تعیین طبقات در نمونه‌گیریهای چندمنظوره یا طبقه بندی شده

عباس گرامی[۱] محمدباقر سخاوت[۲] محمدباقر حقیقی انارکی[۳]

چکیده

سودبخشی نمونه‌گیری طبقه‌بندی شده درمقایسه با نمونه‌گیری تصادفی ساده، در صورتی بیشتر است که واحدهای آماری درون هر طبقه دارای همگنی بسیار زیاد و واحدهای مربوط به طبقات متفاوت از همگنی کمتری برخوردار باشند. در ادبیات مربوطه، وقتی که وسیله تعیین مرز طبقات تنها یک متغیر کمکی باشد، مطالبی ارائه شده است که علاقه‌مندان را به [۱] رجوع می‌دهیم. در این مقاله، شرایطی بررسی شده است که به جای استفاده از یک متغیر کمکی، به روشهای مختلف چندین متغیر به کار برده شده‌اند و سپس در یک مطالعه موردی، قابلیت این روشهای پیشنهادی مورد مقایسه قرار گرفته اند و نشان داده شده است که روشهای مذکوردر مقایسه با روشهای استاندارد، از کارآیی بالاتری برخوردارند. حتی در شرایطی که بخواهیم تنها یک متغیر کمکی را وسیله تعیین مرز طبقات قرار دهیم، روشهای پیشنهادی دارای مزیت بیشتری می‌باشند.

واژه‌های کلیدی: نمونه‌گیری چندمنظوره، نمونه‌گیری طبقه‌بندی شده، تحلیل خوشه‌ای، تحلیل مؤلفه‌های اصلی، ضریب تغییرات و معیارهای مقایسه.

۱. مقدمه

سالیانه مقادیر زیادی از نیرو و امکانات کشور صرف تهیه و تولید آمارهای مورد نیاز دستگاههای برنامه‌ریزی، اجرایی و نیز مدیریت جامعه در بخش‌های دولتی و خصوصی می گردد. اکثریت آمارهای مورد نیاز از طریق تهیه و اجرای طرح‌های نمونه‌گیری حاصل می‌شود و دقت آمارهای به دست آمده از این طریق، به عوامل زیادی از جمله طرح مورد استفاده بستگی دارد. مسلماً طرح‌های با حجم نمونه نسبتاً کم و دقت بالا در صرفه‌جویی امکانات تأثیر بسزایی دارد و بنابراین در نمونه‌گیریها همواره کوشش می‌شود از طرح‌هایی استفاده شود که با حجم نمونه معین، دقت بیشتری به همراه داشته باشند.

از جمله طرح‌های بسیار معمول و مورد استفاده در نمونه‌گیریها من‌جمله نمونه‌گیریهای کشاورزی، نمونه‌گیری طبقه‌بندی شده است که در آن یک متغیر کمکی یا ترکیب خطی ساده‌ای از چند متغیر کمکی(مثلاً مجموع) به عنوان معیار طبقه‌بندی و تعیین طبقات مورد استفاده قرار می گیرد. کارآیی و قابلیت این کار، زمانی است که متغیرهای کمکی مورد استفاده در طبقه بندی با متغیرهای مورد بررسی

[۱] عضو هیئت علمی پژوهشکده آمار

[۲و۳] آمارشناسان مرکز آمار ایران

در نمونه‌گیری، دارای همبستگی بالا باشند. اما ممکن است در عمل چنین وضعیتی وجود نداشته باشد. بنابراین لازم است روشهای دیگری مورد استفاده قرار گیرد که در صورت برقرار نبودن وضعیت مورد بحث، از کارآیی بیشتری برخوردار باشند.

در نمونه‌گیریهای طبقه‌بندی شده، برای تعیین طبقات، تعداد نمونه، انتساب نمونه‌ها به طبقات و انتخاب واحدهای نمونه، موارد زیر باید بررسی شوند

الف - پیداکردن یک یا چند متغیر کمکی مناسب برای طبقه‌بندی واحدهای جامعه

ب - روش ساختن طبقات

ج - تعیین تعداد طبقات

د - تعیین تعداد نمونه

هـ - روش انتساب نمونه به طبقات

در مقاله حاضر بررسی بندهای الف وب از بندهای فوق یعنی تعیین طبقات مورد توجه است. در این راستا ابتدا نحوه تعیین متغیر یا متغیرهای کمکی برای طبقه‌بندی واحدهای جامعه، مورد بحث قرار می‌گیرند و سپس روشهایی برای ساختن طبقات پیشنهاد می‌شود . در انتها با معرفی چند معیار، روشهای کلاسیک موجود برای ساختن طبقات با روشهای پیشنهادی مقایسه می‌شوند.

