مشخص شدن رژیم‌ها، کافی بودن مشاهدات هر گروه بسیار مهم است. اگر مشاهدات یک گروه تعیین‌شده از تعداد پارامترهای مدل کمتر باشد، آن گروه و مشاهدات آن به عنوان داده دورافتاده از محاسبات خارج می‌شوند. در نهایت تعریف مفید برای تشریح همگرایی، تابع تغییرات وزن یعنی

$$d(w) = \max_{i,j} |\bar{w}^{(\ell-۱)}_{ij} - w^{\ell}_{ij}|, \qquad (۶.۳)$$

است. این تابع، $d(w)$، هم‌زمان با تثبیت شدن فرآیند همگرایی و اتمام مرحله اول، به صفر میل می‌کند.

قبل از ورود به مرحله دوم رهیافت و تشریح آن، مطالب گفته‌شده در این بخش، یعنی مراحل فرآیند همگرایی و تعیین رژیم‌ها به روش IGWR را به‌طور خلاصه بیان می‌کنیم.

- تعریف وزن‌های اولیه $\mathbf{W}_i$ توسط یک هسته که تابعی از فاصله هر دو نقطه در فضا و یک مقدار پهنای باند است؛ و استفاده از این وزن‌ها در (۳۷.۲) برای برآورد ضرایب رگرسیون جغرافیایی.

[۵] Memory parameter

- پس از برآورد ضرایب اولیه، فرآیند همگرایی با جایگذاری وزن‌های جدید در قطر اصلی ماتریس $\mathbf{W}_i$ حاصل می‌شود. وزن‌های جدید به واسطه یک فرآیند به‌روزرسانی وزن‌ها با روش AWS محاسبه می‌شوند.

- فرآیند به‌روزرسانی وزن‌ها، یک فرآیند هموارساز است که به تدریج وزن‌های اولیه را به واسطه ترکیب آن‌ها با یک وزن تعریف‌شده بر اساس آماره آزمون χ_{ij}، معادله (۴.۳)، و سپس دوباره به‌روزرسانی وزن حاصل با محاسبه میانگین وزنی آن‌ها و وزن‌های نهایی در تکرار قبل، معادله (۵.۳)، تعدیل می‌کند.

- با قرار دادن $\eta = ۰/۵$ در (۵.۳)، یک ساختار تدریجی برای وزن‌های جدید تضمین می‌شود. به‌روزرسانی وزن‌ها برای استفاده مجدد در (۳۷.۲) تا همگرایی، یعنی زمانی‌که $d(w) \to ۰$، ادامه می‌یابد.

همان‌طور که می‌بینیم، این روش هیچ گونه محدودیتی لحاظ نمی‌کند و کاملا سازگار است. یعنی به اطلاعات پیشین درباره ساختار فضایی داده‌ها نیاز ندارد.

۳.۳ مرحله دوم: مدل‌های اقتصادسنجی با رژیم فضایی درونی

پس از اتمام مرحله اول و مشخص شدن ناهمگنی فضایی تحت رژیم‌های مختلف فضایی، نوبت به قرار دادن آن‌ها در مدل‌های فراموضعی است. مدل‌های حاصل، مدل‌هایی مشابه قبل هستند با این تفاوت که حضور رژیم‌های شناسایی‌شده توسط روش IGWR با ایجاد افرازهایی در متغیر پاسخ، متغیرهای تبیینی و سایر اجزای مدل، موجب برآورد ضرایب رگرسیونی مختلف می‌شود. در این پایان‌نامه، مدل‌های اقتصادسنجی مورد استفاده هستند. مدل‌های پرکاربرد اقتصادسنجی در فصل دوم به‌طور کامل معرفی شدند. از آن میان برای تشریح مرحله دوم رهیافت، تنها به مدل‌های SAC و SDM می‌پردازیم، چراکه همان‌طور که گفته شد اثرات فضایی به‌طور کامل‌تری در آن‌ها لحاظ شده است و همچنین این دو مدل حاصل ترکیب دو مدل SLM و SEM (به دو روش مختلف) هستند. یعنی در حالت خاص $\rho = ۰$، مدل SEM با رژیم فضایی درونی و اگر $\lambda = ۰$، مدل SLM با رژیم فضایی درونی حاصل می‌شود.

۱.۳.۳ مدل ترکیبی فضایی با رژیم درونی (ESR-SAC)

برای شروع، صورت کلی مدل ترکیبی فضایی را یادآوری می‌کنیم. ساختار سلسله‌مراتبی مدل SAC به‌صورت زیر است:

$$\begin{aligned} \mathbf{y} &= \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\upsilon}, \\ \boldsymbol{\upsilon} &= \lambda\mathbf{M}\boldsymbol{\upsilon} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim N(\circ, \sigma^{\mathsf{۲}}\mathbf{I}_n). \end{aligned}$$

در این مدل $\mathbf{y}$، $\boldsymbol{\upsilon}$ و $\boldsymbol{\varepsilon}$، بردارهای با بعد n، $\mathbf{W}$ و $\mathbf{M}$ ماتریس‌های وزن مربعی از بعد n، $\boldsymbol{\beta}$ برداری با بعد $p+\mathsf{۱}$ و $\mathbf{X}$ یک ماتریس $n \times (p+\mathsf{۱})$ بعدی است.

پس از اتمام فرآیند IGWR و مشخص شدن تعداد رژیم‌ها و عضوهای هر کدام، بردارها و ماتریس‌های ذکرشده به شکل‌های زیر افراز می‌شوند:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_{\mathsf{۱}} \\ \vdots \\ \mathbf{y}_c \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{\mathsf{۱}} \\ \vdots \\ \boldsymbol{\beta}_c \end{pmatrix} \quad \boldsymbol{\upsilon} = \begin{pmatrix} \boldsymbol{\upsilon}_{\mathsf{۱}} \\ \vdots \\ \boldsymbol{\upsilon}_c \end{pmatrix} \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathsf{۱}} \\ \vdots \\ \boldsymbol{\varepsilon}_c \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{\mathsf{۱}} & \cdots & \mathbf{o} \\ \vdots & \ddots & \vdots \\ \mathbf{o} & \cdots & \mathbf{X}_c \end{pmatrix} \tag{۷.۳}$$

به‌عنوان مثال $\mathbf{y}_j$، $j = \mathsf{۱}, \ldots, c$ بردار مقادیر متغیر پاسخ و $\mathbf{X}_j$ ماتریس مقادیر متغیرهای تبیینی در رژیم jام هستند که بعد آن‌ها به ترتیب $(n_j) \times \mathsf{۱}$ و $(n_j) \times (p+\mathsf{۱})$ است.

ملاحظه ۱.۳.۳. توجه داشته باشید که $\sum_{j=\mathsf{۱}}^{c} n_j = n$ و بعد ماتریس $\mathbf{X}$، $n \times (c(p+\mathsf{۱}))$ است.

بنابراین مدل SAC پس از درونی کردن رژیم‌های فضایی به شکل زیر خواهد بود:

$$\begin{aligned} \begin{pmatrix} \mathbf{y}_{\mathsf{۱}} \\ \vdots \\ \mathbf{y}_c \end{pmatrix} &= \rho\mathbf{W} \begin{pmatrix} \mathbf{y}_{\mathsf{۱}} \\ \vdots \\ \mathbf{y}_c \end{pmatrix} + \begin{pmatrix} \mathbf{X}_{\mathsf{۱}} & \cdots & \mathbf{o} \\ \vdots & \ddots & \vdots \\ \mathbf{o} & \cdots & \mathbf{X}_c \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{\mathsf{۱}} \\ \vdots \\ \boldsymbol{\beta}_c \end{pmatrix} + \begin{pmatrix} \boldsymbol{\upsilon}_{\mathsf{۱}} \\ \vdots \\ \boldsymbol{\upsilon}_c \end{pmatrix} \\ \begin{pmatrix} \boldsymbol{\upsilon}_{\mathsf{۱}} \\ \vdots \\ \boldsymbol{\upsilon}_c \end{pmatrix} &= \lambda\mathbf{M} \begin{pmatrix} \boldsymbol{\upsilon}_{\mathsf{۱}} \\ \vdots \\ \boldsymbol{\upsilon}_c \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathsf{۱}} \\ \vdots \\ \boldsymbol{\varepsilon}_c \end{pmatrix} \end{aligned} \tag{۸.۳}$$

که در آن $i = \mathsf{۱}, \mathsf{۲}, \cdots, c$، $\boldsymbol{\varepsilon}_j \sim N\left(\circ, \sigma^{\mathsf{۲}}_{\varepsilon_j}\mathbf{I}_{n_j}\right)$.

۲.۳.۳ مدل دوربین فضایی با رژیم درونی (ESR-SDM)

شکل افرازشده جملات معادله (۳۰.۲) برای مدل SDM به‌صورت

$$\mathbf{y}=\begin{pmatrix}\mathbf{y}_1\\ \vdots\\ \mathbf{y}_c\end{pmatrix}\quad \boldsymbol{\beta}=\begin{pmatrix}\boldsymbol{\beta}_1\\ \vdots\\ \boldsymbol{\beta}_c\end{pmatrix}\quad \boldsymbol{\alpha}=\begin{pmatrix}\boldsymbol{\alpha}_1\\ \vdots\\ \boldsymbol{\alpha}_c\end{pmatrix}\varepsilon=\begin{pmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_c\end{pmatrix}$$

$$\mathbf{X}=\begin{pmatrix}\mathbf{X}_1 & \cdots & \mathbf{o}\\ \vdots & \ddots & \vdots\\ \mathbf{o} & \cdots & \mathbf{X}_c\end{pmatrix}\quad \mathbf{X}_v=\begin{pmatrix}\mathbf{X}_{v_1} & \cdots & \mathbf{o}\\ \vdots & \ddots & \vdots\\ \mathbf{o} & \cdots & \mathbf{X}_{v_c}\end{pmatrix} \tag{۹.۳}$$

است. مشابه معادله (۲۹.۲)، $\mathbf{X}_v$ ماتریسی مشابه $\mathbf{X}$ است با این تفاوت که عرض از مبدأ از آن حذف شده است.

نکات گفته‌شده در مدل ESR-SAC، در این مدل نیز برقرار هستند. به‌علاوه بردار $\boldsymbol{\alpha}$ از بعد cp و ماتریس $\mathbf{X}_v$ دارای بعد $n\times cp$ هستند.

مدل دوربین با رژیم فضایی درونی‌شده، به‌صورت زیر نوشته می‌شود:

$$\begin{aligned}\begin{pmatrix}\mathbf{y}_1\\ \vdots\\ \mathbf{y}_c\end{pmatrix}=&\rho\mathbf{W}\begin{pmatrix}\mathbf{y}_1\\ \vdots\\ \mathbf{y}_c\end{pmatrix}+\begin{pmatrix}\mathbf{X}_1 & \cdots & \mathbf{o}\\ \vdots & \ddots & \vdots\\ \mathbf{o} & \cdots & \mathbf{X}_c\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_1\\ \vdots\\ \boldsymbol{\beta}_c\end{pmatrix}\\ &+\mathbf{M}\begin{pmatrix}\mathbf{X}_{v_1} & \cdots & \mathbf{o}\\ \vdots & \ddots & \vdots\\ \mathbf{o} & \cdots & \mathbf{X}_{v_c}\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}_1\\ \vdots\\ \boldsymbol{\alpha}_c\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_c\end{pmatrix}\end{aligned} \tag{۱۰.۳}$$

که در آن $j=1,2,\cdots,c$ $\varepsilon_j\sim N\left(0,\sigma^2_{\varepsilon_j}\mathbf{I}_{n_j}\right)$.

ماتریس‌های وزن $\mathbf{W}$ و $\mathbf{M}$ را می‌توان در مدل‌های اقتصادسنجی با رژیم فضایی درونی، همانند مدل‌های متناظر فراموضعی خود، برابر درنظر گرفت. برآورد پارامترهای این مدل‌ها نیز مشابه حالت فراموضعی خود محاسبه می‌شود.

فصل ۴

ارزیابی مدل پیشنهادی

۱.۴ مقدمه

هدف از این فصل آشکار کردن حضور ناهمگنی فضایی در داده‌های اقتصادی است. پس از این نمایش، نوبت به اثبات کارایی رهیافت IGWR، در مدل‌بندی این ناهمگنی به شکل رژیم‌های مختلف فضایی درون مدل‌های فراموضعی اتورگرسیو است. به این منظور از دو مجموعه داده واقعی با نام‌های بالتیمور[۱] و مسکن[۲] استفاده می‌کنیم. هر دوی این مجموعه داده‌ها در بسته‌افزار spdep[۳] (بیوند و همکاران، ۲۰۱۵) در نرم‌افزار R (گروه اصلی R، ۲۰۱۸) موجود و قابل دسترس هستند.

۲.۴ مثال کاربردی اول

مجموعه داده بالتیمور شامل اطلاعات قیمت فروش مسکن و برخی خصوصیات آن‌ها نظیر متراژ منزل، تعداد پارکینگ‌ها، واقع بودن در بخش شهری یا حوالی آن و غیره است. متغیرهای موجود در این مجموعه در جدول ۱.۴ فهرست شده‌اند. این اطلاعات مربوط به ۲۱۱ مسکن

[1] baltimore
[2] house
[3] Spatial dependence

در کلان شهر بالتیمور واقع در ایالت ماری‌لند کشور امریکا است. شکل ۱.۴ موقعیت کلان شهر بالتیمور را در ایالت ماری‌لند و ایالات متحده امریکا نمایش می‌دهد. این مجموعه داده اولین بار توسط دوبین (۱۹۹۲) مورد بررسی قرار گرفت و برای اولین بار مشکل پارامترهای مختلف در فضا از اهمیت برخوردار شد.

(آ) موقعیت در ایالت ماری‌لند (ب) موقعیت در ایالات متحده امریکا

شکل ۱.۴: موقعیت شهر بالتیمور

جدول ۱.۴: متغیرهای مجموعه داده بالتیمور

	نام متغیر	توضیحات
مجموعه داده بالتیمور	price	قیمت خانه_ متغیر پاسخ
	dwell	متغیر ظاهری_ ۱ برای خانه‌های تک خانوار
	nbath	تعداد حمام‌ها
	patio	متغیر ظاهری_ ۱ برای خانه‌های دارای پاسیو
	firepl	متغیر ظاهری_ ۱ اگر شومینه موجود است
	ac	متغیر ظاهری_ ۱ اگر تهویه هوا موجود است
	bment	تعداد زیرزمین‌ها
	gar	تعداد گاراژها
	citcou	متغیر ظاهری_ ۱ برای خانه‌هایی که در بخش بالتیمور واقع هستند
	lotsz	متراژ مسکن به هزار فوت‌مربع*

* هر فوت‌مربع برابر ۰/۰۹۲۹۰۳۰۴ مترمربع

۱.۲.۴ انجام آزمون‌های آماری و برازش مدل‌های فراموضعی فضایی

به منظور بررسی وجود خودهمبستگی فضایی در داده‌ها از آزمون‌های معرفی‌شده در فصل اول استفاده می‌کنیم. توابع این آزمون‌ها در بسته‌افزار spdep موجود هستند.

