

مقایسه‌ی روش طبقه‌بندی دالینیوس-هاجز با خوشه‌بندی فازی در داده‌های چوله

آرمان بیداربخت‌نیا،* واحد معروفی

مرکز آمار ایران

چکیده. طرح نمونه‌گیری طبقه‌بندی‌شده در بسیاری از آمارگیری‌های نمونه‌ای مورد استفاده قرار می‌گیرد. روشی که از دیرباز برای طبقه‌بندی عناصر چارچوب به کار برده می‌شود، روش طبقه‌بندی دالینیوس-هاجز است که به‌شکلی قوی به توزیع جامعه وابسته است. در این مقاله به بررسی فرضیات روش طبقه‌بندی دالینیوس-هاجز پرداخته، روش خوشه‌بندی فازی را به‌عنوان یکی از روش‌هایی که فرض‌های محدودکننده روی توزیع جامعه ندارد با روش دالینیوس-هاجز مقایسه می‌کنیم. نتایج بررسی‌ها در این مقاله نشان می‌دهد که روش خوشه‌بندی فازی در داده‌های چوله روشی کاراتر است که می‌تواند در طرح‌های آمارگیری مرکز آمار ایران به کار برده شود.

۱- مقدمه

طبقه‌بندی واحدهای جامعه با استفاده از یک یا چند متغیر طبقه‌بندی مرتبط با پارامتر(های) مورد نظر، یک روش بسیار معمول برای انتخاب واحدهای نمونه‌گیری در آمارگیری‌های نمونه‌ای است. با استفاده از این روش، واحدهای جامعه به طبقات همگن طبقه‌بندی شده، سپس واحدهای نمونه‌گیری بر اساس یک خط مشی انتساب، متناسب با اطلاعات در دسترس از اندازه‌ی طبقات، پراکندگی صفت در هر طبقه یا هزینه‌ی آمارگیری انتخاب می‌شوند. با این کار سعی می‌شود تا واحدهای نمونه‌گیری انتخاب‌شده نمایانگر بهتری از جامعه‌ی مورد آمارگیری باشند. به این ترتیب با اندازه‌ی نمونه‌ی

واژگان کلیدی: طرح نمونه‌گیری طبقه‌بندی‌شده؛ روش طبقه‌بندی دالینیوس-هاجز؛ خوشه‌بندی فازی.

* نویسنده‌ی عهده‌دار مکاتبات

یکسان، برآورد دقیق‌تری نسبت به نمونه‌گیری تصادفی ساده ارائه می‌کند. یک روش ساده و بسیار کاربردی برای طبقه‌بندی توسط دالینوس و هاجز [۴] معرفی شده است که عموماً توسط طراحان آمارگیری‌ها به کار گرفته می‌شود. در این مقاله نشان می‌دهیم که انحراف از فرض‌های این روش که معمولاً در داده‌های واقعی اتفاق می‌افتد، می‌تواند همگن بودن طبقه‌ها و همچنین مناسب بودن طبقه‌بندی برای انتساب نیمین را خدشه‌دار کند.

در بخش ۲ به بررسی روش طبقه‌بندی دالینوس-هاجز [۴] و فرض‌های این روش می‌پردازیم. بخش ۳ به مرور روش خوشه‌بندی فازی اختصاص دارد و در بخش ۴ نتایج مربوط به مقایسه‌ی طبقه‌بندی‌های حاصل از به‌کارگیری دو روش بر روی داده‌های واقعی با توزیع‌های متفاوت و دو مجموعه از داده‌های تولید شده از توزیع‌های نرمال و یکنواخت ارائه می‌شود.