۲. متغیرهای کمکی مناسب برای ساختن طبقات

نمونه‌گیریها اغلب چند منظوره هستند، بدین مفهوم که برآورد پارامترهای مربوط به چندین متغیر یا چندین پارامتر از یک متغیر از جامعه هدف ویا تلفیقی از این دو نوع را مورد نظر دارند. در صورتی که از طرح نمونه‌گیری طبقه بندی شده استفاده شود، موضوع مهم، تعیین متغیر یا متغیرهای کمکی است که بر اساس آن طبقات تعریف می‌شوند. بدیهی است که باید منطقاً بین متغیرهای کمکی با متغیرهای مورد نظر در طرح، نوعی وابستگی وجود داشته باشد، تا طبقات تشکیل شده موجب افزایش کارآیی برآوردها شود. در بعضی نمونه‌گیریها مثلاً نمونه‌گیریهای کشاورزی، استفاده از مجموع سطح زیر کشت محصولات مورد آمارگیری در آخرین سرشماری کشاورزی به عنوان یک متغیر کمکی، منطقی به نظر می‌رسد ولی این سوال مطرح می‌شود که اگر قرارباشد تنها از یک متغیر کمکی برای تعیین طبقات استفاده شود آیا مجموع ساده سطح زیر کشت در آخرین سرشماری می‌تواند بهترین متغیر کمکی باشد؟ بدیهی است که استفاده از تنها یک متغیر کمکی برای ساختن طبقات درنمونه‌گیریهای چند منظوره نمی‌تواند کارآیی بالایی به همراه داشته باشد. بنابراین تعداد مطلوب متغیر کمکی و نحوه تعیین آنها سؤال دیگری است که مطرح می‌شود. استفاده از مقادیر همه متغیرهای مورد بررسی براساس اطلاعات آخرین سرشماری به عنوان متغیرهای کمکی یک انتخاب معقول به نظر می‌رسد. با این وجود ممکن است حجم زیاد محاسبات در تعیین طبقات، کاربرد آن را مشکل و یا حتی غیر ممکن سازد. در این صورت باید تعداد کمتری متغیر کمکی را مورد توجه قرار داد. در این بخش پیشنهاداتی در مورد تعیین متغیر یا متغیرهای کمکی برای ساختن طبقات به شرح زیر ارائه می‌شود.

- انتخاب متغیری از بین متغیرهای کمکی ممکن که دارای بیشترین تغییرات است.
- استفاده از تحلیل مؤلفه های اصلی[۴]، PCA، برای ساختن متغیرهای کمکی.

۲ . ۱ انتخاب متغیر کمکی با بیشترین تغییرات

چنانچه یک جامعه آماری را که برای هر واحد آن، اطلاعات p متغیر کمکی در اختیار داریم بخواهیم با استفاده از یک متغیر طبقه‌بندی کنیم، انتظار داریم که متغیر انتخابی، آن متغیری باشد که بیشترین واریانس یا ضریب تغییرات را داشته باشد، گرچه استفاده از این متغیر برای طبقه‌بندی لزوماً منجر به یک طرح با خواص بهینه نخواهد بود. در این ارتباط می‌توان به این نکته نیز اشاره کرد که اگر بخواهیم یک جامعه آماری را که برای هر واحد آن p متغیر کمکی در دست است به کمک تعداد کمتری متغیر، مثلاً m $(m<p)$ متغیر، طبقه‌بندی کنیم، در این صورت نیز مناسب به نظر می‌رسد تا از m متغیری که بیشترین واریانس یا بیشترین ضریب تغییرات را دارند استفاده کنیم.

---

[۴] Principal Component Analysis (PCA)

۲-۲ تحلیل مؤلفه‌های اصلی و استفاده از آن در تعیین متغیرهای کمکی مناسب

هدف اصلي تحليل مؤلفه‌هاى اصلى، كاهش حجم داده‌هاى چند متغيره‌اى است كه داراى وابستگى‌هاى خطى درونى قابل توجهى هستند، به طورى كه داده‌هاى كاهش يافته بيشترين اطلاعات داده‌هاى اصلى را منتقل كنند. اين روش مبتنى بر ويژه بردارهاى[۵] ماتريس واريانس - كوواريانس متغيرها يا ماتريس ضريب همبستگى مربوط به آنهاست. چنانچه در اين تحليل، مشاهدات به شكل اصلى خود مورد استفاده قرار گيرد، از ماتريس واريانس - كوواريانس و در حالتى كه اين مشاهدات به صورت استاندارد درآيند از ماتريس ضريب همبستگى استفاده مى‌شود.

شايان ذكر است كه تحليل مؤلفه‌هاى اصلى، معمولاً گام آخر تجزيه و تحليل داده‌ها را تشكيل نمى‌دهد، بلكه اين تحليل ابزارى جهت انجام تجزيه و تحليل‌هاى بعدى نيز ارائه مى‌كند كه بعداً به رابطه آن با تحليل خوشه‌اى[۶] اشاره مى‌كنيم. در صورتى كه حجم داده‌هاى موجود از نظر تعداد متغير بسيار زياد باشد و بخواهيم از تحليل خوشه‌اى استفاده كنيم، ممكن است ابزارهاى محاسباتى در دسترس، قابليت انجام محاسبات روى حجم بالیي داده‌ها را نداشته باشند. در چنين حالتي به جاى استفاده از كليه متغيرها ، از تعداد اندكى از مؤلفه‌هاى اصلى مهم استفاده مى‌شود كه امكانات محاسباتى آن در اختيار باشد.