جدول ۲.۴: نتایج آزمون‌های خودهمبستگی فضایی برای داده‌های بالتیمور

آزمون‌های خودهمبستگی فضایی	مقدار آماره	p ـ مقدار
موران I	۰/۰۸۷۲	۰/۰۰۰۱
نسبت درستنمایی	۱۰/۵۷	۰/۰۰۱۱
ضریب لاگرانژ خطا	۸/۷۵۶۳	۰/۰۰۳
ضریب لاگرانژ تأخیر	۲۰/۷۱۹	$< ۰/۰۰۰$

نتایج آزمون‌های خودهمبستگی نشان‌دهنده وجود خودهمبستگی فضایی در داده‌های بالتیمور است. آزمون ضریب لاگرانژ تأخیر فضایی با قدرت بیشتری، نسبت به سایر آزمون‌ها، فرضیه صفر خود را رد کرده است. یعنی با احتمال بیشتری خودهمبستگی فضایی، در متغیرهای پاسخ داده‌های بالتیمور وجود دارد. با توجه به این نتایج، برآورد به روش OLS برای این مجموعه داده مناسب نبوده و باید از مدل‌های اتورگرسیو فضایی برای برازش داده‌ها استفاده کرد.

برای این داده‌ها از چهار مدل اتورگرسیو خطای فضایی، تاخیر فضایی، دوربین فضایی و ترکیبی فضایی استفاده کردیم. در این مدل‌ها وزن‌ها $(\mathbf{W})$ به روش k ـ نزدیک‌ترین همسایگی با $k = ۱۰$ محاسبه شدند. همچنین در مدل‌های SAC و SDM، $\mathbf{W} = \mathbf{M}$ در نظر گرفته‌شد. برای برازش مدل از توابع موجود در بسته‌افزار spdep در نرم‌افزار R استفاده کردیم. نتایج این مدل‌ها در جدول ۳.۴ گزارش شده‌اند.

جدول ۳.۴: برآورد پارامترهای مدل‌های فراموضعی برای مجموعه داده بالتیمور

ضرایب	SLM		SEM		SAC		SDM	
عرض از مبدأ	−۶/۱۷۳	.	۳/۶۷۴		−۶/۱۲۵	.	۲۲/۹۲	*
dwell	۷/۳۳۶	***	۹/۲۲۳	***	۷/۵۴۷	***	۸/۴۲۱	***
nbath	۶/۸۳۹	***	۷/۸۵۴	***	۶/۹۶۰	***	۶/۲۲۶	***
patio	۸/۴۴۳	***	۷/۶۶۱	**	۸/۳۴۳	***	۶/۹۸۵	**
firepl	۹/۷۴۰	***	۸/۸۶۸	***	۹/۵۹۲	***	۹/۰۲۴	***
ac	۶/۹۷۰	**	۷/۹۷۴	***	۷/۰۶۱	**	۸/۷۰۴	***
bment	۳/۳۸۰	***	۳/۴۶۹	***	۳/۴۰۵	**	۳/۰۱۱	***
gar	۵/۴۵۶	***	۴/۸۶۶	**	۵/۴۰۸	***	۵/۰۴۶	***
citcou	۹/۴۲۶	***	۱۳/۱۳۰	***	۹/۶۰۹	***	۱۱/۶۰۳	***
lotsz	۰/۰۳۴	*	۰/۰۳۷	*	۰/۰۳۴	*	۰/۰۳۸	**
M-dwell							−۱۳/۶۴۵۱	.
M-nbath							−۱۰/۵۲۲۴	*
M-patio							۱۲/۷۶۵۵	.
M-firepl							۲۱/۶۵۳۳	**
M-ac							۵/۴۱۵۵	
M-bment							−۴/۵۸۰۳	
M-gar							۶/۰۱۶۱	
M-citcou							−۳/۷۷۵۱	
M-lotsz							۰/۱۱۸۳	.
ρ	۰/۳۳۴	***	–		۰/۳۲۴	***	۰/۰۶۲	
λ	–		۰/۴۹۰	**	۰/۰۵۰		–	
AIC	۱۶۶۴/۱۴		۱۶۷۲/۵۷		۱۶۶۶/۰۹۲		۱۶۵۱/۶۰۸	

معنی دار بودن: ۱/۰ < . ۰/۰۵ < * ۰/۰۱ < ** ۰/۰۰۱ < ***

بنا به نتایج جدول ۳.۴ در همه مدل‌های فراموضعی در سطح معنی‌داری ۰/۰۵، تمامی متغیرهای تبیینی معرفی‌شده در جدول ۱.۴ دارای اثرات معنی‌دار و مثبت در تعیین قیمت مسکن شهر بالتیمور هستند. در مدل SDM علاوه‌بر این متغیرها، متغیرهای تبیینی تأخیر فضایی نیز حضور دارند که تعدادی از آن‌ها دارای اثر منفی معنی‌دار هستند. برای مثال، متغیر تعداد حمام‌ها با تأخیر فضایی دارای اثر منفی معنی‌دار در سطح ۰/۰۵ است. یعنی تعداد زیاد حمام‌های مسکن i موجب افزایش قیمت فروش این مسکن اما کاهش قیمت فروش مسکن‌های مجاور آن می‌شود.

ضریب خودهمبستگی ρ در مدل SDM با ورود متغیرهای تبیینی تأخیر فضایی نسبت به مدل SLM به‌طور چشمگیری (از ۰/۳۳۴ به ۰/۰۶۲) کاهش یافته است و این بدان معناست که بی‌توجهی به همبستگی فضایی متغیرهای تبیینی و حذف اثرات آن‌ها در مدل تاخیر

فضایی موجب ایجاد خودهمبستگی ظاهری و کاذب در مدل شده است. به عبارتی ناهمگنی حاصل از حذف این متغیرها (جنبه اول ـ واریانس نامتجانس) در مدل SLM به اشتباه تحت خودهمبستگی قوی وارد مدل‌بندی شده است.

برای انتخاب بهترین مدل از چهار مدل برازش داده‌شده، از معیار اطلاع آکاییک استفاده شد. این معیار که به صورت

$$\mathrm{AIC} = ۲k - ۲\ell \qquad (۱.۴)$$

تعریف شده است، اندازه‌ای از میزان نیکویی برازش یک مدل ارائه می‌دهد. در معادله (۱.۴)، k تعداد پارامترهای مدل و ℓ لگاریتم تابع درستنمایی ماکسیمم است. برای مجموعه‌ای از مدل‌ها، مدلی که کمترین مقدار AIC را داشته باشد، بهتر است. همان‌طور که از نتایج آزمون‌های خودهمبستگی نتیجه گرفتیم، خودهمبستگی فضایی در متغیر پاسخ وجود دارد. با مقایسه نتایج سطر AIC جدول ۳.۴ خواهیم دید که مدل‌هایی که خودهمبستگی را به این شکل لحاظ کرده‌اند، مدل‌های بهتری هستند (مقایسه SEM با SLM و SDM).

با توجه به نتایج سطر AIC در جدول ۳.۴، مدل SDM دارای کمترین AIC است. اما باید توجه داشت که ضریب خودهمبستگی (ρ) در این مدل معنی‌دار نیست، همان‌طور که گفته شد بی‌توجهی به تاثیر متغییرهای تأخیر تبیینی موجب القای حضور خودهمبستگی فضایی متغیرهای پاسخ شده بود، لذا مدل مناسب برای داده‌های بالتیمور مدل رگرسیون خطی (۲.۱) با ۱۹ متغیر تبیینی ردیف شده در جدول ۳.۴، است. مقدار AIC برای این مدل ۱۶۴۹/۸ محاسبه شد که گواه این ادعا است.

اما همان‌طور که در فصل دوم گفته شد در این مدل‌ها به امکان وجود ناهمگنی فضایی حاصل از پارامترهای مختلف در فضا توجه نشده است و ناهمگنی همچون خودهمبستگی فضایی به‌طور مستقیم وارد مدل نشده‌است.

۲.۲.۴ ناهمگنی در داده‌ها

برای آزمون ناهمگنی در حضور خودهمبستگی فضایی در مدل‌های فراموضعی فضایی از آزمون بریچ ـ پیگان (۱۹.۱) استفاده کردیم. نتایج در همه مدل‌ها حاکی از وجود ناهمگنی در مقادیر خطا است.

پس از تایید وجود ناهمگنی فضا، به منظور نمایش ناهمگنی فضایی، رگرسیون موزون جغرافیایی را به داده‌ها برازش دادیم. با این کار ناهمگنی فضایی خود را در قالب اختلاف بارز در ضرایب مشاهدات، نشان می‌دهد. اما همان‌طور که گفته شد نتایج قابل تحلیل و آزمون نیستند.

جدول ۴.۴: نتایج آزمون‌های ناهمگنی فضایی دادهای بالتیمور

مدل‌های فضایی	آماره بریچ_پیگان	درجه آزادی	p_مقدار
SLM	۳۹/۷۲۸	۹	< ۰/۰۰۰۱
SEM	۳۳/۲۵۶	۹	۰/۰۰۰۱
SAC	۳۹/۳۰۶	۹	< ۰/۰۰۰۱
SDM	۴۵/۸۲۷	۱۸	۰/۰۰۰۳

رگرسیون موزون جغرافیایی با تابع gwr.basic در بسته‌افزار GWmodel (گلینی و همکاران، ۲۰۱۳) در نرم‌افزار R در دسترس است و نتایج برآورد پارامترهای مختلف نقاط جغرافیایی در جدول ۵.۴ نمایش داده شده‌اند. به‌عنوان مثال، ضرایب رگرسیونی متغیر DWELL برای ۲۱۱ خانه از مقدار ۱/۹۰ تا ۱۳/۹۰ تغییر می‌کند که گواه ناهمگنی فضایی به شکل پارامترهای مختلف در فضا (جنبه دوم ناهمگنی) در داده‌های بالتیمور است.

جدول ۵.۴: نتایج برازش مدل GWR برای داده‌های بالتیمور

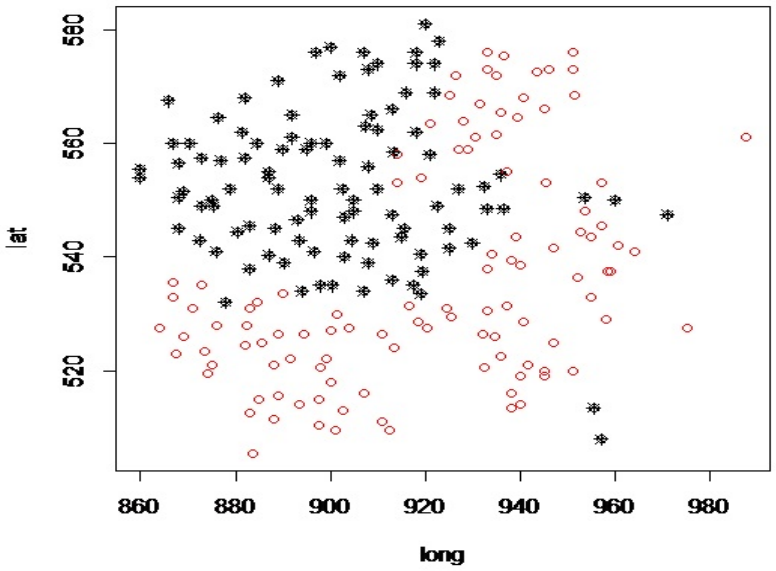
ضرایب	مینیمم	چارک اول	میانه	چارک سوم	ماکسیمم
Intercept	−۴/۸۹۲۸۱۵۳	−۰/۱۵۲۶۲۷۲	۳/۴۵۸۴۲۸۸	۵/۸۵۸۶۱۱۰	۸/۲۵۹۳
DWELL	۱/۹۰۲۶۲۲۳	۳/۸۸۴۴۰۴۶	۸/۰۵۳۱۴۷۲	۱۲/۵۰۲۷۶۰۳	۱۳/۹۰۵۳
NBATH	۴/۶۵۶۸۵۶۲	۵/۷۰۲۰۱۱۵	۶/۸۶۹۵۶۵۳	۷/۷۱۱۵۰۰۲	۱۰/۰۳۸۵
PATIO	۲/۵۲۵۰۷۶۷	۶/۰۶۱۷۷۰۳	۸/۴۲۱۳۵۵۳	۱۰/۶۴۷۸۳۴۰	۱۲/۵۷۴۷
FIREPL	۵/۰۲۷۶۳۳۰	۱۰/۲۶۸۵۰۵۱	۱۲/۶۴۳۴۰۳۷	۱۴/۲۶۹۹۶۱۶	۱۷/۰۲۶۹
AC	۴/۶۵۱۰۴۵۵	۶/۳۶۸۳۵۷۴	۷/۲۸۱۵۷۴۴	۸/۱۳۶۹۳۱۴	۱۱/۱۴۲۱
BMENT	۲/۲۷۵۱۷۴۱	۳/۱۰۸۴۲۲۷	۳/۸۶۸۸۹۰۶	۵/۱۰۰۱۱۸۲	۶/۵۰۰۲
GAR	۱/۴۷۸۱۴۰۲	۲/۳۷۲۷۲۴۴	۳/۷۵۶۳۵۸۲	۵/۶۸۹۰۱۳۵	۱۲/۲۹۵۱
CITCOU	۸/۸۵۱۰۲۵۶	۱۰/۸۶۴۰۰۱۱	۱۳/۸۰۷۱۰۷۰	۱۴/۸۵۲۹۶۸۰	۱۷/۱۳۲۶
LOTSZ	−۰/۰۵۱۵۰۱۴	۰/۰۰۱۱۸۲۶	۰/۰۴۴۷۰۵۳	۰/۰۷۶۴۸۷۷	۰/۰۹۲۳

۳.۲.۴ برازش مدل پیشنهادی

بنابر نتایج گفته‌شده، به دلیل حضور هم‌زمان ناهمگنی و خودهمبستگی فضایی مدل OLS و مدل‌های فراموضعی فضایی برای داده‌های مورد مطالعه مناسب نیستند؛ لذا از رهیافت پیشنهادی خود که ناهمگنی فضایی را در قالب رژیم‌های مختلف فضایی درون مدل‌های خودهمبستگی فضایی لحاظ می‌کند، استفاده می‌کنیم و نتایج را با مدل‌های فراموضعی مقایسه می‌کنیم.

