خوشه‌بندی چند متغیره داده‌های فضایی بر مبنای مقیاس‌های منطقه‌ای همبستگی فضایی

ناصر طوسی [۱] - ابراهیم زینوند لرستانی [۲]

[۱] آمار، دانشکده علوم، دانشگاه گلستان
[۲] گرواکولوژی، دانشگاه علوم کشاورزی و منابع طبیعی گرگان

چکیده: تاکنون معمولا برای خوشه‌بندی داده‌های فضایی از خوشه‌بندی کلاسیک استفاده شده است، در حالی که خوشه‌بندی کلاسیک فقط بر مبنای اطلاعات چندمتغیره داده‌هاست، لذا برای خوشه‌بندی داده‌های فضایی نیاز به رویکرد و الگوریتمی می‌باشد که علاوه بر استفاده از اطلاعات چندمتغیره داده‌ها، توزیع فضایی داده‌ها را نیز مد نظر قرار دهد. در این مقاله یک روش تقسیم بندی شناخته شده به نام k-means را که با ساختار فضایی داده‌ها همگام‌سازی شده است، مطرح می‌شود. در این راستا برای لحاظ کردن موقعیت مکانی داده‌ها در خوشه‌بندی، از مقیاس‌های منطقه‌ای «خود همبستگی فضایی» و شاخص های موران و ضریب «گری» و آماره‌ی جی عمومی استفاده شده است.

واژه‌های کلیدی: تحلیل فضایی، خوشه‌بندی، شاخص‌های موران و گری، آماره جی عمومی، k-means

۱ مقدمه

فرض کنید n مشاهده در s مکان موجود باشد که از هر واحد p متغیر اندازه گیری شده است. الگوریتم خوشه‌بندی فضایی به دنبال آن است تا هر واحد را به یکی از k خوشه تخصیص دهد. در عمل داده‌ها یی وجود دارد که به موقعیت جغرافیایی وابسته اند. خود همبستگی فضایی برای نشان دادن شباهت و نزدیکی مکان داده‌ها استفاده می‌شود .

وجود خود همبستگی فضایی، بین مکان هایی که در همسایگی هم هستند را می توان با نشانگر خود همبستگی فضایی کل تعیین کرد (Cliff and Ord, 1981). همچنین نیاز باعث شده است که علاوه بر یافتن خود همبستگی کل، خود همبستگی به صورت محلی نیز بدست آورده شود. اگر خود همبستگی کل می تواند خلاصه ای از همبستگی

فضایی کل منطقه مورد مطالعه را ارائه دهد، خود همبستگی محلی، میزان انحراف از همبستگی کل را نشان داده و در یافتن داده‌های پرت مفید می‌باشد (Boots,2002). خوشه‌بندی یکی از مهمترین مباحث آنالیز آماری می‌باشد که در اکثر موارد داده‌ها از هم مستقل فرض می‌شوند، در حالی که ممکن است به لحاظ مکانی به هم وابسته باشند و این وابستگی باید در خوشه‌بندی لحاظ شود.

در بخش دوم تحلیل مدل خودهمبستگی فضایی[۱] بیان می‌شود. در این راستا مفاهیم کلی و شاخص های خودهمبستگی فضایی معرفی می‌شود. در بخش سوم الگوریتم خوشه‌بندی فضایی پیشنهادی که با لحاظ کردن موقعیت مکانی داده‌ها به خوشه‌بندی می‌پردازد، بیان می‌شود. در بخش چهارم داده‌های به کار برده شده معرفی می‌شوند و در بخش نهایی عملکرد الگوریتم بر روی داده‌های واقعی مشاهده می‌شود.