در پاره‌اى از نمونه‌گيريها ، مخصوصاً نمونه‌گيريهاى كشاورزى، از يك تركيب خطى، مثلاً مجموع سطح كشت تمامى محصولات هر آبادى، به عنوان معيار طبقه‌بندى استفاده مى‌شود. اگر قرار باشد تنها از يك تركيب خطى استفاده كنيم، شايد اين تركيب كه همان مجموع سطح زيركشت هر آبادى است، معيار بدى نباشد. اما معمولاً استفاده از يك تركيب خطى به تنهايى، افت اطلاعاتى زيادى را به همراه دارد. اگر قرار باشد از دو تركيب خطى يا بيشتر استفاده كنيم كه افت اطلاعاتى كمتر شود، پيداكردن اين تركيبات خطى بدون استفاده از مؤلفه‌هاى اصلى، مشكل و يا غيرممكن مى‌نمايد. در چنين مواقعى مى‌توان تركيبات خطى متناظر با ويژه بردارهاى ماتريس واريانس - كوواريانس يا مؤلفه‌هاى اصلى مهم را كه از خواص آمارى مناسبى برخوردار هستند مورد استفاده قرار داد. به علاوه حتى زمانى كه بخواهيم بر اساس تنها يك تركيب خطى، طبقه‌بندى را انجام دهيم، باز هم استفاده از مهمترين مؤلفه اصلى(اولين مؤلفه اصلى[۷]) در مقايسه با تركيب خطى ساده مجموع صفات، اطلاعات بيشترى را انتقال خواهد داد. براى توضيح بيشتر پيرامون اين تحليل ، به منبع [۵] مراجعه شود.

۳. معرفی روشهای ساختن طبقات

روشهاى ساختن طبقات كه در منابع نمونه‌گيرى مورد بحث قرار مى‌گيرند، عمدتا برمبناى يك متغير كمكى است، مثل سطح زير كشت محصولات كشاورزى در آخرين سرشمارى كشاورزى كه قبلا توضيح داده شد. استفاده از تنها يك متغير كمكى براى ساختن طبقات از مزاياى سادگى در محاسبات و تعيين(حدود) طبقات برخورداراست. اما مسلما نمى‌توان گفت كه اين متغير به تنهائى بهترين طبقه‌بندى را بدست مى‌دهد. در بخش قبل متغير يا متغيرهاى كمكى كه مى‌توانند براى طبقه‌بندى استفاده شوند، مورد بحث قرار گرفت. در صورتى كه تنها يك متغير كمكى براى طبقه بندى موجود باشد يا تنها يك متغير مناسب باشد، روش كلاسيك دالينيوس[۸][۱] مورد استفاده قرار مى‌گيرد. گرچه سادگى اين روش و مناسب بودن آن در شرايط خاص، از مزاياى آن محسوب مى‌شود، اما سؤال مهم اين است كه آيا اين روش، بهترين طبقه‌بندى ممكن را بدست مى‌دهد؟ گرچه پاسخ روشنى براى اين سؤال وجود ندارد، اما امكان استفاده از اين روش براى شرايط با بيش از يك متغير كمكى، وجود ندارد. در زير به معرفى بعضى تكنيكهاى تحليل خوشه‌اى كه امكان طبقه‌بندى جامعه بر اساس بيش از يك متغير كمكى را نيزفراهم مى‌كند پرداخته مى‌شود.

٤. روشهای ساختن طبقات با استفاده ازتحلیل خوشه‌ای

هدف از تحليل خوشه‌اى عبارت از تقسيم n فرد بين k خوشه است، به گونه‌اى كه بر اساس معيارى كه تعريف مى‌شود، افراد داخل

[۵] Eigenvectors

[۶] Cluster Analysis

[۷] First Principal Component

[۸] Dalenius

یک خوشه از لحاظ p متغیر مورد بررسی $x_1, x_2, \cdots, x_p$، که از این به بعد با بردار X نشان داده می‌شود، دارای تفاوتهای اندکی با یکدیگر باشند، درحالی که افراد از خوشه‌های مختلف، تفاوتهای زیادی با هم داشته باشند. روشهای مختلفی برای این نوع تحلیل وجود دارد که قالب کلی آنها مختصراً در زیر توضیح داده می‌شود.

معمولترین روشهای تحلیل خوشه‌ای را می‌توان در دو قالب کلی، روشهای سلسله مراتبی[۹] و روشهای تقسیم‌بندی[۱۰]، گروه‌بندی کرد.

٤-۱ روشهای سلسله مراتبی

دراین روشها ، اساس کار بر ترکیب یا تقسیم متوالی افراد داخل خوشه‌های مختلف یا خوشه‌های یکسان است. این روشها نیز همان گونه که از تعریف بر می‌آید خود به دو گروه اصلی تقسیم‌بندی می‌شوند.