۲- روش طبقه‌بندی دالینوس-هاجز

دالینوس [۳] یک روش طبقه‌بندی را برای انتخاب بهترین مرزها برای انتساب نیمین و با معلوم بودن تعداد طبقه‌ها ارائه داد. در طبقه‌بندی به‌روش دالینوس توزیع فراوانی داده‌ها به‌عنوان اصلی‌ترین ویژگی، مورد استفاده قرار می‌گیرد. در این روش اگر مقدارهای متغیر اصلی مورد بررسی در دست نباشند آن‌گاه طبقه‌بندی بر اساس متغیر دیگری که دارای همبستگی زیاد با متغیر اصلی است انجام می‌شود. در این مرحله برای معرفی روش، فرض می‌کنیم مقدارهای متغیر اصلی، y ، موجود است. همچنین تعداد طبقه‌ها، L ، معلوم و y_0 و y_L به‌عنوان کوچک‌ترین و بزرگ‌ترین مقدارهای y در جامعه در نظر گرفته می‌شوند. در این طبقه‌بندی معیار پیدا کردن مرز سایر طبقه‌ها، $(y_1, y_2, \dots, y_{L-1})$ ، مینیمم کردن مقدار $\text{var}(\bar{y}_{st})$ (واریانس برآوردگر میانگین جامعه) است. از آن‌جا که مینیمم واریانس برای انتساب نیمین با تعداد کل نمونه‌ی ثابت n در نمونه‌گیری تصادفی طبقه‌بندی شده عبارت است از

$$\text{var}(\bar{y}_{st}) = \frac{1}{n} \left(\sum_{h=1}^L W_h S_h \right)^2 + \frac{1}{N} \sum_{h=1}^L W_h S_h^2$$

که در آن S_h^2 واریانس درون طبقه‌ی h ام، N اندازه‌ی جامعه، $W_h = N_h/N$ و N_h اندازه‌ی طبقه‌ی h ام است، بنا بر این کافی است برای مینیم کردن $\text{var}(\bar{y}_{st})$ به مینیم کردن $\sum_{h=1}^L W_h S_h$ بپردازیم. دالینوس و هاجز در سال ۱۹۵۹ یک روش تقریبی برای مینیم کردن رابطه‌ی بالا ارائه دادند.

رابطه‌ی $Z(y) = \int_y^y \sqrt{f(t)} dt$ را در نظر بگیرید. با فرض این‌که $f(y)$ تابع فراوانی پیوسته‌ی متغیر y در جامعه باشد، آن‌گاه اگر طول طبقه‌ها کم و تعداد آن‌ها بسیار زیاد باشد انتظار می‌رود که $f(y)$ در داخل هر طبقه یک مقدار ثابت شود. بنا بر این

$$W_h = \int_{y_{h-1}}^{y_h} f(t) dt = f_h (y_h - y_{h-1})$$

$$S_h = \frac{1}{\sqrt{12}} (y_h - y_{h-1})$$

$$Z_h - Z_{h-1} = \int_{y_{h-1}}^{y_h} \sqrt{f(t)} dt = \sqrt{f_h} (y_h - y_{h-1})$$

که در آن f_h مقدار $f(y)$ در فاصله‌ی $[y_{h-1}, y_h]$ است. با جایگذاری مقادیرهای بالا داریم:

$$\sqrt{12} \sum_{h=1}^L W_h S_h = \sum_{h=1}^L f_h (y_h - y_{h-1}) = \sum_{h=1}^L (Z_h - Z_{h-1})$$

از فرض‌ها و رابطه‌ی بالا پیداست که مقدار $\sum_{h=1}^L W_h S_h$ زمانی مینیم خواهد شد که $(Z_h - Z_{h-1})$ برای $h = 1, 2, \dots, L$ مقدار ثابت باشد. بنا بر این قاعده‌ی ساختن مرز طبقه‌ها به این صورت خواهد بود که با در دست بودن $f(y)$ ، مقدار تجمعی $\sqrt{f(y)}$ را ساخته، y_h ها را طوری انتخاب می‌کنیم که فاصله‌های مساوی را روی مقادیرهای تجمعی $\sqrt{f(y)}$ اختیار کنند.

در صورتی‌که طول طبقه‌های مفروض، در توزیع اصلی y نامساوی باشند، الگوی بالا مقداری تغییر می‌کند. این حالت زمانی پیش می‌آید که فراوانی بعضی از طبقه‌های اولیه صفر باشد و ناچار به ادغام طبقه‌ها باشیم. به این ترتیب طبقه‌های اولیه با طول‌های متفاوت تولید می‌شود. اگر فاصله از یک طول به اندازه‌ی d به طول به اندازه‌ی ud تغییر

کند آن گاه مقدار $\sqrt{f}$ برای فاصله‌ی دوم در مقدار $\sqrt{u}$ ضرب شده، سپس آن را در مجموع تجمعی به کار می‌برند. برای آگاهی از جزئیات بیشتر این روش می‌توان به [۲] مراجعه کرد.