۲ تحلیل مدل خودهمبستگی فضایی

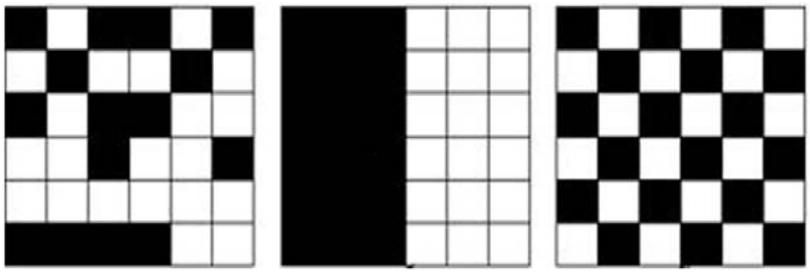
در طبقه بندی الگوهای فضایی، خواه خوشه ای، پراکنده و تصادفی ـ می توان بر چگونگی نظم و ترتیب قرارگیری واحدهای ناحیه ای متمرکز شد. می‌توان مشابهت و نبود مشابهت هر جفت از واحدهای ناحیهای مجاور را اندازه گرفت. وقتی این مشابهت و نبود مشابهت‌ها برای الگوهای فضایی تعیین شود، خود همبستگی فضایی شکل می گیرد (Oldand, 1988). خود همبستگی یا همبستگی سریالی، به ارتباط باقی مانده های معادله رگرسیونی اشاره دارد. به وسیله خود همبستگی، وضعیتی را توصیف میکنیم که درآن هر باقی مانده یا ضریب خطا مرتبط به ضریب های قبلی است (Clark,1986).

خود همبستگی فضایی قوی بدین مفهوم است که ارزش صفات پدیده‌های جغرافیا یی به طور قوی با یکدیگر رابطه دارند (مثبت یا منفی). ضریب ویژگی توزیع پدیده‌های جغرافیایی مجاور، ارتباطات و نظم ظاهری مختلفی دارد که گفته می‌شود دارای ارتباط فضایی ضعیف، قوی و یا دارای الگوی تصادفی می‌باشند (شکل شماره ۱).

مطالعه همبستگی فضایی، پیشنهادهای ضمنی مهمی برای کاربرد تکنیک‌های آماری در تحلیل اطلاعات فضایی در بر دارد. برای اندازه گیری همبستگی فضایی، آماره‌هایی وجود دارد که به ما اجازه می‌دهند با نقاط یا پلی گونها (سطوح نواحی) کار کنیم.

این روشها ممکن است برای اندازه‌گیری تعامل فضایی داده‌های عددی و فاصله‌ای، نسبی به کار روند. بخصوص داده‌های شمارشی پیوسته می‌تواند برای تعامل فضایی در میان پلی گونه‌ای با داده‌های عددی دوتایی استفاده شود. برای داده‌های فاصله‌ای

[۱] Spatial Autocorrelation



شکل ۱: انواع خودهمبستگی فضایی، همبستگی منفی (قاب راست)، همبستگی مثبت (قاب وسط)، عدم همبستگی فضایی (قاب چپ)

نسبی، شاخص موران (Moran's I) و ضریب گری (Gray Ratio) و شاخص محلی G(G-Statistics) به کار می‌روند.

۳ انواع معیارهای تعامل فضایی

مدل‌های متفاوتی برای اندازه‌گیری آماره‌های تعامل فضایی وجود دارد. اگر صفت‌های فضایی یا متغیرهای مورد مطالعه با مقیاس اسمی (Nominal) و دوتایی (Binary) باشند (به عنوان نمونه صفتها فقط دو ارزش ممکن صفر و یک دارند)، پس آماره محاسبات عددی، تعداد اتصالها (Joint Count) می تواند استفاده شود. اگر متغیرهای فضایی اندازه گیری شده، دارای مقیاس فاصله ای یا نسبی باشند، آماره های ارتباط فضایی مناسب شاخص موران (Moran's I) و ضریب گری (Gary Ratio) می‌باشند و گزینه ممکن دیگر، آماره G عمومی (G-Statistic) است.

یک توزیع یا یک الگوی فضایی می تواند در شرایط مختلف از نظر فضایی ناهمگن باشد. برای توصیف ناهمگنی خود همبستگی فضایی، باید بر معیارهایی متکی باشیم که میتوانند خود همبستگی فضایی را در مقیاس محلی کشف کنند. شاخص محلی تمرکز فضایی (Local Indicator of Spatial Association-LISA) و آماره G محلی (Local G-Statistics) برای این هدف مورد استفاده قرار می گیرد.