نتایج رهیافت IGWR

در این مثال به منظور تعیین وزن‌های اولیه رگرسیون وزنی جغرافیایی از تابع هسته گاوسی با پهنای باند سازوار بر اساس معیار AIC در رابطه (۴۳.۲) استفاده کردیم. مقدار پهنای باند بهینه برای این مجموعه داده ، $b_{Knn}^{AIC} = ۳۳$ نتیجه شد. الگوریتم به‌روزرسانی وزن‌ها با $\omega = ۰/۰۰۰۱$، پارامتر توزین $\tau = ۰/۰۰۱$ و پارامتر حافظه $\eta = ۰/۵$، پس از ۹۶ تکرار به همگرایی رسید و ۲۱۱ مشاهده در $c = ۲$ رژیم همگن با زیرنمونه‌های $n_۱ = ۱۰۱$ و $n_۲ = ۱۱۰$ دسته‌بندی شدند. این دو رژیم فضایی در شکل ۲.۴ با دو نماد متفاوت $*$ و $\circ$ نشان داده شده‌اند. همان‌طور که در شکل ۲.۴ مشخص است، الگوریتم IGWR نواحی مرکزی و شمال‌غربی بالتیمور را از نواحی جنوبی و شمال‌شرقی آن جدا کرده است.



شکل ۲.۴: رژیم‌های فضایی مجموعه داده بالتیمور، نقاط $*$ رژیم فضایی اول و نقاط $\circ$ رژیم فضایی دوم را نشان می‌دهند

برازش مدل‌ها با رژیم فضایی درونی

پس از مشخص شدن رژیم‌های فضایی، داده‌ها را به صورت‌های (۷.۳) و (۹.۳) افراز کرده و مدل‌های (۸.۳) و (۱۰.۳) را به آن‌ها برازش دادیم. مدل (۸.۳) را در حالت‌های خاص $\rho = ۰$ و $\lambda = ۰$ نیز در نظر گرفتیم و به ترتیب مدل‌های خطای فضایی و تأخیر فضایی با رژیم درونی را نیز برازش دادیم. در این مدل‌ها وزن $\mathbf{W}$ همانند مدل‌های فراموضعی به روش ۱۰ـ نزدیکترین همسایگی محاسبه شدند؛ با این تفاوت که رژیم‌های مشخص‌شده در آن لحاظ شدند. یعنی

بردارهای حاوی اطلاعات مختصات داده‌ها نیز افراز و سپس عملیات وزن‌دهی انجام شد. نتایج در جدول ۶.۴ گزارش شده‌اند. مشابه نتایج قبل، *** برای p ـ مقدار کمتر از ۰/۰۰۱، ** برای p ـ مقدار کمتر از ۰/۰۱، * و . به ترتیب برای معنی‌دار بودن در سطح ۰/۰۵ و ۰/۱ استفاده شده‌اند.

جدول ۶.۴: برآورد پارامترهای مدل‌های با رژیم فضایی درونی برای مجموعه داده بالتیمور

ضرایب	SLM		SEM		SAC		SDM	
عرض از مبدأ ۱	−۱۶/۴۰۹	***	−۵/۲۷۴		−۱۶/۶۷۰	***	−۵/۳۶۷	
عرض از مبدأ ۲	۰/۲۰۶		۱۱/۳۷۴	*	−۰/۶۰۲		۹/۳۰۳	
dwell 1	۶/۰۰۴	*	۸/۶۳۵	**	۵/۲۶۹	.	۷/۸۱۵	**
dwell 2	۷/۵۷۰	**	۷/۸۶۴	**	۷/۴۴۰	**	۸/۷۸۱	**
nbath 1	۵/۴۰۶	**	۵/۵۴۶	**	۵/۳۳۰		۴/۳۲۹	**
nbath 2	۸/۵۲۷	***	۸/۴۱۳	**	۸/۶۳۵	**	۷/۹۶۵	**
patio 1	۱۰/۳۲۸	***	۸/۹۲۴	**	۱۰/۶۹۶	***	۷/۸۵۰	**
patio 2	۱/۰۱۰		۲/۷۴۴		۰/۳۹۲		۰/۷۷۷	
firepl 1	۱۲/۴۲۰	***	۱۲/۲۰۳	***	۱۲/۷۳۳	***	۱۲/۱۴۱	***
firepl 2	۶/۰۸۷	.	۴/۹۳۴		۶/۴۴۱	*	۷/۶۳۲	*
ac 1	۷/۶۲۰	**	۷/۹۹۹	**	۷/۵۱۷	**	۱۰/۰۶۳	***
ac 2	۴/۸۰۰	.	۵/۷۹۴	*	۴/۴۷۸		۳/۵۲۶	
bment 1	۷/۸۸۵	***	۷/۸۵۳	***	۷/۶۹۹	***	۷/۴۰۱	***
bment 2	۱/۴۰۳		۱/۵۰۳		۱/۳۹۰		۱/۸۱۴	.
gar 1	۷/۳۴۲	***	۶/۷۸۲	***	۷/۴۰۲	***	۷/۰۸۵	***
gar 2	−۰/۰۸۵		−۰/۴۷۹		۰/۰۹۰		۰/۲۴۱	
citcou 1	۱۴/۱۳۴	***	۲۰/۲۵۲	***	۱۳/۱۴۹	***	۲۰/۰۴۶	***
citcou 2	۶/۲۱۱	*	۸/۸۱۰	**	۵/۹۵۴	*	۸/۰۵۳	.
lotsz 1	۰/۰۳۳	.	۰/۰۳۳	.	۰/۰۳۴	.	۰/۰۳۰	
lotsz 2	۰/۰۳۶		۰/۰۴۱	.	۰/۰۳۵		۰/۰۳۲	
M-dwell 1							−۲۰/۳۲۸	.
M-dwell 2							۸/۸۵۸	
M-nbath 1							۲/۹۱۹	
M-nbath 2							−۵/۳۸۷	
M-patio 1							۱۳/۷۵۹	
M-patio 2							−۳/۰۹۰	

ادامه جدول

جدول ۶.۴ – *ادامه جدول*

ضرایب	SLM		SEM		SAC		SDM
M-firepl 1							۲۲/۱۰۵ .
M-firepl 2							۱۴/۹۰۵
M-ac 1							۱۵/۷۳۰
M-ac 2							−۱/۷۷۷
M-bment 1							−۴/۲۸۹
M-bment 2							۳/۷۷۳
M-gar 1							۷/۵۳۲
M-gar 2							۹/۱۹۱
M-citcou 1							−۶/۴۵۷
M-citcou 2							−۴/۲۸۳
M-lotsz 1							۰/۱۱۹
M-lotsz 2							۰/۰۶۷
ρ	۰/۳۲۷	***	–	***	۰/۳۵۴	***	−۰/۱۱۰
λ	–		۰/۴۹۴	**	−۰/۱۳۹		–
AIC	۱۶۴۶/۴۸۶		۱۶۵۵/۱۶۶		۱۶۴۸/۳۰۰		۱۶۴۵/۳

در خصوص این نتایج می‌توان بیان کرد که

۱. در مدل‌های فراموضعی همه متغیرهای تبیینی دارای اثر معنی‌دار بر قیمت مسکن بودند؛ اما پس از اعمال رژیم‌های فضایی این اثرات تغییر کرده‌اند. اغلب متغیرهای تبیینی فقط در رژیم اول دارای اثر معنی‌دار هستند و در رژیم دوم هیچ تأثیری بر تعیین قیمت مسکن ندارند. این امر می‌توان ناشی از این حقیقت باشد که بخش مرکزی و شمال‌غربی شهر بالتیمور، منطقه‌ای مرفه‌نشین است و متغیرهای تجملی نظیر تعدد گاراژ و داشتن پاسیو برای افراد ساکن در آن حائز اهمیت است و موجب افزایش قیمت مسکن می‌شود.

۲. مقدار AIC در همه موارد در ESRMs، یعنی پس از اعمال رژیم‌های فضایی و مدل‌بندی ناهمگنی، کوچک‌تر شده است. یعنی مدل‌های اقتصادسنجی فضایی با اعمال رژیم‌های فضایی در مقایسه با مدل‌های مشابه فراموضعی خود، مدل‌های بهتری هستند.

۳. همچنین در این مثال در همه مدل‌ها پس از ورود رژیم‌های فضایی، ضرایب خودهمبستگی فضایی تغییر چندانی نکرده‌اند؛ یعنی اثرات خودهمبستگی واقعی بوده و تحت تاثیر ناهمگنی ضرایب بزرگ‌تر برآورد نشده‌اند.

۴. با مقایسه دو مدل SLM و SDM مشاهده می‌کنیم که ضریب خودهمبستگی فضایی با ورود متغیرهای تبیینی تأخیر فضایی به‌طور چشمگیری کاهش یافته است. یعنی مشابه آن‌چه در مورد مدل‌های فراموضعی گفته‌شد، بیشتر اثرات خودهمبستگی فضایی توسط متغیرهای تبیینی حذف‌شده ناشی از عدم توجه به همبستگی فضایی متغیرهای تبیینی به SLM القا شده است.

۵. همچنین علامت اثر خودهمبستگی نیز تغییر کرده است؛ بدون درنظر گرفتن رژیم‌ها در مدل SDM خودهمبستگی مثبت ۰/۳۲۷ بین قیمت مسکن‌های شهر باتیمور برآورد شده بود، در حالی‌که این مقدار پس از اعمال رژیم‌ها به ۰/۱۱۰− رسید. اما همان‌طور که در جدول ۶.۴ گزارش شد، این اثر معنی‌دار نیست.

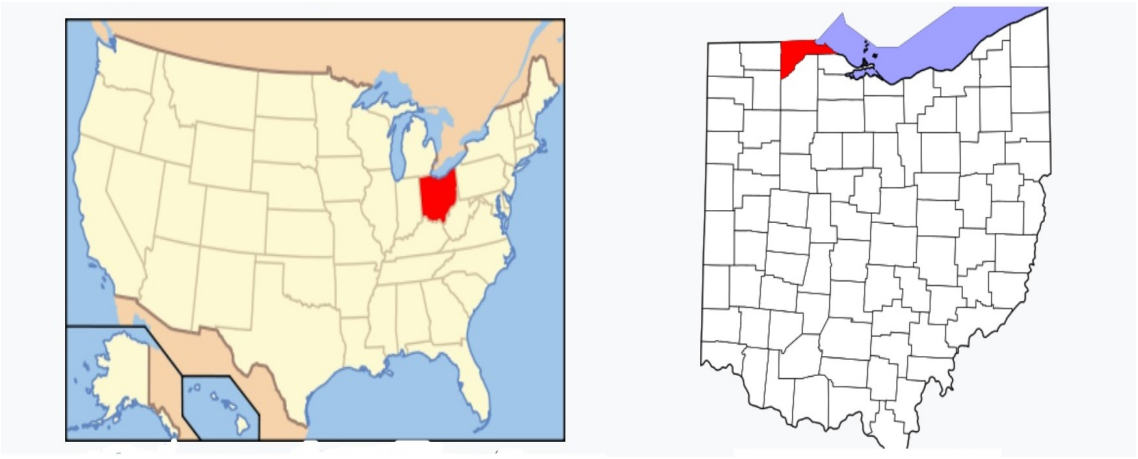
بنابر آنچه گفته‌شد، بهترین مدل برای داده‌های مسکن شهر بالتیمور مدل رگرسیون خطی با حضور متغیرهای تأخیر فضایی و اعمال رژیم‌های فضایی مشخص‌شده، است؛ یعنی مدلی به‌صورت

$$\begin{pmatrix}\mathbf{y}_1\\ \vdots\\ \mathbf{y}_c\end{pmatrix}=\begin{pmatrix}\mathbf{X}_1 & \cdots & \mathbf{o}\\ \vdots & \ddots & \vdots\\ \mathbf{o} & \cdots & \mathbf{X}_c\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_1\\ \vdots\\ \boldsymbol{\beta}_c\end{pmatrix}+\mathbf{M}\begin{pmatrix}\mathbf{X}_{v_1} & \cdots & \mathbf{o}\\ \vdots & \ddots & \vdots\\ \mathbf{o} & \cdots & \mathbf{X}_{v_c}\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}_1\\ \vdots\\ \boldsymbol{\alpha}_c\end{pmatrix}+\begin{pmatrix}\boldsymbol{\varepsilon}_1\\ \vdots\\ \boldsymbol{\varepsilon}_c\end{pmatrix}$$

که در آن $\varepsilon_j \sim N\left(\mathbf{0}, \sigma^2_{\varepsilon_j}\mathbf{I}_{n_j}\right)$ و با ۳۸ متغیر تبیینی که در جدول ۶.۴ ردیف شده‌اند. مقدار AIC این مدل ۱۶۴۳/۷ است. برای درک بهتر این مدل را با مدل (۱۰.۳) مقایسه کنید.