۳- مروری بر خوشه‌بندی فازی

روش‌های شناخته شده‌ی خوشه‌بندی داده‌ها مثل خوشه‌بندی افرازی، هر واحد جامعه را بر اساس یک تابع معیار مشخص به‌طور قطعی به یک خوشه نسبت می‌دهد. بزدک [۱] با یک نگاه جدید به روش تابعی پیشنهاد کرد که به‌طور خودکار فضای داده‌ها را به ذرات فازی تقسیم‌بندی کند. این ذرات فازی در واقع ابزار کار یک سیستم خوشه‌بندی فازی هستند. در یک مدل فازی، تابع‌های عضویت همان نقش تابع‌های معیار در روش‌های متداول خوشه‌بندی را ایفا می‌کنند. به این ترتیب تفاوت عمده‌ی خوشه‌بندی افرازی با روش فازی در این است که در روش فازی یک واحد آماری به‌طور قطع به یک خوشه متناسب نمی‌شود و بر اساس مقدار تابع عضویت می‌تواند با درجه‌های عضویت مختلف به چند خوشه‌ی متفاوت تعلق داشته باشد.

در روش بزدک [۱] که درجه‌ی عضویت داده‌ی z ام $(j = 1, \dots, n)$ به خوشه‌ی i ام، $(i = 1, \dots, c)$ ، به صورت $u_{i,j} \in [0, 1]$ تعریف می‌شود، با فرض معلوم بودن تعداد خوشه، به دنبال مجموعه‌ای از $u_{i,j}$ ها هستیم که تابع $J_m = \sum_{j=1}^n \sum_{i=1}^c u_{i,j}^m d_E^v(x_j, p_i)$ را مینیمم کند. x_j بردار p -بعدی است که مقدارهای p متغیر طبقه‌بندی برای عضو z ام را شامل می‌شود، d_E نشان‌دهنده‌ی فاصله‌ی اقلیدسی بین x_j و p_i است و p_i مرکز خوشه‌ی i ام است. مقدار p_i و $u_{i,j}$ از رابطه‌های زیر به دست می‌آیند:

$$p_i = \frac{\sum_{j=1}^n u_{i,j}^m x_j}{\sum_{j=1}^n u_{i,j}^m} \quad \text{و} \quad u_{i,j} = \frac{1}{\sum_{k=1}^c \left(\frac{d_E^v(x_j, p_i)}{d_E^v(x_j, p_k)} \right)^{\frac{1}{m-1}}}$$

برای مینیمم کردن تابع J_m قیدهای زیر نیز باید برقرار باشند

$$0 < \sum_{j=1}^n u_{i,j} < n \quad \text{و} \quad \sum_{i=1}^c u_{i,j} = 1$$

m یک مقدار ثابت در محدوده‌ی $(1, \infty)$ است که به‌طور دلخواه انتخاب می‌شود و مقدارهای عضویت را تحت تأثیر قرار می‌دهد.

مقدارهای $u_{i,j}$ و p_i هر دو نامعلوم هستند و همان‌طور که از رابطه‌های بالا پیداست، لازم و ملزوم یکدیگرند. لذا مینیمم کردن تابع J_m فقط به‌طور تکراری و با انتخاب یک مقدار اولیه‌ی تصادفی برای مجموعه‌ی مقدارهای $u_{i,j}$ انجام می‌شود. مراحل انجام محاسبات را می‌توان به‌صورت زیر خلاصه کرد:

۱- انتساب یک مقدار اولیه (اغلب تصادفی) به $u_{i,j}$ ها و ساخت ماتریس

$$U(\cdot) = [u_{i,j}]$$

۲- محاسبه‌ی مرکز خوشه‌ها، p_i ها، در مرحله‌ی k ام با استفاده از مقدارهای $u_{i,j}$ و

$$P(k) = [p_i]$$

۳- به‌هنگام کردن $U(k)$ و $U(k+1)$.

