میدان‌های تصادفی مارکوفی گاوسی و کاربرد آن در تحلیل داده‌های طرح آمارگیری از قیمت و اجاره مسکن

مجری طرح

فیروزه ریواز

همکاران طرح

دکتر مجید جعفری خالدی

فرشته حق‌شناس تجن‌گوکه

گروه پژوهشی طرح‌های فنی و روش‌های آماری

پژوهشکده‌ی آمار

۱۳۹۰

پیشگفتار

در برخی از طرح‌های نمونه‌گیری، داده‌ها به گونه‌ای جمع‌آوری می‌شوند که اطلاعات مکان یا ناحیه‌ای مشخص مانند استان، مناطق، نواحی یا بلوک‌های شهری را منعکس کنند. این داده‌ها، علاوه بر این‌که معرف اطلاعات ناحیه‌ای مشخص هستند، بین آن‌ها وابستگی نیز وجود دارد. لذا به کارگیری روش‌های معمول تحلیل برای این‌گونه داده‌ها که داده‌های فضایی مشبّکه‌ای نامیده می‌شوند، و در آن‌ها فرض بر استقلال مشاهدات است، مناسب نمی‌باشد. از این رو، برای مدل‌بندی چنین داده‌هایی عموماً از یک میدان تصادفی مارکوفی گاوسی که یک خانواده مهم از توزیع‌هاست، استفاده می‌شود. یکی از طرح‌هایی که در حال حاضر در مرکز آمار ایران در حال اجرا است و اطلاعات متوسط اجاره‌بها در سطح مناطق شهرداری در مراکز استان‌ها جمع‌آوری می‌شود، طرح آمارگیری از قیمت و اجاره‌ی مسکن است. در این طرح پژوهشی با استفاده از طرح آمارگیری از قیمت و اجاره‌ی مسکن، عملکرد دو روش بسامدی و بیزی در تحلیل داده‌های فضایی مشبکه‌ای مورد بررسی قرار می‌گیرد.

پژوهشکده‌ی آمار با توجه به رسالت خود در زمینه‌ی اجرای طرح‌های پژوهشی با هدف تجزیه و تحلیل آمار و اطلاعات، اجرای طرح مذکور را در دستور کار خود قرار داده است. در گروه مطالعاتی طرح مذکور، خانم دکتر فیروزه ریواز عضو هیأت علمی دانشگاه شهید بهشتی به عنوان مجری طرح، آقای دکتر مجید جعفری خالدی و خانم فرشته حق‌شناس تجن‌گوکه به‌عنوان همکاران طرح، عضویت داشتند که بدین‌وسیله از زحمات یکایک این افراد تشکر و قدردانی می‌شود. داوری طرح نیز به عهده‌ی جناب آقای دکتر فرید روحانی عضو هیأت علمی دانشگاه شهید بهشتی بوده است که بدین وسیله از ایشان تشکر و قدردانی می‌شود.

از خوانندگان محترم تقاضا می‌شود، نظرات اصلاحی خود در ارتباط با محتوای مجموعه‌ی حاضر را به گروه پژوهشی طرح‌های فنی و روش‌های آماری منعکس نمایند.

گروه پژوهشی طرح‌های فنی و روش‌های آماری

پیشگفتار

بخش عمده‌ای از داده‌هایی که توسط مراکز و سازمان‌های مختلف جمع‌آوری می‌شوند، داده‌هایی هستند که منعکس کننده اطلاعات مکان یا ناحیه‌ای مشخص مانند استان، مناطق، نواحی یا بلوک‌های شهری هستند. این داده‌ها علاوه بر این‌که معرف اطلاعات ناحیه‌ای مشخص هستند، اغلب بین آنها وابستگی‌ای از جنس مکان قرارگیری آنها در فضای مورد مطالعه نیز وجود دارد. به عبارت دیگر، مشاهداتی که در همسایگی یکدیگر قرار دارند، مشابه‌تر هستند. لذا به کارگیری روش‌های معمول تحلیل، که در آنها فرض بر استقلال مشاهدات است، برای تحلیل این‌گونه داده‌ها که داده‌های فضایی مشبکه‌ای نامیده می‌شوند، مناسب نمی‌باشد.