الف- روشهای ترکیبی یا تراکمی[۱۱]

ب- روشهای تقسیمی[۱۲]

در فرایند روشهای ترکیبی، ابتدا هر فرد به تنهائی یک خوشه را تشکیل می‌دهد، یعنی در بدو امر به تعداد افراد خوشه وجود دارد. سپس در مراحل بعدی، در هر مرحله، این خوشه‌ها درهم ترکیب شده و خوشه‌های بزرگتری را تشکیل می‌دهند، به گونه‌ای که رفتن از یک مرحله به مرحله بعدی، موجب کم شدن یک خوشه نسبت به تعداد خوشه‌های مرحله قبلی خواهد شد، و در آخرین مرحله کلیه افراد در داخل یک خوشه قرار می گیرند.

در روشهای تقسیمی ، برعکس روشهای تراکمی، ابتدا همه افراد دریک خوشه منظور می‌شوند و سپس در هر مرحله خوشه جدیدی اضافه می‌شود به طوری که تعداد خوشه‌های هر مرحله یکی بیشتر از تعداد خوشه‌های مرحله بلافاصله قبل از آن است. بدین ترتیب در آخرین مرحله، هر فرد در داخل یک خوشه قرار می گیرد و تعداد خوشه‌ها در این مرحله برابر با تعداد افراد است. در خصوص این روشها به دلیل اینکه کمتر در عمل مورد استفاده قرار می گیرند در اینجا توضیح بیشتری ارائه نمی‌شود. توضیحات بیشتر را می‌توان در منابع [۲] ، [۳] و [٦] پیدا کرد.

٤-۲ روشهای ترکیبی

این روشها از لحاظ کاربرد بسیار معمول و شامل انواع متفاوتی هستند. متداولترین انواع این روشها عبارتند از

۱- تک اتصالی[۱۳] (SL)

۲- اتصال کامل[۱٤] (CL)

۳- اتصال متوسط[۱۵] (AL)

٤- روش مرکز هندسی[۱٦]

۵- روش وارد[۱۷]

روشهای فوق همگی یک فرایند مشترک دارند ولی در آنها از معیارهای مختلفی برای ادغام خوشه‌ها استفاده می‌شود. در این فرایند دنباله‌ای از ادغامها، روی داده‌ها صورت می‌پذیرد. اولین ادغام، روی n خوشه انجام می گیرد که هر کدام شامل یک عنصر است و آخرین ادغام که صورت گیرد، یک خوشه بدست می‌آید که شامل همه n عنصر می‌شود. در این فرایند اصول عملیات پایه برای تمامی این روشها مشابه وبدین ترتیب است که در شروع ، هر خوشه از یک عنصر تشکیل می‌شود. سپس

- مرحله ۱- نزدیکترین زوج از خوشه‌های مجزا شناسائی و در هم ادغام می‌شوند واز تعداد خوشه‌ها، یکی کاسته می‌شود.
- مرحله۲- در صورتی که تعداد خوشه‌ها در این مرحله یک باشد، فرایند متوقف می‌شود. در غیر این صورت مرحله ۱ دوباره تکرار می‌شود.

در هر یک از مراحل، آن دو خوشه‌ای در هم ادغام می‌شوند و یک خوشه را تشکیل می‌دهند که نزدیکترین(شبیه ترین) خوشه‌ها باشند. تفاوت بین روشهای مختلف، تفاوت در تعریف فاصله (یا شباهت) بین یک خوشه شامل یک عنصر، با خوشه دیگر شامل گروهی از عناصر، یا

---

۹ Hierarchical Techniques

۱۰ Partitioning Techniques

۱۱ Agglomerative Techniques

۱۲ Divisive Techniques

۱۳ Single Linkage

۱٤ Complete Linkage

۱۵ Average Linkage

۱٦ Centriod Method

۱۷ Ward Method

بین یک خوشه با گروهی از عناصر با خوشه‌ای دیگر است که آن نیز شامل گروهی از عناصرمی‌باشد. بدین ترتیب که

- در روش تک اتصالی، فاصله بین دو خوشه، فاصله نزدیکترین افراد، یکی از خوشه اول و دیگری از خوشه دوم است. به همین دلیل، این روش به روش نزدیکترین همسایه[۱۸] نیز معروف است.

- در روش اتصال کامل، بر عکس روش تک اتصالی، فاصله بین دو خوشه، فاصله بین دورترین افراد، یکی از خوشه اول و دیگری از خوشه دوم است. به همین دلیل، این روش به روش دورترین همسایه[۱۹] نیز معروف است.

- در روش اتصال متوسط، فاصله بین دو خوشه، میانگین فاصله کلیه زوج عناصر، یکی ازخوشه اول و دیگری از خوشه دوم است.

- در روش مرکز هندسی، هر خوشه‌ای که تشکیل می‌شود، بردار میانگین کل عناصر آن محاسبه می‌شود و برای تعیین فاصله آن خوشه با خوشه دیگر مورد استفاده قرار می‌گیرد.