۴- اگر $\|U(k+1) - U(k)\| < \varepsilon$ برنامه متوقف می‌شود و در غیر این صورت به

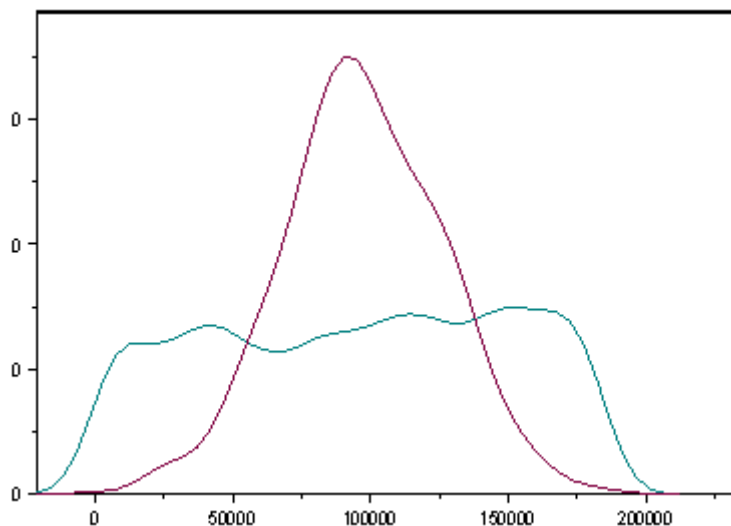
مرحله‌ی ۲ می‌رود، که ε یک آستانه‌ی از پیش تعیین شده و $\|\cdot\|$ نماد فاصله‌ی دو ماتریس است.

برای اطلاع بیش‌تر از جزئیات و مراحل تکرار آن می‌توان به [۵] رجوع کرد. در مقاله‌ی حاضر با توجه به این که خوشه‌بندی فقط با یک متغیر انجام می‌شود، فاصله‌ی اقلیدسی عبارت است از همان قدر مطلق تفاضل متغیر طبقه‌بندی از مرکز خوشه. همان‌طور که از روش پیداست، خوشه‌بندی فازی بر خلاف روش دالینیوس-هاجز هیچ فرضی در مورد توزیع داده‌ها اعمال نمی‌کند، بلکه پایه‌ی این روش مقدارهای مشابهت بین مشاهده‌ها در یک خوشه است که بر اساس یک تابع فاصله به دست می‌آید. همین تفاوت دو روش در فرض‌های آن‌ها در مورد توزیع داده‌ها موجب شده است تا در این مقاله کارایی این روش‌ها برای طبقه‌بندی مشاهده‌ها به‌منظور به‌کارگیری انتساب نیمن در آمارگیری‌های نمونه‌ای طبقه‌بندی‌شده را مورد مقایسه قرار دهیم. لازم به یادآوری است که آنچه در روش‌های چندمتغیره‌ی آماری به‌عنوان خوشه شناخته می‌شود همان

ویژگی‌های مطلوب طبقه در طرح‌های نمونه‌گیری را دارد. به این ترتیب در هر دو روش خوشه‌بندی و طبقه‌بندی به دنبال ساخت گروه‌های جدا از هم متشکل از واحدهای متجانس هستیم و تفاوت فقط در نام‌گذاری است.

برای مقایسه‌ی روش‌های طبقه‌بندی دالینیوس-هاجز با خوشه‌بندی فازی دو گروه از داده‌ها مورد استفاده قرار گرفت. نخست برای ارزیابی دور شدن از فرض توزیع یکنواخت در حالی که داده‌ها همچنان دارای توزیع متقارن باشند، دو سری داده‌های متقارن با اندازه‌ی $N = 750$ و با دامنه‌ی یکسان، یکی از توزیع نرمال و دیگری از توزیع یکنواخت (نمودار ۱) تولید شد. اندازه‌ی N از آن جهت ۷۵۰ انتخاب شده است تا با اندازه‌ی مجموعه داده‌های واقعی که در مرحله‌ی بعد استفاده می‌شود برابر و مقایسه‌پذیر باشد.

از سوی دیگر برای بررسی حساسیت و همچنین میزان کارایی طبقه‌بندی دالینیوس-هاجز برای داده‌های چوله، پنج سری از داده‌های واقعی از سرشماری کشاورزی سال ۱۳۸۲ ایران که مربوط به پنج محصول باغی متفاوت با چولگی‌های متفاوت (۱ و ۳ و ۶/۵ و ۱۵ و ۲۱) هستند، انتخاب شده که هر یک از پنج مجموعه‌ی داده‌ها با استفاده از هر یک از روش‌های طبقه‌بندی و بر اساس اطلاع محصول متناظر (که در چارچوب موجود است) به پنج طبقه تقسیم‌بندی شدند.