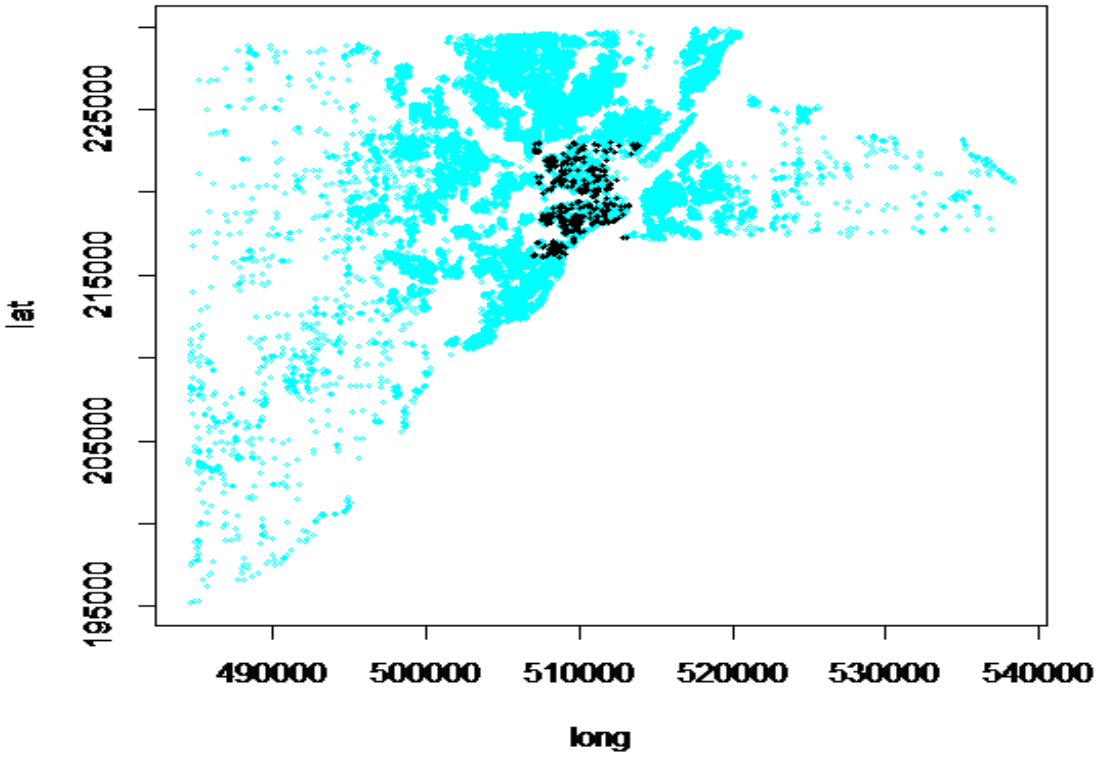
۳.۴ مثال کاربردی دوم

مجموعه داده مسکن نیز شامل اطلاعات قیمت فروش مسکن در بازه سال‌های ۱۹۹۳ـ۱۹۹۸ برای ۲۵۳۵۷ خانه در شهرستان لوکاس ایالت اوهایو کشور امریکا است. شکل ۳.۴ موقعیت ایالت اوهایو و شهرستان لوکاس را نشان می‌دهد. بیل و همکاران (۲۰۱۷) به دلیل هزینه محاسبات الگوریتم IGWR تنها از خانه‌های بخش مرکزی لوکاس که در سال ۱۹۹۳ به فروش رفته‌اند (۳۸۲ مورد)، استفاده کردند. در این پایان‌نامه نیز پس از بررسی‌های لازم و مشخص کردن طول و عرض جغرافیایی بخش مرکزی شهرستان لوکاس، از خانه‌های واقع در آن منطقه برای تحلیل استفاده کردیم . شکل ۴.۴، کل مجموعه داده مسکن و داده‌های انتخابی (لوکاس) را نمایش می‌دهد. متغیرهای این مجموعه داده در جدول ۷.۴ معرفی شده‌اند.



(آ) موقعیت در ایالت اوهایو (ب) موقعیت ایالت اوهایو در ایالات متحده امریکا

شکل ۳.۴: موقعیت شهرستان لوکاس



شکل ۴.۴: داده‌های انتخاب شده از مجموعه داده مسکن

جدول ۷.۴: متغیرهای داده لوکاس

	متغیر	توضیحات
	price	قیمت خانه_ متغیر پاسخ
	yrbuilt	سال ساخت
مجموعه	TLA	جابجایی محل زندگی
داده	baths	تعداد حمام‌ها
لوکاس	halfbaths	تعداد حمام‌های کوچک
	garagesqft	متراژ گاراژ به فوت مربع
	lotsize	متراژ مسکن به فوت مربع*

* هر فوت‌مربع برابر ۰/۰۹۲۹۰۳۰۴ مترمربع

۱.۳.۴ انجام آزمون‌های آماری و برازش مدل‌های فراموضعی

برای این مجموعه داده نیز، ابتدا به بررسی وجود خودهمبستگی فضایی پرداختیم. جدول ۸.۴ نتایج را برای آزمون‌های مختلف خودهمبستگی فضایی نشان می‌دهد.

جدول ۸.۴: نتایج آزمون‌های خودهمبستگی فضایی برای داده‌های لوکاس

آزمون‌های خودهمبستگی فضایی	مقدار آماره آزمون	p_مقدار
موران I	۰/۳۶۷۴	< ۰/۰۰۰
نسبت درستنمایی	۱۱۲/۳۳	< ۰/۰۰۰
ضریب لاگرانژخطا	۲۸۷/۹۱	< ۰/۰۰۰
ضریب لاگرانژ تأخیر	۲۴۶/۲۵	< ۰/۰۰۰

همان‌طور که در جدول ۸.۴ نشان داده شده است، تمامی آزمون‌های خودهمبستگی معنادار بوده و با قدرت، حضور خودهمبستگی فضایی را در داده‌های لوکاس تایید می‌کنند. لذا مشابه داده‌های بالتیمور از مدل‌های اقتصادسنجی برای برازش استفاده می‌کنیم.

برای این داده‌ها از سه مدل تأخیر فضایی، خطای فضایی و ترکیبی فضایی استفاده کردیم. ماتریس‌های وزن در هر سه مدل به روش ۱۰_ نزدیکترین همسایگی محاسبه شدند. نتایج در جدول ۹.۴ گزارش شده‌اند.

جدول ۹.۴: برآورد پارامترهای مدل‌های فراموضعی برای مجموعه داده مسکن لوکاس

ضرایب	SLM		SEM		SDM	
عرض از مبدأ	−۵۵۲۵۹۹/۳۳	***	−۴۳۸۷۳۷/۹۷	***	−۳۰۵۰۵۶/۰۴	*
yrbuilt	۲۷۲/۲۰	***	۲۲۱/۷۲	***	۲۰۹/۵۵	***
TLA	۱۶/۴۳	***	۲۰/۶۸	***	۱۹/۶۴	***
baths	۱۴۲۰۴/۶۴	***	۱۴۵۷۸/۸۲	***	۱۵۱۸۵/۰۰	***
halfbaths	۴۹۷۶/۶۰	***	۵۰۰۷/۱	***	۵۱۹۳/۲۸	***
garagesqft	۱۴/۰۳	***	۱۱/۹۲	***	۱۱/۹۸	**
lotsize	۰/۹۶	***	۱/۰۲	***	۰/۱	***
M-yrbuilt					−۵۵/۲۰	
M-TLA					−۱۸/۴۸	***
M-baths					−۱۰۶۱۴/۳۶	
M-halfbaths					۵۸۳۷/۱۸	
M-garagesqft					۳/۰۲	
M-lotsize					۰/۹۰	
ρ	۰/۵۸۰	***	–		۰/۶۵۵	***
λ	–		۰/۸۳۱	***	–	
AIC	۸۳۰۲/۷۱۹		۸۲۸۹/۳۳۳		۸۲۶۹/۵۴۹	

معنی دار بودن: ۱/۰ < . ۵/۰ < * ۱/۰ < ** ۰۰۱/۰ < ***

بنابر نتایج جدول ۹.۴، تمامی متغیرهای معرفی شده در جدول ۷.۴ دارای اثرات معنی‌دار مثبت بر تعیین قیمت مسکن در بخش مرکزی شهرستان لوکاس هستند. به‌عنوان مثال، ضرایب متغیر yrbuilt که نشان‌دهنده سال ساخت هر خانه است، در هر سه مدل مثبت و دارای اثر معنی‌دار است. یعنی افزایش مقدار این متغیر، که به معنی کاهش قدمت خانه و نوساز بودن آن است، موجب افزایش قیمت خانه می‌شود.

ضرایب خودهمبستگی در هر سه مدل مثبت و معنی‌دار هستند. از میان مدل‌های برازش داده‌شده، مدل دوربین فضایی کمترین مقدار AIC را داراست. در نتیجه بهترین مدل برای برازش به داده‌های مسکن لوکاس، SDM است.

همان‌طور که گفته شد، در این مدل‌ها به جنبه دوم ناهمگنی، یعنی تفاوت اثرات متغیرهای تبیینی وابسته به فضای مشاهدات، توجه نشده است.

۲.۳.۴ ناهمگنی در داده‌ها

مشابه مثال اول، آزمون بریچ–پیگان را برای همه مدل‌های فراموضعی فضایی انجام دادیم تا ناهمگنی فضایی را در حضور انواع مدل‌های خودهمبستگی آزمون کنیم.

جدول ۱۰.۴: نتایج آزمون‌های ناهمگنی فضایی داده‌های مسکن لوکاس

مدل‌های فضایی	آماره بریچ_پیگان	درجه آزادی	p_مقدار
SLM	۸۱/۴۲۵	۶	$< ۰/۰۰۰۱$
SEM	۷۲/۲۳۱	۶	$< ۰/۰۰۰۱$
SDM	۹۰/۲۱۵	۱۲	$< ۰/۰۰۰۱$

با توجه به نتایج جدول ۱۰.۴، در این داده‌ها نیز حضور ناهمگنی با قوت تایید می‌شود. برای درک بهتر حضور ناهمگنی فضایی، مدل رگرسیون موزون جغرافیایی را به داده‌ها برازش دادیم. نتایج رگرسیون موزون جغرافیایی در جدول ۱۱.۴ گزارش شده‌اند. ناهمگنی در داده‌ها و تغییر پارامترها وابسته به فضای مشاهدات، در متغیرهای baths با بیش از ۳۰۰۰۰ واحد تغییر و halfbaths با ۴۰۰۰۰ واحد تغییر، به وضوح قابل مشاهده است.

جدول ۱۱.۴: نتایج برازش مدل GWR برای داده‌های مسکن لوکاس

ضرایب	مینیمم	چارک اول	میانه	چارک سوم	ماکسیمم
Intercept	−۱۳۴۶۴۵۸/۷۶	−۸۸۳۲۴۰/۴۴	−۶۴۷۲۷۲/۴۲	−۳۶۴۵۵۴/۹۲	۶۰۱۴۵/۳۵
yrbuilt	−۱۸/۸۹	۱۸۷/۰۸	۳۳۷/۲۲	۴۴۹/۲۴	۶۹۰/۵۰
TLA	۵/۱۳	۱۱/۶۷	۲۰/۰۵	۲۹/۱۳	۵۳/۸۶
baths	−۴۴۸۴/۷۲	۲۰۱۷/۲۷	۷۳۷۹/۸۳	۱۵۲۷۷/۲۴	۲۷۸۲۴/۵۳
halfbaths	−۲۰۷۵۵/۰۷	−۱۶۶۸/۳۱	۲۲۶۹/۹۷	۱۱۶۹۵/۶۸	۱۹۳۱۲/۰۴
garagesqft	−۶/۷۱	۸/۵۳	۱۲/۴۲	۱۵/۷۵	۳۱/۲۰
lotsize	−۰/۶۴	۰۰/۶۶	۱/۰۷	۱/۶۶	۲/۶۰

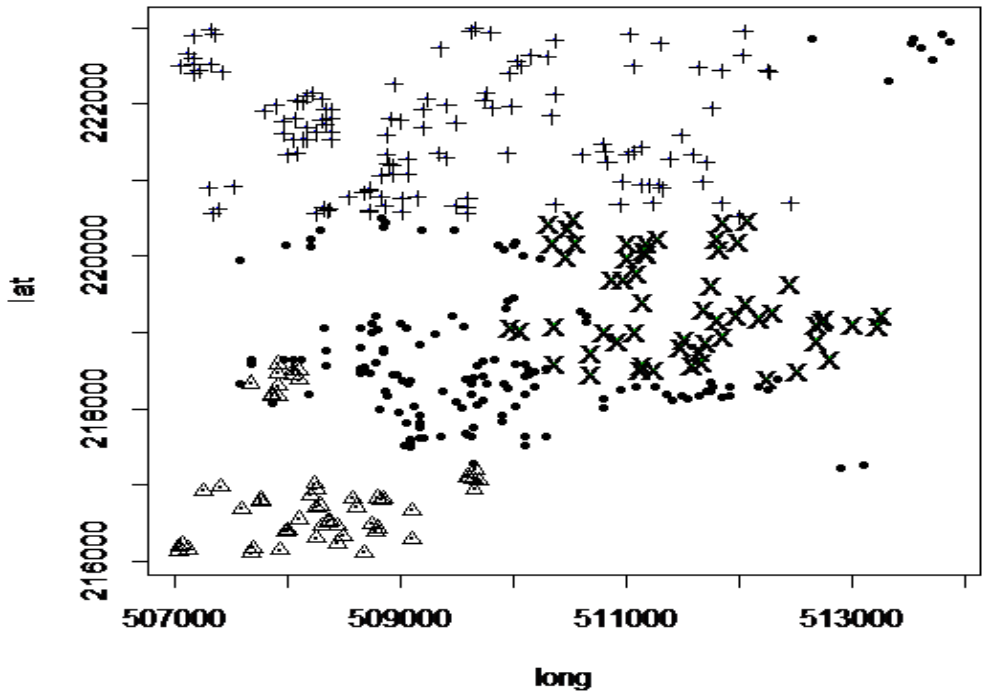
۳.۳.۴ برازش مدل پیشنهادی

بنابر نتایج به‌دست‌آمده، برای این داده‌ها نیز به دلیل حضور هم‌زمان ناهمگنی و خودهمبستگی فضایی، مدل OLS و مدل‌های فراموضعی فضایی مناسب نیستند؛ لذا از رهیافت پیشنهادی خود برای مدل‌بندی داده‌ها استفاده می‌کنیم.

نتایج رهیافت IGWR

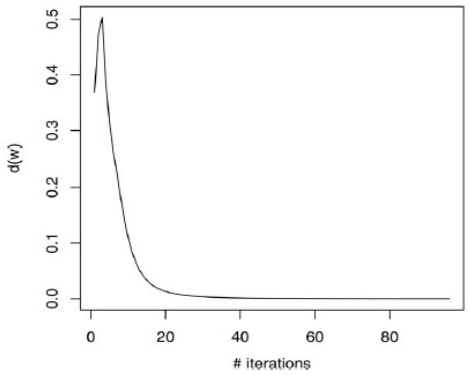
در این مثال نیز مشابه داده‌های بالتیمور، به منظور تعیین وزن‌های اولیه رگرسیون وزنی جغرافیایی از تابع هسته گاوسی با پهنای باند سازوار بر اساس معیار AIC استفاده شد. مقدار پهنای باند اولیه برای مجموعه داده مسکن لوکاس ، $b_{Knn}^{AIC} = ۱۹$ محاسبه شد. الگوریتم به‌روزرسانی وزن‌ها در خصوص این داده‌ها پس از ۱۸۲ تکرار به همگرایی رسید و $c = ۴$ رژیم

(شکل ۵.۴) با زیرنمونه‌هایی به حجم‌های $n_1 = ۵۷, n_2 = ۱۶۱, n_3 = ۵۸, n_4 = ۱۲۶$ تشخیص داده شدند. در این مثال، پارامترهای الگوریتم به‌روزرسانی وزن‌ها مشابه داده‌های بالتیمور در نظر گرفته‌شدند ($\omega = ۰/۰۰۰۱$، $\tau = ۰/۰۰۱$ و $\eta = ۰/۵$). حجم محاسبات در این مثال در مقایسه با داده‌های بالتیمور تقریبا دو برابر بود. این افزایش دو برابری با توجه به حجم داده‌های مسکن لوکاس که تقریبا دو برابر داده‌های بالتیمور است، منطقی است.