۱.۳ شاخص موران وضریب گری

شاخص‌های موران و گری[۲] مشخصه‌های مشترکی دارند، اما خواص آماری آنها متفاوت است. اکثر تحلیلگران با شاخص موران موافق‌ترند، که اساساً به خاطر توزیع مشخصاتش، مطلوب تر است (Cliff and Ord.1973, 1981).

[۲] Moran I & Gary Rtio C

هنوز هر دو روش بر مقایسه ارزش‌های همسایگی واحدهای ناحیه‌های متکی هستند. اگر واحدهای ناحیه ای همسایگی در طول ناحیه ارزش‌های مشابهی داشته باشند، آماره ها (مدل‌ها) بر یک خود همبستگی فضایی قوی دلالت داشته‌اند. اگر واحدهای ناحیه‌ای همسایگی ارزشهای خیلی نامشابهی داشته‌باشند، آماره‌ها باید یک خودهمبستگی فضایی منفی خیلی قوی را نشان دهند. به هر حال دو مدل، روش‌های متفاوتی را برای مقایسه ضرایب همسایگی‌ها به کار می گیرند.

۱.۱.۳ شاخص موران

شاخص موران[۳] ، به شر ح ذیل است:

$$I = \frac{N \sum_{i=۱}^{n} \sum_{j=۱}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{(\sum_{i=۱}^{n} \sum_{j=۱}^{n} w_{ij}) \sum_{i=۱}^{n} (x_i - \bar{x})^{۲}} \qquad (۱)$$

که در آن N تعداد کل داده‌ها ، Xi داده مکان i ام، Xj داده مکان j ام، Wij شاخص وزنی که مکان i و j را مرتبط می‌سازد.

ضریب موران بین ۱− تا ۱ متغیر است. ۱− برابر تعامل فضایی منفی و ۱ برابر تعامل فضایی مثبت به کار می‌رود. اگر تعامل فضایی وجود نداشته باشد، ضریب‌های مورد انتظار موران، برابر است با:

$$E(I) = -\frac{۱}{n-۱} \qquad (۲)$$

وقتی شاخص موران محاسبه می‌شود ، ماتریس‌های وزنی فضایی مورد استفاده ماتریس های دوتایی و تصادفی می‌باشند. اگر شاخص دوتایی (زوجی) استفاده شود، W در مخرج کسر اساساً دو برابر مرزهای مشترک در کل ناحیه مورد مطالعه خواهد بود یا J۲. به هر حال، این امکان وجود دارد که انواع دیگری از ماتریس‌های وزنی را به کار ببریم.

۲.۱.۳ ماتریس وزن[4]

یکی از مهمترین ابزار های آمار فضایی ماتریس وزن می‌باشد، چرا که میزان ارتباط فضایی بین داده ها به وسیله ماتریس وزن به داده ها نسبت داده می شود و ما با توجه به ساختار فضایی داده‌ها ماتریس متناسب را روی داده ها اعمال می کنیم. متداولترین

[۳] Moran's I
[4] Weight matrix

حالت‌های ماتریس وزن عبارت است از: ماتریس معکوس فاصله، زمان گردش، فاصله ثابت و K همسایگی مجاور و...

در موضوع مورد مطالعه، ماتریس وزنی که لحاظ شده است، ماتریس وزن، براساس معکوس فاصله می‌باشد به این شکل که با توجه به مختصات داده‌ها فاصله آنها را از هم محاسبه و این فاصله را عکس می‌کنیم تا بدین ترتیب هر چه داده‌ها از هم دورتر شوند وزن کمتری بگیرند و بالعکس

$$W_{ij} = \frac{۱}{d_{ij}}$$

۳.۱.۳ ضریب «گری»[5]

مشابه روش شاخص موران برای اندازه گیری خود همبستگی فضایی، ضریب C گری می‌تواند یک عبارت حاصل ضرب ضربدری را سازگار کند. ضریب «گری» به صورت فرمول زیر است:

$$C = \frac{(n-۱)\sum_i \sum_j W_{ij}(X_i - X_j)^۲}{۲\sum_i \sum_j W_{ij}(X_i - \bar{X})^۲} \qquad (۳)$$

شبیه شاخص موران، ضریب «گری» می‌تواند با هر نوع ماتریس وزنی فضایی به کار رود؛ گرچه عمومی‌ترین آنها، ماتریس‌های دوتایی و تصادفی می‌باشند.