شکل ۱ - نمودار داده‌های نرمال و یکنواخت شبیه‌سازی شده

برای مقایسه‌ی روش‌های طبقه‌بندی از سه معیار متفاوت استفاده شده است:

(آ) **عدم تجانس‌ها:** در هر یک از طبقه‌های ساخته شده، میانگین معیار عدم تجانس اعضای آن طبقه محاسبه شده است و همچنین میانگین معیار عدم تجانس برای تمام اعضای هر یک از جامعه‌های مورد بررسی نیز برای مقایسه‌ی کلی به کار رفته است. اگر y_{ih} مقدار واحد i ام جامعه در طبقه‌ی h ام باشد مقدار عدم تجانس برای y_{ih} برابر است با

$$d_{ih} = \sum_{j \neq i} |y_{ih} - y_{jh}| \quad (i, j = 1, 2, \dots, N_h)$$

که N_h تعداد واحدهای جامعه در طبقه‌ی h ام است.

(ب) **فرض برابری $N_h S_h$ ها:** همان‌گونه که در بخش ۲ اشاره شد یکی از هدف‌های مهم طبقه‌بندی دالینیوس-هاجز ساختن طبقه‌هایی است که تا حد ممکن مقدار $N_h S_h$ های مساوی داشته باشند. برای مقایسه‌ی میزان برابری $N_h S_h$ ها بین طبقه‌های ساخته شده از هر یک از روش‌های طبقه‌بندی، میزان عدم تجانس بین $N_h S_h$ ها را مورد مقایسه قرار داده‌ایم. عدم تجانس بین $N_h S_h$ ها نیز به همان روشی که در بالا توضیح داده شد محاسبه می‌شود. یعنی عدم تجانس برای $N_h S_h$ عبارت است از:

$$d_h = \sum_{k \neq h} |N_k S_k - N_h S_h| \quad (k, h = 1, 2, \dots, L)$$

(پ) **خطای استاندارد:** برای نشان دادن میزان کارایی روش‌های طبقه‌بندی در برآورد مقدار کل صفت مورد نظر، خطای استاندارد برآورد با استفاده از یک نمونه‌ی ۱۰۰ تایی و ۱۰۰۰ تکرار محاسبه شده است. به این ترتیب پس از انتساب بهینه‌ی نمونه‌ی ۱۰۰ تایی به طبقه‌ها در هر یک از روش‌های طبقه‌بندی و ۱۰۰۰ بار تکرار نمونه‌گیری، خطای استاندارد برآورد مقدار کل صفت در هر روش، با استفاده از برآوردهای مقدار کل حاصل از همه‌ی تکرارها به‌سادگی قابل محاسبه است.

۴- نتایج مقایسه‌ی روش‌ها

نتایج مقایسه‌ها نشان می‌دهد همان‌طور که انتظار می‌رفت از آن‌جا که روش طبقه‌بندی دالینیوس-هاجر همواره، تا حد امکان، به دنبال برابر نگه داشتن مقدارهای $N_h S_h$ در بین طبقات است، در داده‌های چوله کارایی طرح نمونه‌گیری را به‌شدت تحت تأثیر قرار می‌دهد.

جدول ۱- میانگین عدم تجانس‌های مقادیر $N_h S_h$ و میانگین S_h ها به تفکیک توزیع متغیر طبقه‌بندی (محصول)

میانگین عدم تجانس	میانگین انحراف استاندارد	طبقه‌بندی	نوع داده
۱۶۲۴۲۶۸	۱۰۶۹۳	فازی	یکنواخت
۸۶۷۴۶	۱۰۹۲۶	دالینیوس-هاجر	
۲۶۰۴۰۲۴	۸۱۹۱	فازی	نرمال
۲۶۷۶۷۵	۸۷۰۳	دالینیوس-هاجر	
۳۲۴۶۹۵	۵۱۵۶	فازی	$X(1)$
۸۲۱۱۴	۵۱۷۲	دالینیوس-هاجر	
۲۳۹۶۰	۳۲۳	فازی	$X(3/6)$
۱۲۰۱۴	۳۴۴	دالینیوس-هاجر	
۶۱۱۱۷۱	۵۵۹۸	فازی	$X(6/5)$
۱۲۳۱۷۶	۶۷۶۷	دالینیوس-هاجر	
۱۲۶۶۵۲۰۰	۱۱۶۸۲۶	فازی	$X(15)$
۴۸۷۴۲۷۳	۱۸۲۱۰۲	دالینیوس-هاجر	
۷۴۶۴۰۰	۱۹۹۸۳	فازی	$X(21)$
۸۹۵۱۶۷	۵۹۴۶۵	دالینیوس-هاجر	