دو موضوع مهم در تحلیل داده‌های فضایی مشبکه‌ای مدل‌بندی ساختار همبستگی بین مشاهدات و تهیه نقشه هموار از کمیت مورد مطالعه است. برای این منظور، عموماً با در نظر گرفتن داده‌ها به عنوان تحقق‌هایی از یک میدان تصادفی مارکوفی گاوسی به تحلیل آنهاپرداخته می‌شود. به عبارت دیگر، با در نظر گرفتن یک ساختار مارکوفی برای بیان همبستگی بین مشاهدات و فرض نرمالیتی برای توزیع‌های تمام‌شرطی کمیت مورد مطالعه، داده‌ها مورد تحلیل قرار می‌گیرند.

دو روش معمول برای انجام این مهم، روش بسامدی و بیزی است. در روش بسامدی تحلیل، تابع درستمایی پارامترهای مدل تشکیل شده و برآوردهای ماکسیمم درستنمایی آنها تعیین می‌شود.

این برآوردها معمولاً از خواص مجانبی مناسبی برخوردار می‌باشند. اما برای اندازه نمونه‌های کوچک، کارایی آن‌ها مورد سؤال است. ضمن این‌که اگر در مورد پارامترها اطلاعی از قبل در اختیار باشد، نمی‌توان از آن در انجام استنباط‌ها استفاده کرد و کارایی برآوردها و پیشگوها را افزایش داد. بنابراین استفاده از روش بیزی می‌تواند رهگشا باشد. در روش بیزی تحلیل، با در نظر گرفتن عدم حتمیت پارامترها به کمک توزیع پیشینی و ترکیب آن با تابع درستنمایی، توزیع پسینی پارامترها بدست می‌آید. سپس با متوسط گیری از توزیع پیشگو با وزن‌های پسینی، توزیع پیشگوی بیزی تعیین می‌شود. برای انجام تحلیل بیزی، ابتدا توزیع‌های پیشینی عینی برای پارامترهای این کلاس از مدل‌ها تعیین می‌گردد. برای این منظور، با به‌کارگیری رهیافت ارائه شده در برگر و همکاران (۲۰۰۱) برای مدل‌های زمین آماری گاوسی، پیشین‌های مرجع و جفریز بدست آمده و سپس خواص این پیشین‌ها در ارتباط با توزیع پسینی حاصل از آن‌ها بررسی می‌شود. سپس یک پیشین بیز تجربی مبتنی بر ماکسیمم‌سازی تابع درستنمایی حاشیه‌ای با استفاده از الگوریتم EM ارائه می‌شود. در ادامه خواص بسامدی این پیشین‌ها در یک مطالعه شبیه‌سازی مورد بررسی قرار می‌گیرد.

در طرح حاضر، ضمن معرفی میدان‌های تصادفی مارکوفی گاوسی و ارائه ویژگی‌های آن‌ها، نحوه مدل‌بندی داده‌های فضایی مشبکه‌ای با استفاده از آن بیان می‌شود. سپس با به‌کارگیری روش بسامدی و بیزی، به تحلیل داده‌های فضایی مشبکه‌ای پرداخته می‌شود. در ادامه با استفاده از داده‌های طرح قیمت و اجاره مسکن، عملکرد این دو روش در تحلیل داده‌های فضایی مشبکه‌ای مورد مطالعه قرار می‌گیرد.

واژه‌های کلیدی : توزیع گاوسی، میدان تصادفی، خاصیت مارکوفی، داده‌های مشبکه‌ای، روش ماکسیمم درستنمایی، روش بیزی، پیشین مرجع، پیشین بیز تجربی.

فهرست مندرجات

فصل ۱

کلیات

۱.۱ مقدمه

افزایش روز افزون مطالعاتی که در آن با داده‌هایی وابسته به مکان روبرو هستیم، پژوهشگران را بر آن داشته تا شیوه‌های نوینی برای تحلیل چنین داده‌هایی که داده‌های فضایی نامیده می‌شوند، به‌کار برند.