۲.۳ آماره Gعمومی[6]

شاخص محلی دیگر خودهمبستگی فضایی آماره‌ی G عمومی است (Getis and Ord 1992). آماره G عمومی محلی برای هر واحد ناحیه‌ای محاسبه می‌شود و بر این دلالت دارد که چگونه ارزش واحد ناحیه مورد مطالعه مرتبط به ارزش‌های واحدهای ناحیه ای مجاور، از طریق آستانه مسافت (d) تعریف شده می‌باشد. و از نظر فرمولی، به شرح زیر است:

$$G_i = \frac{\sum_j w_{ij}\, x_i}{\sum_j x_j} \quad i = ۱, ..., n \qquad (۴)$$

در اینجا نیز بهتراست آماره را در بطن آماره استاندارد شده تفسیر کنیم. برای به دست آوردن آماره استاندارد شده، به دانستن امید ریاضی و واریانس آماره نیاز است. امید

[5] Geary's Ratio C
[6] General G-Statistic

ریاضی واریانس به شرح زیر بدست می‌آید:

$$var(G_i) = \frac{w_i(n-w_i)s^2}{n^2(n-1)\bar{x}^2} \ , E(G_i) = \frac{w_i}{n} \qquad (۵)$$

که در آن $\bar{x} = \sum_i x_i/n$ و $w_i = \sum_j w_{ij}$ و $s^2 = \sum_i (x_i - \bar{x})^2/n$ حال با توجه به (۵) آماره استاندارد شده G عمومی به صورت زیر حاصل می‌شود که در الگوریتم خوشه‌بندی که در ادامه مطرح می‌شود از این آماره استفاده شده است.

$$z(G_i) = \frac{\sum_{j=1}^{n} w_{ij}x_j - \bar{x}w_{ij}}{\sqrt{\frac{s^2}{n-1}\left(n\left(\sum w_{ij}^2\right) - w_i^2\right)}} \qquad (۶)$$

۴ تشخیص خوشه‌بندی فضایی بر مبنای آماره‌های محلی برای خود همبستگی فضایی

خوشه داده‌ها در مباحث آنالیز آماری از پیشینه و سابقه طولانی برخوردار است. تحلیل خوشه ای در جستجوی یافتن روندی برای گروه بندی داده‌ها می‌باشد. اساس تحلیل خوشه ای، یافتن مبنایی از شباهت‌ها وعدم شباهت‌ها می‌باشد که در گروه‌بندی لحاظ می‌شود.

فرض کنید n مشاهده در s مکان داشته باشیم که از هر واحد p متغیر اندازه گیری شده است. الگوریتم خوشه‌بندی به دنبال آن است تا هر واحد را به یکی از k خوشه تخصیص دهد. بیشتر الگوریتم های خوشه‌بندی را می توان به دو دسته تقسیم بندی نمود.

۱) الگوی پارتیشن‌بندی واحدها را به تعداد خوشه‌های از پیش تعیین شده تقسیم می‌کند،
۲) روش‌های سلسله مراتبی خوشه‌هایی به صورت پیشامدهای تصادفی تولید می‌کند.

در این مقاله از روش اول یعنی الگوی پارتیشن بندی با تعداد خوشه های از پیش تعیین شده استفاده شده است. یکی از متداول ترین روش ها روش k-means می‌باشد که مراحل این الگوریتم برای k خوشه به شکل زیر می‌باشد:
گام اول : ابتدا k نمونه انتخاب کرده و در k نقطه‌ی مرکزی قرار می دهیم.
گام دوم: فاصله $n-k$ نقطه باقیمانده را نسبت به مرکزها حساب کرده و این واحدها

به خوشه‌هایی که فاصله یشان کمترین است قرار می‌گیرد.
گام سوم : دوباره مرکزها را حساب و فواصل واحدها را تا مرکزها محاسبه کرده و این عملیات تا جایی ادامه پیدا میکند که به یک پایداری در مرکزها برسد.