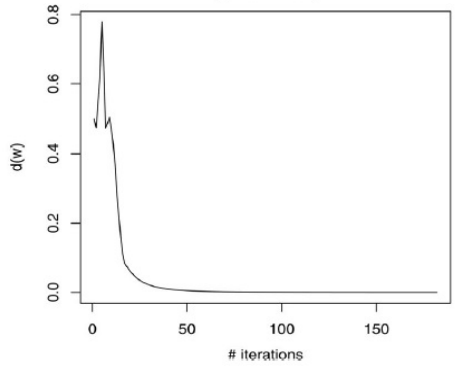


شکل ۵.۴: رژیم‌های فضایی مجموعه داده مسکن لوکاس، نقاط △ رژیم فضایی اول، نقاط • رژیم فضایی دوم، نقاط × رژیم فضایی سوم و نقاط + رژیم فضایی چهارم را نشان می‌دهند.

شکل ۶.۴ تابع تغییرات وزن‌ها (d(w)) را در تکرارهای مختلف نشان می‌دهد. در این نمودار رفتار فرایند همگرایی مشخص است. پس از یک دوره ناپایداری، اختلاف وزن‌ها با افزایش تکرار به صفر میل می‌کنند. بیشترین کاهش معنادار، در داده‌های بالتیمور در بازه تکرارهای ۱۵ – ۵ و در مجموعه مسکن لوکاس در تکرارهای ۳۰ – ۱۵ اتفاق افتاده است.



(آ) تابع تغییرات وزن مجموعه داده بالتیمور



(ب) تابع تغییرات وزن مجموعه داده مسکن لوکاس

شکل ۶.۴: تابع تغییرات وزن‌های روش IGWR

برازش مدل‌ها با رژیم فضایی درونی

در این بخش نیز مشابه مثال قبل، پس از مشخص شدن رژیم‌های فضایی، داده‌ها را به‌صورت (۷.۳) و (۹.۳) افراز کرده و مدل‌های تأخیر فضایی، خطای فضایی و دوربین فضایی با رژیم درونی را به داده‌ها برازش دادیم. ماتریس وزن **W** باز هم به روش ۱۰ـ نزدیکترین همسایگی محاسبه شد. نتایج در جدول ۱۲.۴نمایش داده شدند.

جدول ۱۲.۴: برآورد پارامترهای مدل‌های با رژیم فضایی درونی برای مجموعه داده مسکن شهر لوکاس

ضرایب	SLM		SEM		SDM	
عرض از مبدأ ۱	−۵۰۷۵۰۹/۸۲	**	−۷۵۲۸۶۳/۴۹	***	−۱۲۴۵۷۹۷/۶۴	***
عرض از مبدأ ۲	−۶۲۶۷۹۶/۷۴	***	−۴۹۳۴۴۱/۷۳	***	−۱۱۷۸۰۰۵/۷۱	***
عرض از مبدأ ۳	−۱۵۱۹۱۲/۸۱		−۲۰۸۸۲۴/۶۲		−۸۴۱۰۲۶/۶۸	**
عرض از مبدأ ۴	−۸۵۹۸۰۲/۴۷	***	−۶۳۹۹۳۴/۵۳	***	−۱۳۲۱۰۳۷/۸۰	***
yrbuilt 1	۲۵۴/۱۱	*	۳۹۱/۳۲	***	۳۰۵/۶۰	**
yrbuilt 2	۳۲۴/۲۵	***	۲۶۱/۶۶	***	۲۷۷/۹۶	***
yrbuilt 3	۷۶/۷۵		۱۱۱/۶۵		۱۰۲/۳۲	
yrbuilt 4	۴۲۸/۹۰	***	۳۲۱/۱۸	***	۳۳۸/۴۶	***
TLA 1	۱۲/۲۷	*	۱۲/۹۷	*	۱۳/۸۰	.
TLA 2	۱۱/۷۰	***	۱۳/۱۸	***	۱۲/۴۷	***
TLA 3	۸/۵۱	.	۷/۹۲	.	۷/۶۰	
TLA 4	۲۳/۹۵	***	۲۸/۲۲	***	۲۷/۷۳	***

ادامه جدول

جدول ۱۲.۴ – *ادامه جدول*

ضرایب	SLM		SEM		SDM	
baths 1	۲۲۸۴۱/۲۳	**	۲۰۱۳۴/۱۱	**	۲۲۷۱۷/۹۹	***
baths 2	۲۹۸۸/۹۱		۱۱۵۰/۷۵		۳۷۶۰/۲۴	
baths 3	۸۳۰/۷۶		۲۸۲۶/۹۵		۵۸۴۰/۱۹	
baths 4	۱۱۲۶۲/۴۹	*	۱۱۱۲۹/۵۹	***	۱۲۹۱۷/۰۴	***
halfbaths 1	۱۹۴۱/۸۹		−۱۵۳۴/۵۲		−۲۴۹۴/۸۵	
halfbaths 2	۲۶۷۶/۹۱		۳۸۶/۶۰		۲۰۶۹/۱۱	
halfbaths 3	−۱۶۹۷/۷۲		−۴۸۲۸/۶۹		۱۱۱۵/۱۴	
halfbaths 4	۸۱۹۴/۴۳	***	۶۱۹۴/۴۳	***	۷۵۱۸/۸۰	***
garagesqft 1	۱۵/۸۶		۲۰/۹۳	.	۸/۴۴	
garagesqft 2	۱۱/۷۳	*	۱۰/۰۴	*	۷/۵۳	
garagesqft 3	۱۲/۸۳	.	۹/۸۴		۱۳/۶۶	.
garagesqft 4	۱۰/۳۸	.	۱۲/۸۲	*	۷/۹۰	
lotsize 1	۱/۰۳	*	۱/۰۸	**	۰/۹۴	.
lotsize 2	۰/۵۵		۰/۹۳	*	۰/۶۸	
lotsize 3	۰/۱		۱/۲۴		۰/۰۸	
lotsize 4	۱/۰۲	*	۱/۳۵	***	۱/۰۳	**
M-yrbuilt 1					۳۱۵/۷۰	**
M-yrbuilt 2					۳۴۳/۱۰	***
M-yrbuilt 3					۳۴۴/۲۴	**
M-yrbuilt 4					۳۴۱/۱۴	**
M-TLA 1					۴۶/۲۶	
M-TLA 2					−۷/۹۴	
M-TLA 3					−۹/۵۶	
M-TLA 4					−۱۱/۳۴	
M-baths 1					۴۶۴۲۴/۶۰	
M-baths 2					۸۵۶۸/۶۴	
M-baths 3					۱۵۱۵۹/۶۸	
M-baths 4					−۳۵۵۶/۳۹	
M-halfbaths 1					۱۹۴۷۲/۴۸	
M-halfbaths 2					۳۰۱۴/۳۲	
M-halfbaths 3					۲۹۴۹۷/۴۲	

ادامه جدول

جدول ۱۲.۴ – *ادامه جدول*

ضرایب	SLM		SEM		SDM	
M-halfbaths 4					۲۲۰۵۶/۶۰	***
M-garagesqft 1					−۱۶۶/۲۱	**
M-garagesqft 2					−۱۴/۸۰	
M-garagesqft 3					۶۳/۱۱	*
M-garagesqft 4					−۲۷/۵۶	.
M-lotsize 1					−۲/۱۷	.
M-lotsize 2					−۱/۴۳	
M-lotsize 3					−۸/۴۸	
M-lotsize 4					۰/۲۶	
ρ	۰/۴۴۴	***	–		۰/۲۵	**
λ	–		۰/۷۹۱	***		
AIC	۸۲۵۲/۴۱۷		۸۲۴۶/۱۳۲			

معنی دار بودن: ۱/۰ > . ۰/۰۵ > * ۰/۰۱ > ** ۰/۰۰۱ > ***

در مقایسه نتایج مدل‌های فراموضعی و مدل‌ها با اعمال رژیم فضایی، مشاهده می‌کنیم که مقدار AIC در همه موارد در ESRMs کمتر است. در مورد داده‌های مسکن لوکاس نه‌تنها نتایج ESRMs از نظر مقدار AIC بهتر شده‌اند، بلکه در آن‌ها اثرات سرریز فضایی ρ نیز رقیق شده‌اند. این کاهش در مدل دوربین فضایی از ۰/۶۵۵ به ۰/۲۴۸ کاملا مشهود است. این نتیجه به این معنی است که اطلاعات اضافه‌شده توسط رژیم‌های فضایی ما را از اثرات سرریز فضایی آشکار به اثرات سرریز فضایی واقعی می‌رساند.

به عنوان نتیجه می‌توان گفت مدل دوربین فضایی با رژیم فضایی درونی نسبت به باقی مدل‌ها از قدرت بیشتری برخوردار است؛ زیرا طوری طراحی شده است که اثرات آشکار فضایی را به ۳ قسمت: اثرات سرریز فضایی واقعی، متغیرهای همبسته فضایی تبیینی و ناهمگنی فضایی، مدل‌بندی می‌کند.

۴.۴ نتیجه و آینده تحقیق

در این پایان‌نامه یک روش دو مرحله‌ای، که در مرحله اول آن رژیم‌های فضایی تعیین و در مرحله دوم ناهمگنی (تحت عنوان رژیم‌های فضایی) و وابستگی فضایی به‌طور هم‌زمان برآورد شدند، معرفی شد. با استفاده از دو مجموعه داده واقعی مربوط به تعیین قیمت مسکن، نشان دادیم که این روش می‌تواند نواحی همگن را شناسایی کند. مدل‌های اقتصادسنجی

فضایی که با اعمال رژیم‌بندی‌ها برآورد شدند نسبت به حالت فراموضعی خود ارجح‌تر هستند. همچنین اکثر نتایج نشان داد که اثرات خودهمبستگی فضایی در داده‌ها وجود دارد اما این اثرات کمتر از مقداری است که در مدل‌های فراموضعی برآورد می‌شود. این اختلاف در برآورد اثر خودهمبستگی، علاوه بر تأثیر متغیرهای حذف‌شده، ناشی از عدم توجه به ناهمگنی فضایی است. بنابراین برای برازش یک مدل فضایی منعطف، مدل‌بندی هم‌زمان وابستگی و ناهمگنی یک گام اساسی است.

در مرحله اول این رهیافت برای تعیین رژیم‌ها هیچ محدودیتی لحاظ نشده است و به اطلاعات پیشین در مورد ساختار فضایی داده‌ها، که برای اغلب پدیده‌ها در دسترس نیست، نیازی ندارد. لذا به‌راحتی برای هر مدلی که امکان برآورد موضعی آن توسط کمترین توان‌های دوم موزون وجود دارد، قابل تعمیم است. یک تغییر شدنی و جالب، تعمیم این رهیافت برای مجموعه داده‌های پانلی به منظور لحاظ کردن تغییرات زمانی (علاوه بر ساختار فضایی) است.

پیوست آ

برنامه‌های رایانه‌ای با نرم‌افزار R

آ.۱ برنامه‌های داده بالتیمور

دستورات لازم برای تعیین ماتریس وزن W و انجام آزمون‌های خودهمبستگی فضایی

```
library(spdep)
data(baltimore)
View(baltimore)
plot(baltimore$X,baltimore$Y,cex=1,pch=19,xlab=long,ylab=lat)

coords<-cbind(baltimore$X,baltimore$Y)
knn    <- knn2nb(knearneigh(coords,k=10))
listw_10nn_dates <- nb2listw(knn)
OLS <- lm(PRICE~DWELL+NBATH+PATIO+FIREPL+AC+BMENT+GAR+CITCOU+LOTSZ,
data=baltimore)
summary(OLS)

balt.moran <- lm.morantest(OLS,listw_10nn_dates)
```

```
balt.moran

lmtest<-lm.LMtests(OLS,listw_10nn_dates,test=c("LMerr","LMlag"))
lmtest
```

برازش مدل‌های اقتصادسنجی فضایی

```
SLM <-lagsarlm(PRICE~DWELL+NBATH+PATIO+FIREPL+AC+BMENT+GAR+CITCOU+LOTSZ,
data=baltimore, listw_10nn_dates, type="lag")
summary(SLM)

SDM <- lagsarlm(PRICE~DWELL+NBATH+PATIO+FIREPL+AC+BMENT+GAR+CITCOU+LOTSZ,
data=baltimore, listw_10nn_dates, type="durbin")
summary(SDM)

SEM <- errorsarlm(PRICE~DWELL+NBATH+PATIO+FIREPL+AC+BMENT+GAR+CITCOU+LOTSZ,
data=baltimore, listw_10nn_dates, method="eigen", quiet=FALSE)
summary(SEM)

SAC <- sacsarlm(PRICE~DWELL+NBATH+PATIO+FIREPL+AC+BMENT+GAR+CITCOU+LOTSZ,
data=baltimore, listw_10nn_dates)
```

آزمون ناهمگنی در حضور خودهمبستگی فضایی (آزمون بریچ‌ـ پیگان)

```
bptest.sarlm(SLM)

bptest.sarlm(SDM)

bptest.sarlm(SEM)

bptest.sarlm(SAC)
```

دستورات لازم برای برازش مدل رگرسیون موزون جغرافیایی با پهنای باند بهینه با استفاده از معیار اطلاع آکاییک

```
library(GWmodel)
```

```
dmat <- gw.dist(coords,focus=0,p=2,theta=0,longlat=F)
bw <- bw.gwr(PRICE~DWELL+NBATH+PATIO+FIREPL+AC+BMENT+GAR+CITCOU+LOTSZ,
data=SpatialPointsDataFrame(coords,baltimore),approach="AIC",kernel="gaussian",
adaptive=T,p=2,theta=0,longlat=F,dMat=dmat)     # 33 nearnighber
gwr.res <- gwr.basic(PRICE~DWELL+NBATH+PATIO+FIREPL+AC+BMENT+GAR+CITCOU+LOTSZ,
data=SpatialPointsDataFrame(coords,baltimore), bw=bw,
kernel = "gaussian",adaptive=TRUE, dMat=dmat)
gwr.res
```