معمولاً برای مدل‌بندی داده‌های فضایی، از یک میدان تصادفی استفاده می‌شود. یک میدان تصادفی، مجموعه‌ای از متغیرهای تصادفی مانند $Z(\cdot) = \{Z(s); s \in D\}$ است که در آن s موقعیت فضایی متعلق به مجموعهٔ اندیس گذار D، $D \subseteq R^d$، $d \geq ۱$، است.

به‌طور کلی، سه نوع داده‌ی فضایی وجود دارد. اگر موقعیت داده‌ها به صورت پیوسته روی ناحیهٔ مورد مطالعه D تغییر کند، آن‌گاه داده‌ها، زمین‌آماری[۱] نامیده می‌شوند. داده‌های زمین‌آماری می‌توانند پیوسته یا گسسته باشند. به عنوان مثال، میزان بارندگی در یک ناحیهٔ جغرافیایی مشخص

[۱] Geostatistics

مانند شهر تهران، برای یک ماه، داده‌های زمین‌آمار حاصل از یک متغیر پیوسته هستند. در این حالت مدل‌های گاوسی و تعمیم‌هایی از آن همچون مدل گاوسی تبدیل‌یافته (دی‌الیویرا و همکاران، ۱۹۹۷) یا مدل چوله گاوسی (کیم و مالیک، ۲۰۰۴؛ ژانگ و ال شراوی، ۲۰۱۰) برای تحلیل داده‌ها مورد استفاده قرار می‌گیرند. اگر این داده‌ها به‌صورت بارش یا عدم بارش ثبت شده باشند، آن‌گاه داده‌های زمین‌آماری حاصل از یک متغیر گسسته هستند. در این‌گونه مسائل مدل‌های زمین آمار خطی تعمیم‌یافته می‌تواند برای تحلیل چنین متغیرهایی بکار رود (کریستینسن و همکاران، ۲۰۰۶).

در صورتی‌که داده‌ها مربوط به مکان‌های ناحیه‌ای باشند، داده‌های مشبکه‌ای[۲] نامیده می‌شوند. در این حالت D زیرمجموعه‌ای ثابت و شمارا از R^d است که می‌تواند به‌صورت منظم یا نامنظم افراز شده باشد. متغیر مورد مطالعه مشابه با حالت زمین‌آمار نیز می‌تواند پیوسته یا گسسته باشد. به عنوان مثال، تعداد بیماران مبتلا به سرطان لب و دهان در استان‌های مختلف کشور، داده‌های مشبکه‌ای حاصل از یک متغیر تصادفی گسسته هستند. همچنین داده‌های مربوط به متوسط نرخ اجاره مسکن در هر یک از مناطق شهر تهران، داده‌های حاصل از یک متغیر تصادفی پیوسته هستند.

اگر موقعیت‌های فضایی، خود متغیر تصادفی باشند آنگاه داده‌ها الگوی نقطه‌ای[۳] نامیده می‌شوند. در این طرح، تحلیل داده‌های فضایی مشبکه‌ای حاصل از یک متغیر تصادفی پیوسته مورد نظر است که به اختصار داده‌های مشبکه‌ای پیوسته نامیده می‌شود.

اصطلاح مشبکه اولین بار توسط بیرخوف (۱۹۶۷) و سپس هامرسلی و مازارینو (۱۹۸۳) در متون ریاضی ارائه گردید و توسط بیسگ (۱۹۷۴) در تحلیل داده‌های فضایی به کار گرفته شد. بیسگ (۱۹۷۴) مدل‌های اتوگاوسی، اتولوژستیک، اتوپواسون مبتنی بر میدان‌های تصادفی مارکوفی را برای تحلیل داده‌های مشبکه‌ای به کار برد.

---

[۲] Lattice Data

[۳] Point Pattern

در بین این مدل‌ها، مدل اتوگاوسی با خاصیت مارکوفی که به میدان تصادفی مارکوفی گاوسی (GMRF) معروف است، به دلیل فرم تابع درستنمایی مربوط به آن و به‌علاوه ویژگی‌های مناسبی که توزیع نرمال دارد، بسیار مورد توجه قرار دارد. $GMRF$ها برای مدل‌بندی داده‌های مشبکه‌ای به دو صورت بکار برده می‌شوند:

۱) به‌عنوان مدلی برای توزیع نمونه‌گیری داده‌های مشاهده شده (بل و برملینگ، ۲۰۰۰؛ کرسی و همکاران، ۲۰۰۵؛ فریرا و دی‌الیویرا، ۲۰۰۷، سانگ و همکاران، ۲۰۰۸؛ کرسی و ورزلن، ۲۰۰۸؛ دی‌الیویرا و فریرا، ۲۰۱۱).