حال باتوجه به شناخت معیارهای خودهمبستگیفضایی و روش خوشه‌بندی k-means الگوریتم زیر برای خوشه‌بندی داده‌های فضایی پیشنهاد می‌شود:

۱) ماتریس وزن W_{ij}
را برای داده‌ها محاسبه کرده که یک ماتریس $n \times n$ می‌باشد. برای هر متغیر، آماره ی استاندارد شده ی گتیس و اُرد محاسبه می‌شود فرمول (۶) $z(G_j(x_i))$ را برای متغیر j ام $(j = ۱,...,p)$ در i امین واحد $(i = ۱,...,n)$ محاسبه کرده و در نهایت ماتریس Z با بعد $n \times p$ حاصل می‌شود. هر ستون از Z نشانگر مولفههای خودهمبستگی فضایی محلی برای یک متغیر است در حالی که هر سطر آن نمایانگر تمرکز خوشهای در اطراف هر واحد است.

۲) الگوریتم خوشه‌بندی k-means را روی داده‌های جدید که در ماتریس Z ذخیره شده است، اعمال می‌شود. این مرحله باعث می‌شود خوشه‌بندی پیشنهادی هم براساس اطلاعات فضایی داده‌ها و هم بر اساس اطلاعات خود متغیرها انجام شود. الگوریتم k-means تا جایی ادامه پیدا میکند که تابع زیر کمینه گردد.

$$WSS = \sum_{k=۱}^{K} \sum_{C(i)=k} \|z_i - \bar{z}_k\|^۲ \qquad (۷)$$

که در آن z_i ، i اُمین سطر از ماتریس z و $\bar{z}_k(j = ۱,...,k)$ مرکز خوشه ها، $\|x - y\| = \sqrt{\sum_{j=۱}^{n}(x_j - y_j)^۲}$ فاصله اقلیدسی می‌باشد.

۳) مراحل بالا برای k مقدار محاسبه می‌شود و برای بدست آوردن تعداد خوشه های بهینه از آماره‌ی Gap استفاده می‌شود.

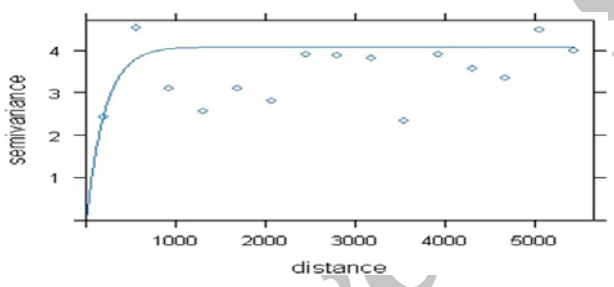
۵ داده‌های استفاده شده

داده‌ها یی که در این مطالعه مورد بررسی قرار گرفته است، از یک بررسی در سال 1389 در شهرستان گرگان (استان گلستان، حوزه قره سو) می‌باشد، که 79 مزرعه گندم به طور تصادفی از شمال، جنوب، غرب و شرق حوزه انتخاب و با استفاده از کادر ۰/۲۵ متر مربعی نمونه‌برداری شده‌اند. در هر کادر بیماری‌های قارچی لکه قهوه‌ای گندم $(Bipolarissorokinianc)$ و زنگ قهوه‌ای $(Pucciniareconditaf.sp.tritici)$

با توجه به علائم موجود بر روی اندام گیاه، مشخص و فراوانی آن مورد شمارش قرار گرفته‌اند. سپس با توجه به تعداد کادر پرتاب شده در هر مزرعه میانگین فراوانی هر بیماری قارچی محاسبه گردیده است.

۶ نتایج، بحث و کارهای آینده

برای خوشه‌بندی در محدوده مورد مطالعه به این صورت عمل شد. با استفاده از نرم افزار R و با اعمال الگوریتم پیشنهادی، آماره (۶) برای داده‌ها محاسبه گردید. در این راستا نیاز به ماتریس وزن بوده که از روش معکوس فاصله، درایه‌های ماتریس محاسبه شد. با توجه به وریوگرام رسم شده (شکل ۲) مشاهده شد که تقریباً از فاصله ی ۵۰۰ متری ساختار فضایی از بین می رود، لذا درایه‌هایی از ماتریس وزن را که کمتر از (۱/۵۰۰) بود، برابر صفر قرار داده شد. باتوجه به ماتریس جدید، مقادیر Z محاسبه و خوشه‌بندی انجام گرفت. نتایج خوشه‌بندی به صورت جدول شماره (۱) می‌باشد که نتایج گام دوم که دارای خطای کمتری بود، انتخاب گردید.