دستورات رهیافت IGWR

```
library(spdep)
library(GWmodel)
coords<-cbind(baltimore$X,baltimore$Y)
dmat<-gw.dist(coords,focus=0,p=2,theta=0,longlat=F)
bw<-bw.gwr(PRICE~DWELL+NBATH+PATIO+FIREPL+AC+BMENT+GAR+CITCOU+LOTSZ,
data=SpatialPointsDataFrame(coords,baltimore),approach="AIC",kernel="gaussian",
adaptive=T,p=2,theta=0,longlat=F,dMat=dmat)
y<-as.matrix(baltimore$PRICE)
x<-cbind(baltimore$DWELL,baltimore$NBATH,baltimore$PATIO,baltimore$FIREPL,
baltimore$AC,baltimore$BMENT,baltimore$GAR,baltimore$CITCOU,baltimore$LOTSZ)

modparcov <- function (y, x, w)   #GWR/LWR
{
  xpw        <- t(x)%*%w                          #w matrice diagonale
  invxpwx    <- solve(xpw%*%x)
  b          <- invxpwx%*%(xpw%*%y)               #kx1 weighted least squares
  resid      <- y - x %*% b                       #nx1
  qresid     <- t(resid)%*%w%*%resid
  c          <- invxpwx%*%xpw
  s          <- x%*%c
  v1         <- sum(diag(s))
  v2         <- sum(diag(t(s)%*%s))
  varresid   <- qresid/(nrow(x)-2*v1+v2)          #sigma2
  ccp        <- c%*%t(c)
```

```
  res        <- list()
  res$par    <- b                                #kx1
  res$cov    <- ccp*as.numeric(varresid)         #kxk
  res
}

testpar <- function (p1, p2, cov1, cov2, w1, w2) #wald test
{
  n1      <- as.numeric(sum(diag(w1)))
  n2      <- as.numeric(sum(diag(w2)))
  pcov    <- (n1*cov1+n2*cov2)/(n1+n2)          #pooled var-cov beta
  t       <- t(p1-p2)%*%solve(pcov)%*%(p1-p2)
  t
}

awsreg <- function (y,x,coords,bw,tau,niter,conv,eta,numout,sout)  #conv=omega
{
x      <- as.matrix(x)
y      <- as.matrix(y)
codic  <- 1:nrow(x)
clas   <- rep(0,nrow(x))
woutl  <- rep(0,nrow(x))

vdist  <- as.matrix(gw.dist(coords,focus=0,p=2,theta=0,longlat=F)) #dmat
knn    <- knn2nb(knearneigh(coords,k=bw))
nbdist <- nbdists(knn,coords)

 #w<-matrix(0,nrow(vdist),ncol(vdist))
 #for (i in 1:nrow(vdist)){
 #for (j in 1:ncol(vdist)){
 #   w[i,j] <- exp(-vdist[i,j]/max(nbdist[[i]]))      #esponenziale
 #}}

  w<-matrix(0,nrow(vdist),ncol(vdist))
```

```
for (i in 1:nrow(vdist)){
for (j in 1:ncol(vdist)){
   w[i,j] <- exp(-0.5*(vdist[i,j]/max(nbdist[[i]]))^2)   #gaussiana
}}

#w<-matrix(0,nrow(vdist),ncol(vdist))                 #bisquare
#for (i in 1:nrow(vdist)){
#for (j in 1:ncol(vdist)){
#     if (vdist[i,j]<max(nbdist[[i]]))
# w[i,j] <- (1-(vdist[i,j]/max(nbdist[[i]]))^2)^2
#     else          w[i,j] <- 0.00001
#}}

#w<-matrix(0,nrow(vdist),ncol(vdist))               #tricube
#for (i in 1:nrow(vdist)){
#for (j in 1:ncol(vdist)){
#     if (vdist[i,j]<max(nbdist[[i]]))
#w[i,j] <- (1-(vdist[i,j]/max(nbdist[[i]]))^3)^3
#     else                              w[i,j] <- 0.00001
#}}

wnew   <- matrix(0,nrow(x),nrow(x))
parm   <- matrix(0,nrow(x),ncol(x))   #matrice parametri nxk
qresid <- rep(0,length=nrow(x))
listcov <- list()

nout   <- 0
j      <- 0
differ <- 1
while (differ > conv && (j <- j+1)<=niter)
{
  listw<-list()
  for (i in 1:ncol(w))
  {
```

```
    listw[[i]]<-diag(w[,i])  #w[i,] pesi x riga
   }
   for (i in 1:nrow(x))
   {
     wls         <- modparcov(y, x, listw[[i]])
     parm[i,]    <- wls$par
     listcov[[i]]<- wls$cov
   }
   for (u in 1:(nrow(x)-1))                    #test
   {
     wnew[u,u] <- 1
     for (v in (u+1):nrow(x))
  {
  if (w[u,v] < conv)
  {
   wnew[u,v] <- 0
  wnew[v,u] <- wnew[u,v]
 }
  else
  {
#wnew[u,v] <- exp(-testpar(parm[u,],parm[v,],
#listcov[[u]],listcov[[v]],listw[[u]],listw[[v]])*tau)        #esponenziale
 wnew[u,v] <- exp(-0.5*(testpar(parm[u,],parm[v,],
 listcov[[u]],listcov[[v]],listw[[u]],listw[[v]])*tau)^2) #gaussiana
  wnew[v,u] <- wnew[u,v]
       }
     }
   }
   wnew[nrow(x),nrow(x)] <- 1

 #wnew2<-matrix(0,nrow(vdist),ncol(vdist))                #esponenziale
 #for (i in 1:nrow(vdist)){
 #for (h in 1:ncol(vdist)){
 #   wnew2[i,h]<-wnew[i,h]*exp(-vdist[i,h]/(max(nbdist[[i]])*j))
```

```
#}}
#wnew<-wnew2

  wnew2<-matrix(0,nrow(vdist),ncol(vdist))                    #gaussiana
  for (i in 1:nrow(vdist)){
  for (h in 1:ncol(vdist)){
    wnew2[i,h]<-wnew[i,h]*exp(-0.5*(vdist[i,h]/(max(nbdist[[i]])*j))^2)
  }}
  wnew<-wnew2

  differ <- max(abs(w-wnew))
  w      <- eta*w+(1-eta)*wnew

  for (i in 1:nrow(x))            #controllo outlier
  {
    wcontr    <- w[,i]
    woutl[i]  <- length(wcontr[wcontr > sout])
  }
  selez  <- which(woutl[1:nrow(x)] < numout)
  if (length(selez) > 0)
  {
    nout   <- nout + length(selez)
    x      <- x[-selez,]
    y      <- y[-selez,]
    codic  <- codic[-selez]
    vdist  <- vdist[-selez,-selez]
    w      <- w[-selez,-selez]
    wnew   <- wnew[-selez,-selez]
  }
  print(c(j,differ,nout))
}
ncl <- 0
for (i in 1:(nrow(x)-1))
{
```

```
    if (clas[codic[i]] < 1)
    {
      ncl     <- ncl + 1
      clas[codic[i]] <- ncl
      for (j in (i+1):nrow(x))
      {
       if (w[i,j] > 0.9) clas[codic[j]] <- clas[codic[i]]
      }
    }
  }
  if (clas[nrow(x)] < 1) clas[nrow(x)] <- ncl + 1
clas[clas == 0] <- NA
clas
}

#            alfa/2
tau<-qchisq(p=0.100,df=1)  #0.015
tau<-qchisq(p=0.050,df=1)  #0.004
tau<-qchisq(p=0.025,df=1)  #0.001 tau alfa=0.05

system.time(
kk <- awsreg(y,x,coords=coords,bw=bw,tau=0.001,
niter=200,conv=0.0001,eta=0.5,numout=20,sout=1e-05) #eta=memory parameter
)
kk

kk[is.na(kk)] <- max(kk,na.rm = TRUE)+1

# if more than 8 groups (no missing)
kk1<-kk
kk1[kk>8]<-0
plot(baltimore$X,baltimore$Y,cex=1,pch=20,col=kk1)
kk2<-kk
kk2[kk<9]<-0
```

```
points(baltimore$X,baltimore$Y,cex=1,pch=12,col=kk2)
plot(baltimore$X,baltimore$Y,cex=1,pch=1,col=kk,xlab="long",ylab="lat")
```

برنامه‌های لازم برای افراز داده‌ها و رژیم‌بندی آن‌ها

```
group1 = baltimore[kk==1,]
group1x<-cbind(group1$DWELL,group1$NBATH,group1$PATIO,group1$FIREPL,
group1$AC,group1$BMENT,group1$GAR,group1$CITCOU,group1$LOTSZ) # X regim1

group2 = baltimore[kk==2,]
group2x<-cbind(group2$DWELL,group2$NBATH,group2$PATIO,group2$FIREPL,
group2$AC,group2$BMENT,group2$GAR,group2$CITCOU,group2$LOTSZ) #  X regim2

newy = cbind(t(baltimore$PRICE[kk==1]),t(baltimore$PRICE[kk==2]))
newy = t(newy)
newx = rbind(cbind(1,group1x,matrix(0,nr=nrow(group1x),nc=ncol(group2x)+1)),
cbind(matrix(0,nr=nrow(group2x),nc=ncol(group1x)+1),1,group2x))

newbalt = cbind(newy,newx)     #new  baltimore data set

newbalt = data.frame(newbalt)

names(newbalt) = c("PRICE","intercept1","DWELL1","NBATH1","PATIO1",
"FIREPL1","AC1","BMENT1","GAR1","CITCOU1","LOTSZ1","intercept2",
"DWELL2","NBATH2","PATIO2","FIREPL2","AC2",
"BMENT2","GAR2","CITCOU2","LOTSZ2")
```

دستورات لازم برای معرفی ماتریس وزن جدید و برازش مدل‌های اقتصادسنجی با رژیم فضایی درونی

```
coordinat1 = cbind(group1$X,group1$Y)
coordinat2 = cbind(group2$X,group2$Y)
coords1 = rbind(coordinat1,coordinat2)  #new coordinat for baltimor data set
knn1    <- knn2nb(knearneigh(coords1,k=10))
```

```
listw_10nn_dates1 <- nb2listw(knn1)

SLM.ES <- lagsarlm(PRICE~-1+intercept1+intercept2+DWELL1+DWELL2+
NBATH1+NBATH2+PATIO1+PATIO2+FIREPL1+FIREPL2+AC1+AC2+
BMENT1+BMENT2+GAR1+GAR2+CITCOU1+CITCOU2+LOTSZ1+LOTSZ2,
data=newbalt, listw_10nn_dates1, type="lag")
summary(SLM.ES)

SDM.ES <- lagsarlm(PRICE~-1+intercept1+intercept2+DWELL1+DWELL2+
NBATH1+NBATH2+PATIO1+PATIO2+FIREPL1+FIREPL2+
AC1+AC2+BMENT1+BMENT2+GAR1+GAR2+CITCOU1+CITCOU2+LOTSZ1+LOTSZ2,
data=newbalt, listw_10nn_dates1, Durbin=~DWELL1+DWELL2+NBATH1+NBATH2+
PATIO1+PATIO2+FIREPL1+FIREPL2+AC1+AC2+BMENT1+
BMENT2+GAR1+GAR2+CITCOU1+CITCOU2+LOTSZ1+LOTSZ2)
summary(SDM.ES)

SEM.ES <- errorsarlm(PRICE~-1+intercept1+intercept2+DWELL1+DWELL2+
NBATH1+NBATH2+PATIO1+PATIO2+FIREPL1+FIREPL2+AC1+AC2+
BMENT1+BMENT2+GAR1+GAR2+CITCOU1+CITCOU2+LOTSZ1+LOTSZ2,
data=newbalt, listw_10nn_dates1, method="eigen", quiet=FALSE)
summary(SEM.ES)

SAC.ES <- sacsarlm(PRICE~-1+intercept1+intercept2+DWELL1+DWELL2+
NBATH1+NBATH2+PATIO1+PATIO2+FIREPL1+FIREPL2+AC1+AC2+
BMENT1+BMENT2+GAR1+GAR2+CITCOU1+CITCOU2+LOTSZ1+LOTSZ2,
data=newbalt, listw_10nn_dates1)
summary(SAC.ES)
```

۲.آ برنامه‌های داده مسکن

دستورات لازم برای انتخاب داده‌های مسکن بخش مرکزی شهرستان لوکاس که در سال ۱۹۹۳ به فروش رسیده‌اند، تعیین ماتریس وزن $\mathbf{W}$ و انجام آزمون‌های خودهمبستگی فضایی

```
library(spdep)
data(house)
plot(coordinates(house),cex=.4,pch=19)

subhouse<-subset(house,house$long>507000 & house$long<514000
& house$lat>216000 & house$lat<223000 & house$s1993==1)

points(coordinates(subhouse),cex=.4,pch=19,col="blue")

coords<- coordinates(subhouse)
knn    <- knn2nb(knearneigh(coords,k=10))
listw_10nn_dates <- nb2listw(knn)
OLS <- lm(price~yrbuilt+TLA+baths+halfbaths+garagesqft+lotsize,
data=subhouse)
summary(OLS)

balt.moran <- lm.morantest(OLS,listw_10nn_dates)
balt.moran

lmtest<-lm.LMtests(OLS,listw_10nn_dates,test=c("LMerr","LMlag"))
lmtest
```

برازش مدل‌های اقتصادسنجی فضایی

```
SLM <- lagsarlm(price~yrbuilt+TLA+baths+halfbaths+garagesqft+lotsize,
data=subhouse, listw_10nn_dates, type="lag",tol.solve = 1.39885e-19)
summary(SLM)

SDM <- lagsarlm(price~yrbuilt+TLA+baths+halfbaths+garagesqft+lotsize,
data=subhouse, listw_10nn_dates, type="mixed",tol.solve = 1.39885e-19)
```

```
summary(SDM)

SEM <- errorsarlm(price~yrbuilt+TLA+baths+halfbaths+garagesqft+lotsize,
data=subhouse, listw_10nn_dates, method="eigen",
quiet=FALSE,tol.solve = 1.39885e-19)
summary(SEM)
```