۲) به‌عنوان مدلی برای توزیع‌های پیشینی فرایندهای پنهان یا اثرات تصادفی با تغییرات فضایی (بیسگ و همکاران ، ۱۹۹۱؛ کلایتون و کالدر، ۱۹۸۷؛ سون و همکاران، ۱۹۹۹، سین و کرسی، ۲۰۰۷).

یکی از روش‌های معمول تحلیل $GMRF$ها، روش ماکسیمم درستنمایی است (کرسی و چن، ۱۹۸۹ ؛ ریچاردسون و همکاران، ۱۹۹۲؛ درایدن و همکاران، ۲۰۰۲؛ پتیت و همکاران، ۲۰۰۳؛ کرسی و همکاران، ۲۰۰۵). اغلب خواص شناخته شده در ارتباط با تحلیل‌های ماکسیمم درستنمایی مبتنی بر رفتار مجانبی آن‌ها است. اما هنگامی که حجم نمونه کوچک است، اطلاع چندانی در مورد رفتار تحلیل‌ها نمی‌توان ارائه نمود. برخلاف روش ماکسیمم درستنمایی، رهیافت بیزی به‌دلیل ترکیب اطلاعات درستنمایی و پیشینی، می‌تواند عملکرد مناسبی در صورت تعیین پیشین‌های مناسب ارائه نماید. اما علی‌رغم این موضوع، رهیافت بیزی به‌منظور تحلیل $GMRF$ها، کمتر مورد توجه قرار گرفته است. دلیل اصلی این موضوع انتخاب توزیع پیشین مناسب است. چون در عمل، معمولاً تفسیر ذهنی پارامتر همبستگی فضایی بسیار دشوار است و اغلب اطلاعات قبلی مناسبی در

خصوص این پارامتر در اختیار نمی‌باشد، تعیین توزیع پیشین با چالش روبرو است. برای این منظور، دو روش معمول استفاده از پیشین‌های ناسره و مبهم است. بکارگیری پیشین‌های ناسره همواره تردید ناسره بودن توزیع پسین را به دنبال دارد و استفاده از پیشین‌های مبهم گرچه مساله سره بودن توزیع پسین را حل می‌کند اما استنباط و پیشگویی به شدت به انتخاب ابرپارامترها حساس است. یک راه جایگزین، استفاده از پیشین‌های عینی است. دو نوع معروف از این پیشین‌ها، پیشین جفریز و مرجع است. فریرا و دی‌الیویرا (۲۰۰۷) با استفاده از این پیشین‌ها به تحلیل بیزی یک مدل $GMRF$ پرداختند. یکی از نقاط ضعف پیشین‌های عینی، دشواری تعیین آن‌هاست به‌ویژه زمانی که تعداد پارامترهای مدل افزایش یابد. به‌علاوه بررسی سره بودن توزیع‌های پسینی متناظر با این پیشین‌ها نیز راحت نیست. در این خصوص می‌توان از پیشین‌های بیز تجربی استفاده کرد که دیگر سره بودن توزیع پسین لازم به بررسی نیست.