- در این روش که توسط وارد[۲۰] [7] معرفی شد و به روش مینیمم واریانس معروف است، ادغامها را به طریقی مورد بررسی قرار می‌دهد که افزایش واریانس حاصل از ادغام به حداقل برسد. در این روش، در هر مرحله ابتدا ادغامهای هردو زوج از خوشه‌ها مورد توجه قرار می‌گیرد و در نهایت آن دو زوجی با هم ادغام می‌شوند که منجر به حد اقل افزایش در واریانس شود.

لازم به توضیح است که در روشهای ۱ تا ۳، ابتدا یک ماتریس فاصله[۲۱] (یا شباهت) تشکیل می‌شود به گونه‌ای که درایه سطر i ام و ستون j ام آن، یک کمیت فاصله‌ای بین فرد i ام و j ام براساس مشاهدات p صفت، این دو فرد را نشان می‌دهد. سپس در کلیه مراحل فرایند، محاسبات لازم بر اساس این ماتریس صورت می‌گیرد. در حالی که در روشهای ٤ و ٥، در کلیه مراحل، اطلاعات همه صفات برای کلیه عناصر مورد استفاده واقع می‌شود.

٤-۳ روشهای تقسیم‌بندی

گرچه در تحلیل خوشه‌ای با استفاده از هر روش، تعیین تعداد مطلوب خوشه، k، نیز خود یک هدف است، اما با توجه به آنچه در بالا راجع به روشهای سلسله مراتبی توضیح داده شد، در فرآیند مربوط به روشهای مذکور، دانستن عدد k، یک ضرورت نیست. حال آنکه برعکس، در روشهای تقسیم‌بندی، داشتن عدد k یک ضرورت است. این موضوع که مقدار k ی مطلوب چقدر باشد بحث مبسوطی را می‌طلبد که دراین مقاله نمی‌گنجد وعلاقه‌مندان می‌توانند از منابع و مراجع ذکر شده در فهرست منابع استفاده کنند. علاوه براین اختلاف، تفاوت فاحش دیگر بین روشهای سلسله مراتبی و تقسیم‌بندی، آن است که در روشهای تقسیم‌بندی به مفهوم دقیق کلمه، به ازاء یک k مشخص، باید ابتدا کلیه حالات ممکن تقسیم‌بندی n فرد بین k خوشه به نحوی که هرخوشه شامل حداقل یک فرد باشد درنظر گرفته شود و سپس از بین کلیه حالات ممکن، براساس معیارهایی، حالت مطلوب مشخص گردد. از همین توضیح مشخص است که تعداد حالات ممکن تقسیم‌بندی فوق، بسیار زیاد خواهد بود و این خود یک محدودیت جدی در استفاده کامل از این روش را به همراه دارد. یک روش تقریبی از این قبیل که به روش k میانگین[۲۲] یا خوشه بندی سریع[۲۳] معروف است، درنرم افزارهای آماری وجود دارد. درتحلیلهای عددی که در این مقاله ارائه گردیده، از این روش استفاده شده است.

۵. معیارهای مقایسه روشهای مختلف ساختن طبقات

به منظور سنجش و مقایسه روشهای مختلف تعیین طبقات و تعیین روش یا روشهای برتر، تعریف معیارهای منطقی ضرورت دارد. هنگامی که از سودبخشی روشی نسبت به روش دیگر صحبت به میان می‌آید، معمولاً شاخص یا شاخصهایی برای بیان میزان سودبخشی مطرح

---

[۱۸] Nearest Neighbour

[۱۹] Furthest neighbour

[۲۰] Ward

[۲۱] Distance Matrix

[۲۲] K-means

[۲۳] Quick Cluster

می‌شود. اهم شاخصهایی که می‌توانند به نوعی سودبخشی روشهای مطرح شده در قسمتهای قبل برای ساختن طبقات را نسبت به یکدیگر ارزیابی کنند، در زیر توضیح داده می‌شود.

۵-۱ مجموع تغییرات متغیرها در داخل طبقات یا $tr(S_W)$

همان طور که قبلاً ذکر شد، در طبقه‌بندی واحدهای جامعه، هدف، کاهش هر چه بیشتر تغییرات داخل طبقات و افزایش آن در بین طبقات می‌باشد. تغییرات کل در یک طبقه بندی را می‌توان به صورت $S_T = S_B + S_W$ تقسیم‌بندی کرد که در آن S_T تغییرات کل، S_B تغییرات بین طبقات و S_W تغییرات داخل طبقات را نشان می‌دهد. بنابراین در مقایسه دو روش، روشی سودبخش‌تر است که S_W مربوط به آن روش در مقایسه با S_W روش دیگر، کوچکتر باشد. در حالت کلی، یعنی وقتی بیش از یک متغیر تحت مطالعه می‌باشد، مجموع تغییرات کلیه متغیرها در داخل طبقات به عنوان ملاک مقایسه دو روش مطرح می‌باشد. بنابراین چنانچه ماتریس مجموع مربعات و مجموع حاصلضربهای متغیرها در داخل طبقات را با S_W نشان دهیم، مجموع عناصر روی قطر این ماتریس یعنی $tr(S_W)$، به عنوان ملاک مقایسه دو روش، مورد استفاده قرار می‌گیرد.