همان‌طور که در جدول ۱ مشاهده می‌شود، مقدار میانگین انحراف استاندارد درون طبقه‌ها برای طبقه‌بندی حاصل از روش فازی در همه‌ی داده‌ها از روش دالینیوس-هاجر کمتر است. دلیل این امر را می‌توان با مشاهده‌ی مقدار میانگین عدم تجانس‌های

بین $N_h S_h$ ها بررسی کرد. این مقدار در مورد همه‌ی انواع داده‌ها برای روش دالینیوس-هاجز بسیار کوچکتر از مقدار حاصل از روش فازی است. به این ترتیب مشاهده می‌شود که قید برابری $N_h S_h$ ها در روش دالینیوس-هاجز موجب کاهش میزان همگنی (تجانس) درون طبقه‌ها شده است که نتیجه‌ی این ناهمگنی را می‌توان در مقدار میانگین انحراف استاندارد درون طبقه‌ها مشاهده کرد. همان‌گونه که در مورد داده‌های $X(21)$ قابل مشاهده است، در مواردی که داده‌ها چولگی زیادی داشته باشند، این موضوع بیشتر اهمیت خود را نشان می‌دهد. در مورد $X(21)$ میانگین انحراف استاندارد درون طبقه‌ها برای روش دالینیوس-هاجز تقریباً ۳ برابر روش فازی است. در این شرایط روش طبقه‌بندی می‌تواند به شدت بر دقت طرح نمونه‌گیری اثر بگذارد.

بررسی معیار عدم تجانس نشان می‌دهد که روش خوشه‌بندی فازی به‌طور قابل توجهی معیار عدم تجانس را در داخل طبقه‌ها نسبت به روش طبقه‌بندی دالینیوس-هاجز کاهش می‌دهد (جدول ۲). این کاهش با میزان چولگی داده‌ها تقریباً نسبت عکس دارد یعنی هرچه به میزان چولگی داده‌ها افزوده می‌شود داده‌های قرار گرفته درون یک طبقه در روش دالینیوس-هاجز نسبت به روش خوشه‌بندی فازی نامتناسج‌تر می‌شوند. در جدول ۲ ستون تفاضل عبارت است از تفاضل عدم تجانس روش فازی از عدم تجانس روش دالینیوس-هاجز. لذا مقدارهای مثبت (خانه‌های با رنگ تیره) در این ستون بیانگر این هستند که داده‌های حاصل از طبقه‌بندی دالینیوس-هاجز در آن طبقه نسبت به داده‌های حاصل از طبقه‌بندی فازی در طبقه‌ی متناظر از تجانس کمتری برخوردار هستند. همان‌طور که مشاهده می‌شود برای داده‌های با چولگی بالاتر (۶/۵ و ۱۵ و ۲۱) در همه‌ی طبقه‌ها عدم تجانس روش طبقه‌بندی دالینیوس-هاجز بیش از روش فازی است به‌طوری که به جز برای ۲ طبقه، در سایر طبقه‌ها این تفاوت تقریباً بیش از ۲ برابر است و در یک مورد به ۸ برابر هم رسیده است.

جدول ۲- تفاضل میانگین عدم تجانس‌ها برای روش‌های طبقه‌بندی دالینیوس-هاجز و فاززی در هر طبقه به‌تفکیک توزیع داده‌ها

یکنواخت				نرمال			
طبقه	دالینیوس-هاجز	فازی	تفاضل	طبقه	دالینیوس-هاجز	فازی	تفاضل
۱	۱۳۰۷۰	۱۰۵۱۵	۲۵۵۵	۱	۱۳۹۵۸	۱۴۴۸۰	-۵۲۲
۲	۱۳۳۱۵	۱۲۴۷۸	۸۳۷	۲	۷۲۲۶	۵۷۴۳	۱۴۸۳
۳	۱۲۶۲۲	۱۳۸۱۸	-۱۱۹۶	۳	۶۲۴۳	۵۶۶۰	۵۸۳
۴	۱۲۴۷۳	۱۳۳۳۵	-۸۶۲	۴	۷۸۵۱	۵۷۸۱	۲۰۷۰
۵	۱۱۸۸۲	۱۱۸۸۲	۰	۵	۱۲۷۲۴	۱۳۳۱۵	-۴۹۱