در طرح پژوهشی حاضر، ضمن ارائه تعاریف و مفاهیم لازم در فصل اول، یک مدل $GMRF$ در فصل دوم معرفی شده و خواص آن مورد بررسی قرار می‌گیرد. سپس تحلیل آن مبتنی بر روش ماکسیمم درستنمایی ارائه می‌شود. در واقع، برآورد پارامترهای مدل با ماکسیمم نمودن تابع درستنمایی تعیین می‌شود. در ادامه با جایگذاری برآوردهای ماکسیمم درستنمایی در توزیع پیشگو، پیشگوی فضایی به‌دست می‌آید. در فصل سوم روش بیزی مبتنی بر پیشین‌های عینی و بیز تجربی، مورد بررسی قرار می‌گیرد. برای این منظور، ابتدا پیشین‌های مرجع و جفریز برای یک میدان تصادفی گاوسی و مارکوفی گاوسی محاسبه شده و سپس خواص توزیع پسین متناظر با این پیشین‌ها بررسی می‌شود. از جمله سره بودن توزیع پسین مورد بررسی قرار می‌گیرد. در ادامه یک پیشین بیز تجربی مبتنی بر ماکسیمم‌سازی تابع درستنمایی حاشیه‌ای با استفاده از الگوریتم EM ارائه می‌شود. در یک مطالعه شبیه‌سازی عملکرد توزیع‌های پیشینی ارائه شده مورد بررسی قرار می‌گیرد.

سپس یک الگوریتم $MCMC$ برای نمونه‌گیری از توزیع‌های پسینی پارامترهای مدل شرح داده می‌شود. در فصل چهارم، داده‌های طرح قیمت و اجاره مسکن مربوط به شهر تهران با به‌کارگیری یک مدل $GMRF$، مدل‌بندی می‌شوند. برای این داده‌ها روش بیزی و بسامدی ارائه شده و مورد بررسی و مقایسه قرار می‌گیرند.

۲.۱ تعاریف و مفاهیم اولیه

در این بخش برخی از تعاریف و مفاهیم لازم، برای تحلیل داده‌های مشبکه‌ای پیوسته بیان می‌شود.

۱.۲.۱ میدان تصادفی گاوسی

برای میدان تصادفی $Z(\cdot)$ عبارت

$$F_{Z(s_1),\dots,Z(s_k)}\left(z_1,\dots,z_k\right)=P\left(Z\left(s_1\right)\leq z_1,\dots,Z\left(s_k\right)\leq z_k\right),\ k\geq 1 \qquad (۱.۲.۱)$$

تابع توزیع متناهی بعد و تابع چگالی متناظر با آن در صورت وجود تابع چگالی متناهی بعد میدان تصادفی نامیده می‌شود.

هرگاه همه توابع چگالی توأم متناهی بعد یک میدان تصادفی، گاوسی چند متغیره باشند، میدان تصادفی گاوسی نامیده می‌شود. توزیع یک میدان تصادفی گاوسی بوسیله تابع میانگین و تابع کواریانس بطور کامل معین می‌شود.

۲.۲.۱ خاصیت مارکوفی

فرایند تصادفی $\{X_t, t \geq 0\}$ را در نظر بگیرید. فرایند را مارکوفی گویند هرگاه به ازای هر مجموعهٔ برل اندازه‌پذیر A، $n \geq 0$ و $t_1 < t_2 < \cdots < t_n < t$ داشته باشیم:

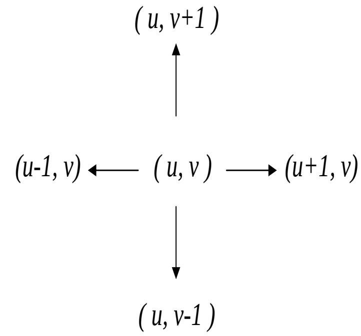
$$P\Big(X(t) \in A | X(t_1) = x_1, X(t_2) = x_2, \cdots, X(t_n) = x_n\Big) = P\Big(X(t) \in A | X(t_n) = x_n\Big). \tag{۲.۲.۱}$$

تعریف فوق بیان می‌دارد که تنها حالت کنونی، اطلاعاتی را در مورد رفتار آینده فرایند ارائه می‌کند. در واقع اطلاعات گذشته فرایند، تأثیری در مورد رفتار آینده فرایند ندارد. بدین ترتیب فرایند دارای خاصیت بی‌حافظگی نامیده می‌شود. اکنون به معرفی خاصیت مارکوفی در داده‌های فضایی می‌پردازیم. این ویژگی همان خاصیت بی‌حافظگی مربوط به مکان‌های دور است. به عبارت دیگر در صورتی که مقدار مکان iام فقط به مقادیر متناظر با موقعیت‌هایی که در همسایگی این مکان قرار دارند، بستگی داشته باشد، گوییم خاصیت مارکوفی برقرار است. در این حالت بیشترین اطلاعات در مورد هر مکان دلخواه توسط مکان‌های نزدیک و همسایه آن ارائه می‌شود.