۵-۲ شاخص *Jarque*

در این قسمت به معرفی معیاری از منبع [۴] می‌پردازیم که آن را بنا بر اسم نویسنده، شاخص *Jarque* نامیده‌اند. فرض کنیم که طبقه‌بندی جامعه بر اساس هر کدام از متغیرها به طور جداگانه انجام شود. در این صورت، این طبقه‌بندی‌ها که با علامت S_1^* و S_2^* و... و S_p^* نشان داده می‌شود، هر کدام برای متغیر مربوطه واریانسی به دست می‌دهد که کمترین مقدار را داراست.

اگر *MSE* حاصل از طبقه‌بندی جامعه بر پایه متغیر jام را با $V_s^*(j)$ و *MSE* حاصل از به کار بردن روش مورد نظر را با $V_s(j)$ نشان دهیم، شاخص *Jarque* عبارت است از

$$F(s)=\sum_{j=1}^{p}\frac{V_s(j)}{V_s^*(j)}$$

برای مقایسه سودبخشی دو روش باید $F(s)$ مربوط را محاسبه کرد. بدیهی است روشی از بین دو یا چند روش، سودبخش‌تر است که $F(s)$ آن روش نسبت به روشهای دیگر کوچکتر باشد.

۵-۳ دترمینان ماتریس S_W ، $\det(S_W)$

دترمینان ماتریس S_W ، $\det(S_W)$ نیز یکی از ملاکهایی است که می‌توانند برای مقایسه سودبخشی مورد استفاده قرار گیرد. در نرم‌افزارهای آماری در تحلیل چند متغیره واریانس یک طرفه[۲۴]، معیار دیگری به نام ضریب *Wilks-Lambda* محاسبه می‌شود که از رابطه $\Lambda = \det(S_W)/\det(S_T)$ به دست می‌آید. اما ازآنجایی که برای یک مجموعه مشخص از داده‌ها $\det(S_T)$ مقدار ثابتی است، لذا مقدار این ضریب، Λ ، بستگی مستقیم به مقدار $\det(S_W)$ دارد، بنابراین روشی سودبخش‌تر است که ضریب Λ مربوط به آن روش، در مقایسه با روشهای دیگر کوچکتر باشد.

۶. کاربرد تحلیل خوشه‌ای

جهت مقایسه کارآیی طبقه‌بندی واحدهای جامعه بر اساس چندصفت، یک مطالعه موردی در سطح شهرستان لردگان از استان چهارمحال و بختیاری و در سطح استان همدان انجام شد. دراین مطالعه، تعداد طبقات ۶ و ۷ درنظر گرفته شده که در نمونه‌گیریهای مرکز آمار ایران، معمول بوده است.

شاخصهای مربوط به طبقه‌بندی بر اساس متغیر کمکی مجموع سطح زیرکشت محصولات سالانه عمده با استفاده از دو روش دالینیوس و K- میانگین محاسبه و در جداول ۱ و ۲ آورده شده است.

مقایسه شاخصهای مختلف در دو روش به خوبی نشانگر سودبخشی بیشتر روش طبقه‌بندی با استفاده از روش K- میانگین نسبت به روش دالینیوس می‌باشد. به عنوان مثال چنانچه با استفاده از روش K- میانگین آبادی‌های استان همدان را به ۶ طبقه تقسیم کنیم، حدود ۲۵ درصد کاهش واریانس نسبت به استفاده از روش دالینیوس خواهیم داشت و این کاهش واریانس در مورد شهرستان لردگان حدود ۵۶ درصد

[۲۴] One-way Multivariate Analysis of Variance

می‌باشد. از آنجا که تعداد نمونه رابطه مستقیمی با واریانس دارد، لذا انتظار می‌رود که با استفاده از طبقه‌بندی به روش K- میانگین بجای روش دالینیوس، تعداد نمونه در شهرستان لردگان حدود ۵٦ درصد و در استان همدان حدود ۲۵ درصد کاهش یابد.

۷. کاربرد تحلیل مؤلفه‌های اصلی

بر اساس تئوری، اولین مؤلفه اصلی در مقایسه با هر ترکیب خطی دیگر، از جمله ترکیب خطی ساده مجموع متغیرها، اطلاعات بیشتری را ارائه می‌دهد. در نتیجه استفاده از اولین مؤلفه اصلی به عنوان تنها متغیر کمکی نیز موجب افزایش کارآیی می‌شود. برای نشان دادن کارآیی اولین مؤلفه اصلی در طبقه‌بندی نسبت به ترکیب خطی ساده مجموع، بر اساس اطلاعات سرشماری کشاورزی سال ٦۷ شهرستان لردگان و استان همدان، شاخصهای سه‌گانه محاسبه و در جداول ۳ و ٤ آورده شده است. همان طورکه جداول ۳ و ٤ نشان می‌دهند، کارآیی مؤلفه اصلی اول درطبقه‌بندی، در مقایسه با ترکیب جمع، ازلحاظ معیارهای $tr(S_W)$ و Λ بیشتر است ولی از لحاظ معیار $F(s)$ این کارآیی قدری کمتر است.