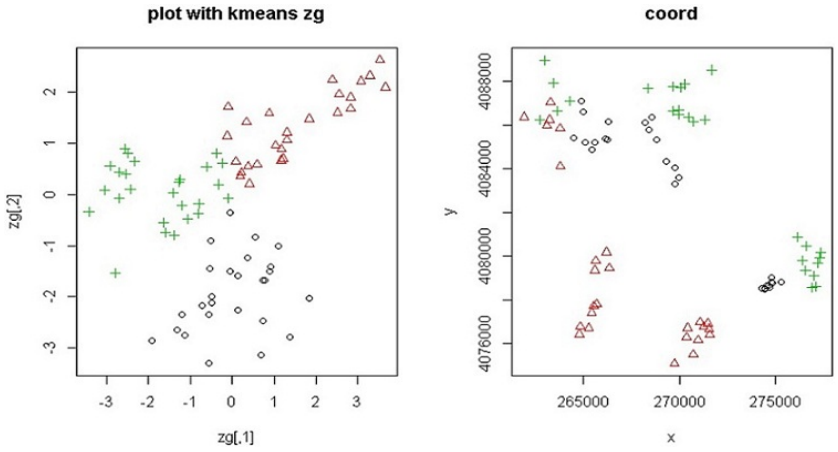


شکل ۲.: وریوگرام بیماری قارچی لکه قهوه‌ای (*Bipolaris sorokinianc*)

جدول ۱: خطای درونی و خطای کل مربوط به سه مرحله اجرای الگوریتم $k-means$

گام	خطای درونی خوشه ها			خطای کل
اول	۳۵/۵۱۲۱۳	۳۳/۳۱۵۷۸	۴۹/۰۸۱۵۴	۱۱۷/۹۰۹۵
دوم	۲۶/۶۹۰۳۲	۵۸/۲۶۹۵۱	۳۲/۹۰۱۰۰	۱۱۷/۸۶۰۸
سوم	۳۳/۸۶۱۷۱	۳۲/۹۰۱۰۰	۵۱/۳۳۱۷۶	۱۱۸/۰۹۴۵

با توجه به شکل (۳)، روش پیشنهادی برای خوشه‌بندی داده‌های فضایی، موقعیت مکانی داده‌ها را به خوبی لحاظ کرده است به طوری که داده‌ها یی که در یک خوشه



شکل ۳: نمودار پراکندگی داده‌ها بر اساس آماره (۶) و خوشه‌بندی فضایی (قاب چپ)، نمودار پراکندگی داده‌ها بر اساس مختصات داده‌ها و خوشه‌بندی فضایی (قاب راست)

قرار گرفته اند تقریباً از لحاظ مکانی نزدیک هم می‌باشد. این موضوع از لحاظ زمین آماری بسیار مهم و مقرون به صرفه می‌باشد. همچنین لازم به ذکر است که در گذشته در این موارد از خوشه‌بندی کلاسیک استفاده شده که خوشه هایی با داده‌های پراکنده از لحاظ مکانی ایجاد کرده است. امید است در آینده با وارد کردن خوشه‌بندی فازی به این الگوریتم دقت این روش را افزایش دهیم.

مراجع

Boots, B. (2002), Local measures of spatial association, Ecoscience, 9(2):168-176.

Cliff A. D. and Ord, J. K. (1981), Spatial processes - models and applications, Pion, London.

Getis, A. and Ord, J. K. (1992), The analysis of spatial association by use of distance statistics, Geographical Analysis, 24:189-206.

Ord J.K., Getis A. (1995) Local spatial autocorrelation statistics: distributional issues and an application. Geographical Analysis, 27: 286-306.

Scrucca, L. (2005), Clustering multivariate spatial data based on local measures of spatial autocorrelation. Geographical Analysis, 38: 34-59.