مثال ۱.۲.۱ : مشبکه‌ای در فضای دوبعدی به‌صورت

$$D = \{s = (u, v)\ ;\ u, v = \cdots, -1, 0, 1, \cdots\}$$

در نظر بگیرید. فرض کنید متغیرها در هر مکان به نزدیکترین موقعیت‌های مجاور به‌صورت شکل ۱.۲.۱ بستگی داشته باشند.



شکل ۱.۲.۱: چها همسایگی مرتبهٔ اول مکان (u, v)

در این صورت مدل احتمال مارکوفی مرتبه اول برقرار است. یعنی

$$P\Big(Z(s) \in A | Z(s_1) = z_1, \cdots, Z(s_n) = z_n\Big) =$$
$$P\Big(Z(u,v) \in A | Z(u, v+1) = z_1, Z(u+1, v) = z_2, Z(u, v-1) = z_3, Z(u-1, v) = z_4\Big).$$

در بخش‌های بعد میدان‌های تصادفی مارکوفی به تفصیل مورد بررسی قرار می‌گیرند.

۳.۲.۱ ساختار وابستگی فضایی

همان طور که در سریهای زمانی، مقدار کمیت مورد مطالعه در آینده با مشاهدات موجود در حال و گذشته نزدیک ارتباط داده می‌شود و مقدار آن پیش‌بینی می‌شود، در میدان‌های تصادفی فضایی مشبکه‌ای نیز مقدار متناظر با کمیت مورد مطالعه در هر ناحیه بر اساس داده‌های همبسته مجاور با آن یعنی مشاهدات واقع در نواحی همسایه، مدل‌بندی و پیشگویی می‌شود. بنابراین، اولین گام در تحلیل داده‌های مشبکه‌ای تعیین ساختار وابستگی یا ارتباط فضایی است که در آن مفهوم همسایگی نقش مهمی ایفا می‌کند.

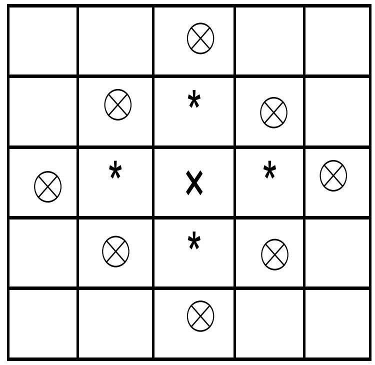
تعریف ۱.۲.۱ : مجموعه مکان‌هایی که بیشترین اطلاع را در خصوص ناحیه دلخواه iام ارائه می‌کنند، مجموعه همسایگی ناحیه iام نامیده شده و با $N(i)$ نمایش داده می‌شود. به عبارت دیگر برای هر ناحیه دلخواه k، $k \neq i$، $k \in N(i)$ است اگر و فقط اگر $i \in N(k)$ باشد.

همسایگی یک ناحیه بر حسب منظم یا نامنظم بودن مشبکه‌ها، به صورت دو موقعیت هم مرز یا دو ناحیه که در یک فاصله معین از هم قرار دارند، تعریف می‌شود. اغلب، همسایگی یک مکان به صورت تجربی تعیین می‌شود. مثال‌های زیر روش‌های مختلف تعیین همسایگی را نشان می‌دهد.

۱) در مثال مدل‌بندی قیمت مسکن در پاریس (تو و همکاران، ۲۰۰۴)، مشبکه‌ها همان ساختمان‌ها در نظر گرفته شدند و ساختمان‌هایی که در فاصله مشخصی از ساختمان iام قرار داشتند، همسایه آن تلقی شدند. وضعیت قرارگیری ساختمان‌ها در شهر پاریس مثالی از یک مشبکه نامنظم است.

۲) برای بررسی میزان کود شیمیایی در یک زمین کشاورزی که به کرت‌های هم اندازه و منظم تقسیم شده است، همسایگی هر خانه ممکن است توسط همسایه‌هایی از مرتبه اول (کرت‌های مماس با هر کرت) و یا از مرتبه اول و دوم در نظر گرفته شده باشند. در شکل ۲.۲.۱، برای مکان $\times$ همسایه‌های مرتبه اول با علامت $*$ و همسایه‌های مرتبه دوم با علامت $\otimes$ مشخص شده‌اند.