آزمون ناهمگنی در حضور خودهمبستگی فضایی (آزمون بریچ_پیگان)

```
bptest.sarlm(SLM)

bptest.sarlm(SEM)

bptest.sarlm(SDM)
```

دستورات لازم برای برازش مدل رگرسیون موزون جغرافیایی با پهنای باند بهینه با استفاده از معیار اطلاع آکاییک

```
library(GWmodel)
dmat<-gw.dist(coords,focus=0,p=2,theta=0,longlat=F)
bw<-bw.gwr(price~yrbuilt+TLA+baths+halfbaths+garagesqft+lotsize,
data=SpatialPointsDataFrame(coords,subhouse@data),
approach="AIC",kernel="gaussian",adaptive=T,p=2,
theta=0,longlat=F,dMat=dmat)   #19 nearnighber

gwr.res1<-gwr.basic(price~yrbuilt+TLA+baths+halfbaths+garagesqft+lotsize,
data=SpatialPointsDataFrame(coords,subhouse@data), bw=bw,
kernel = "gaussian",adaptive=TRUE, dMat=dmat)
gwr.res1
```

دستورات رهیافت IGWR

```
library(spdep)
library(GWmodel)

dmat<-gw.dist(coords,focus=0,p=2,theta=0,longlat=F)
```

```
bw<-bw.gwr(price~yrbuilt+TLA+baths+halfbaths+garagesqft+lotsize,
data=SpatialPointsDataFrame(coords,subhouse@data),
approach="AIC",kernel="gaussian",adaptive=T,p=2,
theta=0,longlat=F,dMat=dmat)   #19 nearnighber

y<-as.matrix(subhouse$price)
x<-cbind(subhouse$yrbuilt,subhouse$TLA,subhouse$baths,
subhouse$halfbaths,subhouse$garagesqft,subhouse$lotsize)

modparcov <- function (y, x, w)   #GWR/LWR
{
  xpw        <- t(x)%*%w                          #w matrice diagonale
  invxpwx    <- solve(xpw%*%x)
  b          <- invxpwx%*%(xpw%*%y)               #kx1 weighted least squares
  resid      <- y - x %*% b                       #nx1
  qresid     <- t(resid)%*%w%*%resid
  c          <- invxpwx%*%xpw
  s          <- x%*%c
  v1         <- sum(diag(s))
  v2         <- sum(diag(t(s)%*%s))
  varresid   <- qresid/(nrow(x)-2*v1+v2)          #sigma2
  ccp        <- c%*%t(c)
  res        <- list()
  res$par    <- b                                 #kx1
  res$cov    <- ccp*as.numeric(varresid)          #kxk
  res
}

testpar <- function (p1, p2, cov1, cov2, w1, w2) #wald test
{
  n1      <- as.numeric(sum(diag(w1)))
  n2      <- as.numeric(sum(diag(w2)))
  pcov    <- (n1*cov1+n2*cov2)/(n1+n2)           #pooled var-cov beta
```

```
  t       <- t(p1-p2)%*%solve(pcov)%*%(p1-p2)
  t
}

awsreg <- function (y,x,coords,bw,tau,niter,conv,eta,numout,sout)  #conv=omega
{
  x      <- as.matrix(x)
  y      <- as.matrix(y)
  codic  <- 1:nrow(x)
  clas   <- rep(0,nrow(x))
  woutl  <- rep(0,nrow(x))

  vdist  <- as.matrix(gw.dist(coords,focus=0,p=2,theta=0,longlat=F)) #dmat
  knn    <- knn2nb(knearneigh(coords,k=bw))
  nbdist <- nbdists(knn,coords)

  #w<-matrix(0,nrow(vdist),ncol(vdist))
  #for (i in 1:nrow(vdist)){
  #for (j in 1:ncol(vdist)){
  #   w[i,j] <- exp(-vdist[i,j]/max(nbdist[[i]]))      #esponenziale
  #}}

  w<-matrix(0,nrow(vdist),ncol(vdist))
  for (i in 1:nrow(vdist)){
  for (j in 1:ncol(vdist)){
     w[i,j] <- exp(-0.5*(vdist[i,j]/max(nbdist[[i]]))^2)   #gaussiana
  }}

  #w<-matrix(0,nrow(vdist),ncol(vdist))                 #bisquare
  #for (i in 1:nrow(vdist)){
  #for (j in 1:ncol(vdist)){
  #     if (vdist[i,j]<max(nbdist[[i]]))
  #w[i,j] <- (1-(vdist[i,j]/max(nbdist[[i]]))^2)^2
  #     else                              w[i,j] <- 0.00001
```

```
#}}

#w<-matrix(0,nrow(vdist),ncol(vdist))              #tricube
#for (i in 1:nrow(vdist)){
#for (j in 1:ncol(vdist)){
#     if (vdist[i,j]<max(nbdist[[i]]))
#w[i,j] <- (1-(vdist[i,j]/max(nbdist[[i]]))^3)^3
#     else                              w[i,j] <- 0.00001
#}}

wnew   <- matrix(0,nrow(x),nrow(x))
parm   <- matrix(0,nrow(x),ncol(x))
qresid <- rep(0,length=nrow(x))
listcov <- list()

nout   <- 0
j      <- 0
differ <- 1
while (differ > conv && (j <- j+1)<=niter)
{
  listw<-list()
  for (i in 1:ncol(w))
  {
    listw[[i]]<-diag(w[,i])
  }
  for (i in 1:nrow(x))
  {
    wls         <- modparcov(y, x, listw[[i]])
    parm[i,]    <- wls$par
    listcov[[i]]<- wls$cov
  }
  for (u in 1:(nrow(x)-1))                    #test
  {
    wnew[u,u] <- 1
```

```
  for (v in (u+1):nrow(x))
  {
    if (w[u,v] < conv)
    {
      wnew[u,v] <- 0
      wnew[v,u] <- wnew[u,v]
    }
    else
    {
      #wnew[u,v] <- exp(-testpar(parm[u,],parm[v,],listcov[[u]],
      #listcov[[v]],listw[[u]],listw[[v]])*tau)        #esponenziale

      wnew[u,v] <- exp(-0.5*(testpar(parm[u,],parm[v,],listcov[[u]],
      listcov[[v]],listw[[u]],listw[[v]])*tau)^2) #gaussiana
      wnew[v,u] <- wnew[u,v] #per simmetria
    }
  }
}
wnew[nrow(x),nrow(x)] <- 1

#wnew2<-matrix(0,nrow(vdist),ncol(vdist))                #esponenziale
#for (i in 1:nrow(vdist)){
#for (h in 1:ncol(vdist)){
#   wnew2[i,h]<-wnew[i,h]*exp(-vdist[i,h]/(max(nbdist[[i]])*j))
#}}
#wnew<-wnew2

wnew2<-matrix(0,nrow(vdist),ncol(vdist))                #gaussiana
for (i in 1:nrow(vdist)){
for (h in 1:ncol(vdist)){
   wnew2[i,h]<-wnew[i,h]*exp(-0.5*(vdist[i,h]/(max(nbdist[[i]])*j))^2)
}}
wnew<-wnew2
```

```
  differ <- max(abs(w-wnew))
  w      <- eta*w+(1-eta)*wnew

  for (i in 1:nrow(x))            #controllo outlier
  {
    wcontr    <- w[,i]
    woutl[i]  <- length(wcontr[wcontr > sout])
  }
  selez  <- which(woutl[1:nrow(x)] < numout)
  if (length(selez) > 0)
  {
    nout   <- nout + length(selez)
    x      <- x[-selez,]
    y      <- y[-selez,]
    codic  <- codic[-selez]
    vdist  <- vdist[-selez,-selez]
    w      <- w[-selez,-selez]
    wnew   <- wnew[-selez,-selez]
  }
  print(c(j,differ,nout))
}
ncl <- 0
for (i in 1:(nrow(x)-1))
{
  if (clas[codic[i]] < 1)
  {
    ncl     <- ncl + 1
    clas[codic[i]] <- ncl
    for (j in (i+1):nrow(x))
    {
      if (w[i,j] > 0.9) clas[codic[j]] <- clas[codic[i]]
    }
  }
}
```

```
  if (clas[nrow(x)] < 1) clas[nrow(x)] <- ncl + 1
  clas[clas == 0] <- NA
  clas
}

#            alfa/2
tau<-qchisq(p=0.100,df=1)  #0.015
tau<-qchisq(p=0.050,df=1)  #0.004
tau<-qchisq(p=0.025,df=1)  #0.001 tau alfa=0.05

system.time(
kk<-awsreg(y,x,coords=coords,bw=bw,tau=0.001,niter=200,
conv=0.0001,eta=0.5,numout=20,sout=1e-05)
)
kk

kk[is.na(kk)] <- max(kk,na.rm = TRUE)+1

# if more than 8 groups (no missing)
kk1<-kk
kk1[kk>8]<-0
plot(coordinates(subhouse),cex=.1,pch=20,col=kk1)

kk2<-kk
kk2[kk<9]<-0
points(coordinates(subhouse),cex=.6,pch=12,col=kk2)

plot(coordinates(subhouse),cex=.6,pch=19,col=kk)
```

برنامه‌های لازم برای افراز داده‌ها و رژیم‌بندی آن‌ها

```
group1 = subhouse[kk==1,]
group2 = subhouse[kk==2,]
group3 = subhouse[kk==3,]
```

```
group4 = subhouse[kk==4,]

newy = cbind(t(subhouse$price[kk==1]),t(subhouse$price[kk==2]),
t(subhouse$price[kk==3]),t(subhouse$price[kk==4]))
newy = t(newy)
dim(newy)

group1x<-cbind(1,group1$yrbuilt,group1$TLA,group1$baths,
group1$halfbaths,group1$garagesqft,group1$lotsize)    # X regim1

group2x<-cbind(1,group2$yrbuilt,group2$TLA,group2$baths,
group2$halfbaths,group2$garagesqft,group2$lotsize)       # X regim2

group3x<-cbind(1,group3$yrbuilt,group3$TLA,group3$baths,
group3$halfbaths,group3$garagesqft,group3$lotsize)          # X regim3

group4x<-cbind(1,group4$yrbuilt,group4$TLA,group4$baths,
group4$halfbaths,group4$garagesqft,group4$lotsize)      # X regim4

newx = as.matrix(bdiag(group1x,group2x,group3x,group4x))
dim(newx)

newsubhouse = cbind(newy,newx)     #new  house data set

newsubhouse = data.frame(newsubhouse)

names(newsubhouse)

names(newsubhouse) = c("price","intercept1","yrbuilt1","TLA1",
"baths1","halfbaths1","garagesqft1","lotsize1","intercept2","yrbuilt2",
"TLA2","baths2","halfbaths2","garagesqft2","lotsize2","intercept3","yrbuilt3",
"TLA3","baths3","halfbaths3","garagesqft3","lotsize3","intercept4",
"yrbuilt4","TLA4","baths4","halfbaths4","garagesqft4","lotsize4"
)
```

دستورات لازم برای معرفی ماتریس وزن جدید و برازش مدل‌های اقتصادسنجی با رژیم فضایی درونی

```
coordinat1 = cbind(group1$long,group1$lat)
coordinat2 = cbind(group2$long,group2$lat)
coordinat3 = cbind(group3$long,group3$lat)
coordinat4 = cbind(group4$long,group4$lat)
coords1 = rbind(coordinat1,coordinat2,coordinat3,
coordinat4)     #new coordinat for baltimor data set
knn1    <- knn2nb(knearneigh(coords1,k=10))
listw_10nn_dates1 <- nb2listw(knn1)

SLM.ES <- lagsarlm(price~-1+intercept1+intercept2+intercept3+intercept4+
yrbuilt1+yrbuilt2+yrbuilt3+yrbuilt4+
TLA1+TLA2+TLA3+TLA4+
baths1+baths2+baths3+baths4+
halfbaths1+halfbaths2+halfbaths3+halfbaths4+
garagesqft1+garagesqft2+garagesqft3+garagesqft4+
lotsize1+lotsize2+lotsize3+lotsize4,
data=newsubhouse, listw_10nn_dates1,
type="lag", tol.solve= 1.58119e-19)
summary(SLM.ES)

options("scipen"=100)

f <- price ~ yrbuilt1+yrbuilt2+yrbuilt3+yrbuilt4+
TLA1+TLA2+TLA3+TLA4+
baths1+baths2+baths3+baths4+
halfbaths1+halfbaths2+halfbaths3+halfbaths4+
garagesqft1+garagesqft2+garagesqft3+garagesqft4+
lotsize1+lotsize2+lotsize3+lotsize4

SDM.ES <- lagsarlm(price~-1+intercept1+intercept2+intercept3+intercept4+
```

```
yrbuilt1+yrbuilt2+yrbuilt3+yrbuilt4+
TLA1+TLA2+TLA3+TLA4+
baths1+baths2+baths3+baths4+
halfbaths1+halfbaths2+halfbaths3+halfbaths4+
garagesqft1+garagesqft2+garagesqft3+garagesqft4+
lotsize1+lotsize2+lotsize3+lotsize4,
data=newsubhouse, listw_10nn_dates1,
Durbin=as.formula(delete.response(terms(f))),
type="mixed", tol.solve= 1.58119e-19)
summary(SDM.ES)

SEM.ES <- errorsarlm(price~-1+intercept1+intercept2+intercept3+intercept4+
yrbuilt1+yrbuilt2+yrbuilt3+yrbuilt4+
TLA1+TLA2+TLA3+TLA4+
baths1+baths2+baths3+baths4+
halfbaths1+halfbaths2+halfbaths3+halfbaths4+
garagesqft1+garagesqft2+garagesqft3+garagesqft4+
lotsize1+lotsize2+lotsize3+lotsize4,
data=newsubhouse, listw_10nn_dates1,
method="eigen", quiet=FALSE,tol.solve= 1.58119e-19)
summary(SEM.ES)
```

مراجع

[۱] محمدزاده م.، (۱۳۹۱)، **"آمار فضایی"** چاپ اول، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

[2] Akaike, H. (1976). An information criteron (AIC), **Mathematical Sciences**, 14, 5-7.

[3] Andreano, M.S., Benedetti, R. & Postiglione, P. (2016). Spatial regimes in regional European growth: an iterated spatially weighted regression approach, **Quality & Quantity**, 51(6), 2665-2684.

[4] Anselin, L. (1980). Estimation methods for spatial autoregressive structures, **Regional Science Dissertation and Monograph Series, Cornel University, Ithaca, NY**.

[5] Anselin, L. (1988a). **Spatial Econometrics: Methods and Models**, Vol. 4, Springer Science & Business Media. London.