۸. کاربرد صفت یا صفات با بیشترین تغییرات

در مثال حاضر، برای بررسی استفاده از صفت یا صفات با بیشترین واریانس یا ضریب تغییرات به عنوان ملاک مناسبی برای گروه‌بندی واحدهای جامعه، آبادی‌های استان همدان در یک مرحله بر پایه ۵ صفت که ضریب تغییرات متفاوتی دارند و در مرحله دیگر بر پایه ۵ صفت که واریانس متفاوتی دارند گروه‌بندی گردید که نتیجه محاسبات انجام شده در مورد شاخص یا شاخصهای مناسب برای سنجش کارآیی در جداول ۵ و ٦ آورده شده است.

ملاحظه روند تغییرات شاخصهای Λ و $F(s)$ در مقایسه با ضریب تغییرات، بیانگر این موضوع است که چنانچه برای کلیه صفات مورد بررسی اهمیت یکسانی قائل شویم، انتخاب صفتی که بیشترین ضریب تغییرات را داشته باشد، بیشترین کارآیی را برای طبقه‌بندی واحدهای جامعه به گروه‌های همگن به همراه دارد.

بررسی روند تغییرات $tr(S_W)$ در مقایسه با واریانس صفات مختلف نیز مؤید این موضوع است که اگر در طرح نمونه‌ای مورد نظر، کم کردن هر چه بیشتر واریانس صفات اهمیت داشته باشد، انتخاب صفتی که بیشترین واریانس را دارد می‌تواند منجر به کاهش $tr(S_W)$ گردد و به عبارت دیگر کارآیی زیادی برای گروه‌بندی مناسب واحدهای جامعه داشته باشد.

۹. نتیجه‌گیری و پیشنهادات

با توجه به نتایج به دست آمده، موارد زیر پیشنهاد می‌شود.

- چنانچه در روش تجزیه و تحلیل خوشه‌ای امکانات محاسباتی لازم برای استفاده از کلیه صفات وجود داشته باشد، بهتر است که همواره از اطلاعات مربوط به تمامی صفات در گروه‌بندی واحدها استفاده شود.
- به کارگیری روش تحلیل مؤلفه‌های اصلی، زمانی توصیه می‌شود که امکان استفاده از خوشه‌بندی بر پایه کلیه صفات فراهم نباشد.
- در صورتی که به منظور استفاده از روش خوشه‌بندی ناچار باشیم که با توجه به امکانات محاسباتی موجود به جای p متغیر از $m < p$ متغیر استفاده کنیم، بهتر است برای این منظور m مؤلفه اصلی اول را به کار بریم. هر چند در این بررسی عملاً در این باره تحقیقی صورت نگرفته است ولی با توجه به نظریه مؤلفه‌های اصلی می‌توان صحت این امر را پیش‌بینی کرد.
- چنانچه بنا باشد از یک ترکیب خطی از کلیه صفات برای گروه‌بندی واحدها استفاده کنیم، مؤلفه اصلی اول بر هر ترکیب خطی دیگر (مثلاً مجموع) ارجحیت دارد.
- در صورتی که برای گروه‌بندی، از یک صفت یا ترکیب خطی ازچند صفت استفاده به عمل می‌آید، توصیه می‌شود که به جای روش دالینیوس، یکی از تکنیکهای خوشه‌بندی مثلاً تکنیک K- میانگین به کار برده شود.
- اگر بخواهیم فقط از یک صفت برای گروه‌بندی استفاده کنیم بهتر است صفتی را به کار بریم که با توجه به هدف بررسی، بیشترین واریانس یا ضریب تغییرات را داشته باشد.

۱۰. تشکر و سپاسگذاری

در خاتمه اضافه می‌نماید که تحقیق حاضر با پشتیبانی مالی و کارشناسی مرکز آمار ایران صورت گرفته است که نگارندگان مقاله مراتب تشکر و قدردانی خود را از آن مرکز اعلام می‌دارند. ضمناً جناب آقای نادر فلاح، عضو گروه علمی دانشگاه شاهد در تدوین و تنظیم مقاله حاضر کمک شایانی نمودند که بدین وسیله از ایشان نیز تشکر می‌گردد.از جناب آقای ابراهیم خدائی عضو گروه علمی سازمان سنجش آموزش کشور که در انجام محاسبات کامپیوتری کمک‌های شایانی نمودند نیز قدردانی می‌شود.