$X(۱)$				$X(۳/۶)$			
طبقه	دالینیوس-هاجز	فازی	تفاضل	طبقه	دالینیوس-هاجز	فازی	تفاضل
۱	۳۱۳۴	۳۰۰۰	۱۳۴	۱	۳۷	۲۲	۱۵
۲	۴۰۴۹	۴۲۸۷	-۲۳۸	۲	۱۲۴	۷۶	۴۸
۳	۶۰۴۹	۵۶۴۸	۴۰۱	۳	۳۷۱	۲۷۹	۹۲
۴	۶۹۲۰	۷۰۵۴	-۱۳۴	۴	۴۲۴	۴۷۱	-۴۷
۵	۹۸۱۰	۹۸۸۴	-۷۴	۵	۹۸۹	۹۸۹	۰

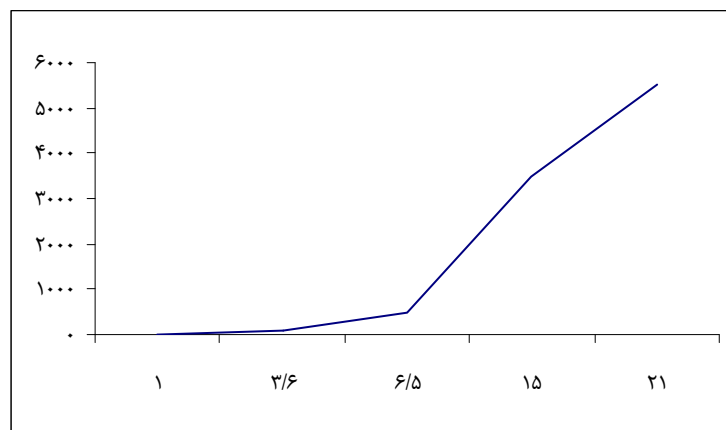
$X(۱۵)$				$X(۶/۵)$			
طبقه	دالینیوس-هاجز	فازی	تفاضل	طبقه	دالینیوس-هاجز	فازی	تفاضل
۱	۲۷۵۷	۹۷۱	۱۷۸۶	۱	۲۷۰	۹۷	۱۷۳
۲	۱۷۹۸۷	۴۸۵۸	۱۳۱۲۹	۲	۱۵۴۶	۵۳۱	۱۰۱۵
۳	۵۳۱۵۳	۲۳۲۸۸	۲۹۸۶۵	۳	۴۱۶۳	۱۷۰۸	۲۴۵۵
۴	۵۲۵۷۷	۵۰۰۶۴	۲۵۱۳	۴	۷۲۸۷	۴۷۵۸	۲۵۲۹
۵	۹۲۴۱۷۸	۳۷۱۷۶۴	۵۵۲۴۱۴	۵	۲۷۸۶۴	۲۳۰۸۴	۴۷۸۰

$X(۲۱)$			
طبقه	دالینیوس-هاجز	فازی	تفاضل
۱	۷۱۶	۲۶۲	۴۵۴
۲	۲۰۱۹	۶۳۸	۱۳۸۱
۳	۶۱۶۶	۱۵۴۱	۴۶۲۵
۴	۱۹۳۳۹	۲۳۸۲	۱۶۹۵۷
۵	۳۸۵۰۴۲	۴۷۲۷۲	۳۳۷۷۰

در نهایت، همان موارد بالا را می‌توان به‌خوبی در مقایسه‌ی واریانس دو روش مشاهده کرد. جدول ۳ برآورد واریانس دو روش برای برآورد صفت مجموع و همچنین کارایی روش خوشه‌بندی فازی نسبت به طبقه‌بندی دالینوس-هاجز را نشان می‌دهد. از این جدول روشن است که در مورد داده‌های با توزیع‌های متقارن یا با چولگی بسیار کم تفاوت چندانی در کارایی حاصل از دو روش مشاهده نمی‌شود، اما برای داده‌های چوله روش فازی به‌شدت کاراتر است و با افزایش میزان چولگی روش فازی با نرخ فزاینده‌تری کارایی خود را نشان می‌دهد. نمودار ۲ که از این اطلاعات حاصل شده است، به‌خوبی نشان می‌دهد که هرچه بر میزان چولگی داده‌ها افزوده شود، کارایی روش خوشه‌بندی فازی نیز افزایش می‌یابد. یکی از عواملی که موجب کاراتر شدن روش فازی می‌شود این است که در جریان انتساب بهینه‌ی اندازه نمونه به طبقه‌ها، معمولاً طبقه(ها)ی آخر سرشماری می‌شود و این امر موجب کاراتر شدن روش فازی نسبت به دالینوس-هاجز شده است.