[6] Anselin, L. (1988b). Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity, **Geographical Analysis**, 20(1), 1-17.

[7] Anselin, L., Florax, R. (Eds.) (1995). **New Directions in Spatial Econometrics**, Springer, New York.

[8] Anselin, L. (1999). **Spatial Econometrics**, Bruton Center, School of Social Sciences, University of Texas at Dallas, Richardson, TX 75083-0688.

[9] Anselin, L. (2010). Thirty years of spatial econometrics, **Papers in Regional Science**, 89, 3–25.

[10] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems, **Journal of the Royal Statistical Society, Series B (Methodological)**, 36(2), 192-236.

[11] Bille, A.G., Benedetti, R. & Postiglione, P. (2017). A two-step approach to account for unobserved spatial heterogeneity, **Spatial Economic Analysis**, 4, 81-91.

[12] Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, A. & Blanchet, G. (2015). R Package 'spdep', ftp://garr.tucows.com/mirrors/CRAN/web/packages/spdep/spdep.pdf

[13] Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates, **Biometrika**, 71(2), 353-360.

[14] Box, G.E. & Cox, D.R. (1964). An analysis of transformations, **Journal of the Royal Statistical Scociety, Series B (Methodological)**, 26(2), 211-252.

[15] Brunsdon, C., Fotheringham, A.S. & Charlton, M.E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity, **Geographical Analysis**, 28(4), 281-298.

[16] Breusch, T.S. & Pagan, A.R. (1979). A simple test for heteroscedasticity and random coefficient variation, **Econometrica: Journal of the Econometric Society**, 47(5), 1287-1294.

[17] Burridge, P. (1980). On the clifford test for spatial correlation, **Journal of the Royal Statistical Society, Series B (Methodological)**, 42, 107-108.

[18] Casetti, E. (1972). Generating models by the expansion method: applications to geographical research, **Geographical Analysis**, 4, 81-91.

[19] Clayton, D. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, **Biometrics**, 43(3), 671-681.

[20] Clayton, D. & Bernardinelli, L. (1992). Bayesian methods for mapping disease risk, **Geographical and Environmental Epidemiology: Methods for Small Area Studies**, 205-220.

[21] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots, **Journal of the American Statistical Association**, 74, 829-836.

[22] Cleveland, W.S. & Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting, **Journal of the American Statistical Association**, 83, 596–610.

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

[24] Dubin, R.A. (1992). Spatial autocorrelation and neighborhood quality, **Regional Science and Urban Economics**, 22, 433–452.

[25] Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2003). **Geographically Weighted Regression: The Analysis of Spatially Varying Relationship**, John Wiley & Sons, New York.

[26] Gollini, I., Lu, B., Charlton, M., Brunsdon, C. & Harris, P. (2013). GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models, arXiv:1306.0413 [stat.AP].

[27] Haining, R. (1990). **Spatial Data Analysis in the Social and Environmantal Sciences**, Cambridge University Press, Combridge.

[28] Hurvich, C.M., Simonoff, J.S. & Tsai, C.L. (1998). Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion, **Journal of the Royal Statistical Society, Series B (Statistical Methodology)**, 60(2), 271-293.

[29] Ibragimov, R. & Muller, U.K. (2010). T-statistic based correlation and heterogeneity robust inference, **Journal of Business and Economic Statistics**, 28(4), 453-468.

[30] Journel, A.G. & Huijbregts, C.J. (1978). **Mining Geostatistics**, Academic Press, London.

[31] Kelejian, H.H. & Prucha, I.R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model, **International Economic Review**, 40(2), 509-533.

[32] Kelejian, H.H. & Prucha, I.R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances, **Journal of Econometrics**, 157(1), 53-67.

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

[34] Matheron, G. (1971). Random sets theory and its application to stereology, **Journal of Microscopy**, 95, 15-23.

[35] Million, E. (2007). The hadamard product, **Course Notes**, 3-6.

[36] Mollie, A. & Richardson, S. (1991). Empirical Bayes estimates of cancer mortality rates using spatial models, **Statistics in Medicine**, 10(1),95-112.

[37] Montgomery, D.C., Peck, E.A. & Vining, G.G. (2012). **Introduction to Linear Regression Analysis**, Vol. 821, John Wiley & Sons, New York.

[38] Moran, P.A. (1948). Some theorems on time series II: the significance of the serial correlation coefficient, **Biometrika**, 35, 255-260.

[39] Ord, K. (1975). Estimation methods for models of spatial interaction, **Journal of the American Statistical Association**, 70, 120-126.

[40] Paelinck, J.H.P. & Klaassen, L.L.H. (1979). **Spatial Econometrics** , Vol. 1, Saxon House.

[41] Postiglione, P., Andreano, M.S. & Benedetti, R. (2013). Using constrained optimization for the identification of convergence clubs, **Computational Economics**, 42, 151–174.

[42] Pede, V.O., Florax, R.J.G.M. & Lambert, D.M. (2014). Spatial econometric STAR models: Lagrange multiplier tests, Monte Carlo simulations and an empirical application, **Regional Science and Urban Economics**, 49, 118–128.

[43] Polzehl, J. & Spokoiny, V. (2000). Adaptive weights smoothing with applications to image restoration, **Journal of the Royal Statistical Society, Series B (Statistical Methodology)**, 62, 335-354.

[44] Polzehl, J. & Spokoiny, V. (2006). Propagation-separation approach for local likelihood estimation, **Probability Theory and Related Fields**, 135, 335–362.

[45] R Core Team. (2018). R: a language and environment for statistical computing, **R Foundation for Statistical Computing**, Vienna, Austria.

[46] Smith, T.E. (2016). **Notebook on Spatial Data Analysis**, Lecture Note, http://www.seas.upenn.edu/ ese502/.

[47] 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.

[48] Wheeler, D.C. & Páez, A. (2010). Geographically weighted regression, **In Handbook of Applied Spatial Analysis, Springer, Berlin, Heidelberg**, 461-486.

[49] Whittle, P. (1954). On stationary processes in the plane, **Biometrika**, 41, 434-449.

واژه‌نامه فارسی به انگلیسی

الف

آزمون ضریب لاگرانژ تأخیر فضایی ..... Lagrange multiplier lag test

آزمون ضریب لاگرانژ خطای فضایی ..... Lagrange multiplier error test

آزمون نسبت درستنمایی ..... Likelihood ratio test

اعتبارسنجی متقابل ..... Cross-validation

آماره بریچ‌ـ پیگان ..... Breusch-Pagan statistic

آماره موران I ..... Moran I statistic

ایستای ذاتی ..... Intrinsic stationary

ایستای قوی ..... Strong stationary

ایستای مرتبه دوم ..... Second order stationary

ب

برآورد درستنمایی ماکسیمم ..... Maximum likelihood estimation

بزرگ‌مقیاس ..... Large scale

پ

پارامتر حافظه ..... Memory parameter

خ

خودهمبستگی فضایی ..... Spatial autocorrelation

د

داده‌های الگونقطه‌ای ..... Point pattern data

داده‌های زمین‌آماری ..... Geostatistical data

داده‌های فضایی ..... Spatial data

داده‌های مشبکه‌ای ..... Lattice data

ر

رژیم‌های فضایی ..... Spatial regimes

رگرسیون موزون جغرافیایی ..... Geographical weighted regression

رگرسیون موزون موضعی ..... Locally weighted regression

رگرسیون موزون موضعی تکراری ..... Iterative geographically weighted regression
روند ..... Trend

ض

ضرب هادامارد ..... Hadamard product

ک

کمترین توان‌های دوم تعمیم‌یافته ..... Generalized least squares
کمترین توان‌های دوم معمولی ..... Ordinary least square
کمترین توان‌های دوم موزون ..... Weighted least squares
کمترین توان‌های دوم موزون موضعی ..... Locally Weighted Least Squares
k ـ نزدیکترین همسایگی ..... k-nearest neighbour
کوچک‌مقیاس ..... Small scale

م

ماتریس وزن فضایی ..... Spatial weight matrix
مجاورت رأسی ..... Vertex contiguity
مجاورت شعاعی ..... Radius contiguity
مجاورت ضلعی ..... Edge contiguity
مدل اتورگرسیو شرطی ..... Conditional autoregressive model
مدل‌های اتورگرسیو فضایی ..... Spatial autoregressive models
مدل اتورگرسیو فضایی با خطاهای اتورگرسیو Spatial autoregressive model with autoregressive disturbances
مدل اتورگرسیو هم‌زمان ..... Simultaneously autoregressive model
مدل تأخیر فضایی ..... Spatial lag model
مدل ترکیبی فضایی ..... Spatial autoregressive combined model
مدل خطای فضایی ..... Spatial error model
مدل دوربین فضایی ..... Spatial durbin model
مدل‌های اقتصادسنجی فضایی ..... Spatial econometric models
مدل‌های با رژیم فضایی درونی ..... Endogenous spatial regime models
مدل‌های فراموضعی ..... Global models
مدل‌های موضعی ..... Local models
معیار اطلاع آکاییک ..... Akaike information criterion
میانگین هندسی ..... Geometric mean
میدان تصادفی گاوسی ..... Gaussian random field

ن

نامنفرد ..... Nonsingular

ناهمگن ..... Heterogeneous
ناهمگنی ..... Heterogeneity
ناهمگنی فضایی ..... Spatial heterogeneity

و

وابستگی فضایی ..... Spatial dependence
واریانس متجانس ..... Homoscedastic

ه

هسته ثابت ..... Fixed kernel
هسته سازوار ..... Adaptive kernel
همگن ..... Homogenuos
هموارسازی وزن‌های سازوار ..... Adaptive weights smoothing

واژه‌نامه انگلیسی به فارسی

A

هسته سازوار ..... Adaptive kernel
هموارسازی وزن‌های سازوار ..... Adaptive weights smoothing
معیار اطلاع آکاییک ..... Akaike information criterion

B

آماره بریچ_پیگان ..... Breusch-Pagan statistic

C

مدل اتورگرسیو شرطی ..... Conditional autoregressive model
اعتبارسنجی متقابل ..... Cross-validation

E

مجاورت ضلعی ..... Edge contiguity
مدل‌های با رژیم فضایی درونی ..... Endogenous spatial regime models

F

هسته ثابت ..... Fixed kernel

G

میدان تصادفی گاوسی ..... Gaussian random field
کمترین توان‌های دوم تعمیم‌یافته ..... Generalized least squares
رگرسیون موزون جغرافیایی ..... Geographical weighted regression
میانگین هندسی ..... Geometric mean
داده‌های زمین‌آماری ..... Geostatistical data
مدل‌های فراموضعی ..... Global models

H

ضرب هادامارد ..... Hadamard product
ناهمگنی ..... Heterogeneity
ناهمگن ..... Heterogeneous
همگن ..... Homogenuos
واریانس متجانس ..... Homoscedastic

I

ایستای ذاتی ..... Intrinsic stationary
رگرسیون موزون موضعی تکراری ..... Iterative geographically weighted regression

K

k ـ نزدیکترین همسایگی ..... k-nearest neighbour

L

آزمون ضریب لاگرانژ خطای فضایی ..... Lagrange multiplier error test
آزمون ضریب لاگرانژ تأخیر فضایی ..... Lagrange multiplier lag test
بزرگ‌مقیاس ..... Large scale
داده‌های مشبکه‌ای ..... Lattice data
آزمون نسبت درستنمایی ..... Likelihood ratio test
مدل‌های موضعی ..... Local models
کمترین توان‌های دوم موزون موضعی ..... Locally Weighted Least Squares
رگرسیون موزون موضعی ..... Locally weighted regression

M

برآورد درستنمایی ماکسیمم ..... Maximum likelihood estimation
پارامتر حافظه ..... Memory parameter
آماره موران I ..... Moran I statistic

N

نامنفرد ..... Nonsingular

O

کمترین توان‌های دوم معمولی ..... Ordinary least square

P

داده‌های الگونقطه‌ای ..... Point pattern data

R

مجاورت شعاعی ..... Radius contiguity

S

ایستای مرتبه دوم ..... Second order stationary
مدل ترکیبی فضایی ..... Spatial autoregressive combined model
مدل اتورگرسیو هم‌زمان ..... Simultaneously autoregressive model
مدل اتورگرسیو فضایی با خطاهای اتورگرسیو Spatial autoregressive model with autoregressive disturbances
کوچک‌مقیاس ..... Small scale
خودهمبستگی فضایی ..... Spatial autocorrelation
مدل‌های اتورگرسیو فضایی ..... Spatial autoregressive models
داده‌های فضایی ..... Spatial data
وابستگی فضایی ..... Spatial dependence
مدل دوربین فضایی ..... Spatial durbin model

مدل‌های اقتصادسنجی فضایی ..... Spatial econometric models
مدل خطای فضایی ..... Spatial error model
ناهمگنی فضایی ..... Spatial heterogeneity
مدل تأخیر فضایی ..... Spatial lag model
رژیم‌های فضایی ..... Spatial regimes
ماتریس وزن فضایی ..... Spatial weight matrix
ایستای قوی ..... Strong stationary

T

روند ..... Trend

V

مجاورت رأسی ..... Vertex contiguity

W

کمترین توان‌های دوم موزون ..... Weighted least squares

Abstract

Many experimental studies need to use variables that are influenced by their geographical locations. In such cases, to consider the spatial structure of data are essential and ignoring this structure results in to lose some relevant information. To deal with such data, the classical linear regression model is not efficient. Indeed, the spatial structure leads to the violation of homogeneity and uncorrelation assumptions. Hence, we should use some appropriate spatial models. To have sufficient flexibility, we need a model that considers both spatial heterogeneity and dependency, simultaneously. However, the real heterogeneity and dependency structure of data is usually unknown. Furthermore, there is no available prior information about this structure.

In this thesis, we use an iterative locally weighted regression approach to determine homogenous spatial regions, named as spatial regimes. Next, we apply these regimes in a class of spatial autoregressive models, known as spatial econometric models. We compare the proposed models with the linear regression model and classical spatial econometric models based on the Akaike information criterion. The results show a better performance of econometric models with the endogenous spatial regimes.

Keywords: Spatial heterogeneity, Spatial dependence, Spatial regimes, Spatial econometrics, Endogenous spatial regime econometric models.

Shahrood University of Technology

Faculty of Mathematical Sciences

MSc Thesis in: Statistics

Spatial Heterogeneity in Production Price Models by Using an Iterative Locally Weighted Regression Approach

By: Roya Payrooliya

Supervisor

Hossein Baghishani

January 2019