۳) در مدل‌های اقتصادی، همسایگی ممکن است بر اساس تبادلات اقتصادی بین دو ناحیه مشخص شود. به عنوان مثال اگر متغیر مورد مطالعه، تولیدات اقتصادی کشورهای مختلف



شکل ۱.۲.۲: مشبکهٔ منظم

باشد، همسایه ناحیه iام، کشورهایی هستند که با آن مبادلات مثبت تجاری دارند. بنابراین دو کشور همسایه، لزوماً مرز مشترک جغرافیایی ندارند.

بعد از تعیین نوع همسایگی، بیان میزان تأثیر این همسایگی دارای اهمیت است. این تأثیر از طریق انتساب وزن به هر یک از همسایه‌ها بدست می‌آید. در ادامه برخی از روش‌های انتساب وزن همسایگی‌ها ارائه می‌شود.

فرض کنید ناحیه مورد مطالعه به n مشبکه افراز شده باشد. معمولاً برای نشان دادن وزن‌های همسایگی از ماتریس

$$H = (g_{ij}) = \begin{pmatrix} g_{11} & g_{12} & \cdot & \cdot & \cdot & g_{1n} \\ g_{21} & g_{22} & \cdot & \cdot & \cdot & g_{2n} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ g_{n1} & g_{n2} & \cdot & \cdot & \cdot & g_{nn} \end{pmatrix} \tag{۳.۲.۱}$$

استفاده می‌شود که آن را ماتریس همجواری[۴] نیز می‌نامند. درایه‌ی g_{ij}، $i \neq j$، بیان‌گر میزان تأثیر

[۴] Proximity Matrix

ناحیه j روی ناحیه i است.

برای مشخص کردن ماتریس H، می‌توان از ویژگی‌های مختلفی مانند فاصله بین مراکز نواحی، مرز مشترک نواحی، جمعیت هر ناحیه، مساحت نواحی و میزان مبادلات فرهنگی و تجاری نواحی استفاده کرد. یکی از ساده‌ترین انتخاب‌های ممکن برای H، تعریف g_{ij}ها به صورت

$$g_{ij} = \begin{cases} 1, & j \in N(i) \\ 0, & j \notin N(i) \end{cases} \tag{۴.۲.۱}$$

می‌باشد، که $N(i)$ مجموعه همسایه‌های ناحیه i است. ماتریس همجواری H را می‌توان با تعریف درایه‌ها به صورت

$$H = \left(\frac{g_{ij}}{g_{i+}}\right) \tag{۵.۲.۱}$$

که $g_{i+} = \sum_j g_{ij}$ است، به یک ماتریس استاندارد تبدیل نمود. در این صورت H ماتریس ردیف تصادفی[5] است یعنی $H\cdot =$. که در آن نمایانگر برداری n بعدی از مقادیر یک است. علی‌رغم سادگی این روش، وزن‌ها فقط با توجه به تعداد همسایه‌ها و انتساب وزن یکسان به هر همسایه در نظر گرفته می‌شوند که بسیاری از مواقع منطقی به نظر نمی‌رسد. زیرا ممکن است برخی از همسایه‌ها تأثیر بیشتری نسبت به سایرین داشته باشند. راه دیگر برای انتساب وزن‌ها، استفاده از فاصله بین نواحی است. به عبارت دیگر، تعریف تابع نزولی مبتنی بر فاصله بگونه‌ای که به همسایه‌های نزدیک، وزن زیاد و به همسایه‌های دور، وزن کم نسبت دهد. در این حالت g_{ij}ها را می‌توان به صورت (۴.۲.۱) نوشت. **به عبارت دیگر با تعریف ۱ $= g_{ij}$، برای کلیه i و jهایی که در یک فاصله مشخص قرار داشته باشند و ۰ $= g_{ij}$ در غیر این صورت، کلیه داریه‌های ماتریس H، ۰ یا ۱ خواهند بود.**

---

[5] Stochastic Row