جدول ۳- واریانس برآورد صفت مجموع برای دو روش طبقه‌بندی فازی و دالینوس-هاجز و کارایی روش فازی نسبت به روش دالینوس-هاجز

داده‌ها	واریانس		کارایی نسبی	چولگی
	دالینوس-هاجز	فازی		
یکنواخت	۵۸۸۸۶۳۱۶۸۹۳۳	۵۸۲۵۳۲۱۴۵۰۰۴	۱	-
نرمال	۲۸۰۲۰۹۸۵۱۴۹۲	۳۱۶۶۶۸۴۹۴۰۲۴	۰/۹	-
$X(1)$	۹۳۲۳۷۴۴۹۵۲۵	۹۶۴۸۸۷۶۰۷۵۴	۱	۱
$X(3/6)$	۳۹۴۸۷۳۶۹۷۰	۳۵۸۳۰۸۹۷	۱۱۰	۳/۶
$X(6/5)$	۵۰۰۰۵۰۱۰۹۰۹۶	۱۰۷۷۶۱۷۹۸۳	۴۶۴	۶/۵
$X(15)$	۱۵۹۶۵۴۸۷۶۳۴۶۹۵۵	۴۶۰۹۰۹۹۶۲۶۸	۳۴۶۴	۱۵
$X(21)$	۲۶۴۳۶۸۶۱۸۲۱۲۸۵	۴۷۹۵۳۰۶۳۳۸	۵۵۱۳	۲۱



شکل ۲- نمودار کارایی روش خوشه‌بندی فازی نسبت به روش دالینیوس-هاجز در مقابل مقدار چولگی داده‌ها

به‌طور خلاصه می‌توان نتیجه گرفت که در انحراف از توزیع یکنواخت در صورتی که تقارن توزیع حفظ شود، همچنان روش دالینیوس-هاجز قابل قبول است. با توجه به این که مرکز آمار ایران در عمل با آمارگیری از جوامعی مواجه است که چوله هستند (به‌طور مثال متغیرهای مربوط به طرح‌های کشاورزی که در چارچوب نمونه‌گیری نیز در دسترس بوده، عمدتاً به‌عنوان متغیر طبقه‌بندی مورد استفاده قرار می‌گیرند) و همچنین با استفاده از نرم‌افزارهای آماری در دسترس، بهره‌گیری از روش‌های مختلف خوشه‌بندی امکان‌پذیر است، پیشنهاد می‌شود در طرح‌های نمونه‌گیری طبقه‌بندی شده در صورت امکان از روش‌هایی استفاده شود که فرض‌های محدودکننده بر توزیع داده‌ها نداشته باشند. روش خوشه‌بندی فازی یکی از این قبیل روش‌ها است که در صورت به‌کارگیری می‌تواند موجب افزایش همگنی درون طبقات و در نتیجه افزایش دقت طرح شود.

مرجع‌ها

- [1] Bezdek, J. C. (1980). A convergence theorem for the fuzzy ISODATA clustering algorithms, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2, 1-8.
- [2] Cochran, W. G. (1977). *Sampling Techniques (Third Edition)*, John Wiley and Sons, New York.
- [3] Dalenius, T. (1950). The problem of optimum stratification. *Skand. Akt.*, 33,

203-213.

- [4] Dalenius, T. and Hodges, J. L. (1959). Minimum Variance Stratification, *Journal of the American Statistical Association*, **54**, 88-101.
- [5] Nascimento, S. Mirkin, B. and Moura Pires, F. (2000). A Fuzzy Clustering Model of Data and Fuzzy c-Means, The Ninth IEEE International Conference on Fuzzy Systems: Soft Computing in the Information Age (FUZZ-IEEE 2000), IEEE Neural Networks Council, San Antonio, TX, May 2000, 302-307.