جدول (۱) شاخصهای سه‌گانه در شهرستان لردگان

روش مورد استفاده	۷ طبقه $F(s)$	۷ طبقه Λ	۷ طبقه $tr(S_W)$	۶ طبقه $F(s)$	۶ طبقه Λ	۶ طبقه $tr(S_W)$
دالینیوس	۵۲۰۰	۰/۱۱۷۴۳	٤/٦٤٥E۱۱	۲۵۸۱	۰/۱٤۰۵۹	٤/۷۸۱۵E۱۱
K- میانگین	۳۹۳۷	۰/۰۰۱۰۹	۱/٦۹۲۵E۱۱	۲۱٦۹	۰/۰۰۲۲٤	۲/۱۳۵۳E۱۱

جدول (۲) شاخصهای سه‌گانه در استان همدان

روش مورد استفاده	۷ طبقه $F(s)$	۷ طبقه Λ	۷ طبقه $tr(S_W)$	۶ طبقه $F(s)$	۶ طبقه Λ	۶ طبقه $tr(S_W)$
دالینیوس	۳۳۰٤۷	۰/۱۲۲۸٦	٤/۳۸۵٤E۱۲	۱۳۷۲٤	۰/۱٤۱۷۱	٤/٤۷٦۷E۱۲
K- میانگین	۳۳۲۳۱	۰/۰۳۵۷۵	۲/۷۱٤۳E۱۲	۱۳۷۹۳	۰/۰۷۹۵۲	۳/۳٦۷۷E۱۲

جدول (۳) مقایسه شاخصهای سه‌گانه سنجش در شهرستان لردگان

روش	سهم از واریانس کل	۷ طبقه $F(s)$	۷ طبقه Λ	۷ طبقه $tr(S_W)$	۶ طبقه $F(s)$	۶ طبقه Λ	۶ طبقه $tr(S_W)$
مولفه اول	۵۳/۸۲	٤۰۳۳	۰/۰۰۲۲۸	۲/۲ E ۱۱	۲۳۳۷	۰/۰۰۷۵٦	۲/۲ E ۱۱
ترکیب جمع	۲۱/۱٦	۳۷۲۹	۰/۰۰۱٦۱	۲/۷ E ۱۱	۲۲۲۹	۰/۰۰۳۲۱	۲/۸ E ۱۱

جدول (٤) مقایسه شاخصهای سه‌گانه سنجش در استان همدان

روش	سهم از واریانس کل	۷ طبقه $F(s)$	۷ طبقه Λ	۷ طبقه $tr(S_W)$	۶ طبقه $F(s)$	۶ طبقه Λ	۶ طبقه $tr(S_W)$
مولفه اول	۷۷/۸۳	۳۳۱٤۲	۰/۰۳۵٤۲	۲/۳ E ۱۲	۱۳٦۱۹	۰/۰۳۷۳۹	۲/۳ E ۱۲
ترکیب جمع	۸/۳٦	۳۳۱۲۱	۰/۰۳۵۵٤	۳/۸ E ۱۲	۱۳٦۰۰	۰/۰٦۳۰٦	٤/۰ E ۱۲

جدول (۵) شاخصهای سنجش کارآیی برای طبقه‌بندی بر پایه ضریب تغییرات

نام صفت	ضریب تغییرات	Λ	$F(s)$
جو دیم	۱۶۷	۰/۰٤۱۹۸	۱۳۸۱۱
گندم دیم	۱۷۱	۰/۰۳۸۲۹	۱۳۷۷۳
هندوانه دیم	۸۱۹	۰/۰۰۸٦۹	۱۳۷۰۲
پیاز	۹۳۵	۰/۰۰۷۲۵	۱۳۷٦۵
برنج	۲۷۲۳	۰/۰۰۰۰۸	۲٤۸۵

جدول (٦) شاخصهای سنجش کارآیی برای طبقه‌بندی بر پایه ضریب تغییرات

نام صفت	واریانس	$tr(S_W)$
پنبه آبی	۲/۸۲۹٦ E ٦	۸/۸۱٤۷ E ۱۲
نخود آبی	۸/۲۸٤۵ E ۸	۸/۳۱۱۹ E ۱۲
عدس آبی	۱/۲۰٦۲ E ۹	۸/۷۷۵۷ E ۱۲
گندم آبی	۸/۰٤٤۳ E ۱۱	۷/۱۳۵۰ E ۱۲
گندم دیم	٦/۷۵٦۳ E ۱۱	۲/۲۷٤۱ E ۱۲

مراجع

[1] Cochran, W.G., 1977. *Sampling Techniques*, 3rd ed, John Wiley & Sons, New Delhi.

[2] Everitt, B.S., 1993. *Cluster Analysis*, Third Edition, John Wiley & Sons.

[3] Dillon, W.R. and Goldstein M., 1984. *Multivariate Analysis. Methods and Applications*, John Wiley & Sons. Torento.

[4] Jarque, C.M., 1981. *A Solution to the Problem of Optimum Stratification in Multivariate Sampling*, Appl. Statist., Vol. 30, N0. 2, PP 163-169.

[5] Jolliffe, I.I., 1986. *Principal Component Analysis*, Springer-Verlag, New York.

[6] Mardia, K.V., Kent, J.T. and Bibby, J.M., 1982. *Multivariate Analysis*, Academic Press, London.

[7] Ward, J.H., 1963. *Hierarchical grouping to optimize an objective function*, J.Amer.Statist., Vol. 58, PP 236-234.