تحلیل خوشه‌بندی فازی داده‌های ترکیبی و مقایسه آن با دندروگرام اکتشافی داده‌های ترکیبی ژئوشیمی رسوبات آبراهه‌ای منطقه انار

حمید معینی*[۱]، فرهاد محمدتراب[۲]، مجید کیخای حسین‌پور[۳]

۱- دانشجوی دکتری، دانشکده مهندسی معدن و متالورژی، دانشگاه یزد

۲- استادیار، دانشکده مهندسی معدن و متالورژی، دانشگاه یزد

۳- دانشجوی دکتری، دانشکده مهندسی معدن و متالورژی، دانشگاه یزد

(دریافت: بهمن ۱۳۹۳، پذیرش: دی ۱۳۹۵)

چکیده

از روش‌های مهم در داده‌کاوی نظارت نشده داده‌های ژئوشیمیایی، انواع روش‌های خوشه‌بندی است که چنانچه روی متغیرها انجام شوند منجر به کاهش ابعاد داده‌ها می‌شوند. در میان انواع روش‌های خوشه‌بندی، نوع فازی آن به دلیل ویژگی‌های خاص منطق فازی و انعطاف بیشتر در تعیین گروه‌های داده مشابه، در سالیان اخیر بسیار مورد توجه قرار گرفته است. در این پژوهش از الگوریتم فازی منعطف به نام *FANNY* به منظور خوشه‌بندی متغیرهای داده‌های ژئوشیمی رسوبات آبراهه‌ای که خاصیت ترکیبی دارند، استفاده شده است. با تحقیقات گسترده محققان علم آمار و ارائه روش‌های جدید بازکردن داده‌های ترکیبی، مشخص شده است که فاصله‌ها و روابط دیگری بر فضای این نوع داده‌ها حاکم است که برای درک بهتر آنها نیاز به انتقال ایزومتریک به فضای اقلیدسی است تا قابل استفاده و تفسیر با روابط کلاسیک آماری باشند. در پژوهش حاضر، پس از آماده‌سازی داده‌های ژئوشیمی رسوبات آبراهه‌ای منطقه انار کرمان (به عنوان مثالی از داده‌های ترکیبی با ابعاد زیاد) ابتدا دندروگرام اکتشافی روی متغیرها در فضای سیمپلکس و با استفاده از پارتیشن دودوئی ترتیبی(*SBP*) پیش فرض، محاسبه و ترسیم شد که با بکارگیری این روش، تعداد ۴ خوشه با متغیرهای مشابه شناسایی شد. سپس دوباره با استفاده از الگوریتم *fanny*، همان متغیرهای داده‌های باز شده با تبدیل *clr* خوشه‌بندی شد. نتایج خوشه‌بندی متغیرها با الگوریتم *fanny* انطباق قابل قبولی با دندروگرام اکتشافی داده‌های ترکیبی نشان داد. در صورتی که *SBP* مورد نیاز برای بالانس‌های دندروگرام اکتشافی در مختصات ایزومتریک با شناخت کامل‌تر از متغیرها و نه بصورت پیش فرض تعیین شود نتایج دندروگرام دقت بسیار بهتری خواهد داشت.

کلید واژه‌ها

رسوبات آبراهه‌ای، منطقه انار، الگوریتم *fanny*، دندروگرام داده‌های ترکیبی، فاصله آیچیسون، تبدیل *clr*

ارجاع به این مقاله:

معینی، ح.، محمدتراب، ف.، کیخای حسین‌پور، م.، (۱۳۹۵)، تحلیل خوشه‌بندی فازی داده‌های ترکیبی و مقایسه آن با دندروگرام اکتشافی داده‌های ترکیبی ژئوشیمی رسوبات آبراهه‌ای منطقه انار، روش‌های تحلیلی و عددی در مهندسی معدن، ۶(۱۲)، ۱۱-۱۹.

* عهده دار مکاتبات: *hamidmoini@stu.yazd.ac.ir*

۱- مقدمه

امروزه اکتشافات ژئوشیمیائی بخشی از طیف وسیع روش‌های اکتشافات معدنی محسوب می‌شود که دارای جایگاه ویژه‌ای در ارزیابی پتانسیل‌های اقتصادی هر منطقه است[۱]. یکی از محیط‌های تحت پوشش اکتشافات ژئوشیمیایی، محیط رسوبات رودخانه‌ای است که یک روش مستقل و مفید برای تشخیص نواحی با پتانسیل بالای معدنی است[۲].

از جمله روش‌هایی که کاربرد فراوانی در مطالعات اکتشافی و بررسی الگوی ژئوشیمیایی عناصر دارد، تحلیل‌های آماری چند متغیره است. این روش‌ها به بررسی تغییرات همزمان چند متغیر و استنباط آماری ناشی از آن می‌پردازد و به دلیل داشتن خطای کمتر و اعتبار بیشتر نسبت به روش‌های آماری تک متغیره یا دو متغیره، کاربرد فراوانی در مطالعات علوم زمین دارد[۳]. از جمله روش‌های داده‌کاوی چند متغیره در مسائل اکتشافی و به خصوص در اکتشافات ژئوشیمیایی می‌توان به روش‌های تحلیل مولفه‌های اصلی، تحلیل فاکتوری و انواع روش‌های خوشه‌بندی اشاره کرد که به طور موفقیت آمیزی توسط محققان مختلف بکار گرفته شده است[۴-۱۵].

نکته‌ای که در مطالعات پیشین کمتر به آن پرداخته شده است، عدم توجه به ماهیت ترکیبی بودن آنالیزهای ژئوشیمیایی است[۱۶، ۱۷]. امروزه ثابت شده است که داده‌های ژئوشیمیایی خاصیت ترکیبی دارند. بدین معنی که متغیرهای اندازه‌گیری شده برای هر نمونه از یکدیگر مستقل نبوده و در واقع بخشی از کل هستند[۱۸]. مطالعات بسیاری به منظور بررسی خواص این نوع داده‌ها و نحوه تحلیل هرچه مناسب‌تر آنها در سالیان اخیر صورت گرفته است[۱۹-۲۹]. در این پژوهش خوشه‌بندی داده‌های رسوبات آبراهه‌ای منطقه اکتشافی انار، با نگرشی خاص به ویژگی ترکیبی بودن داده‌های ژئوشیمیایی و با بکارگیری تحلیل خوشه‌بندی فازی و دندروگرام اکتشافی روی داده‌های ترکیبی صورت گرفته است.

۲- زمین‌شناسی منطقه مورد مطالعه

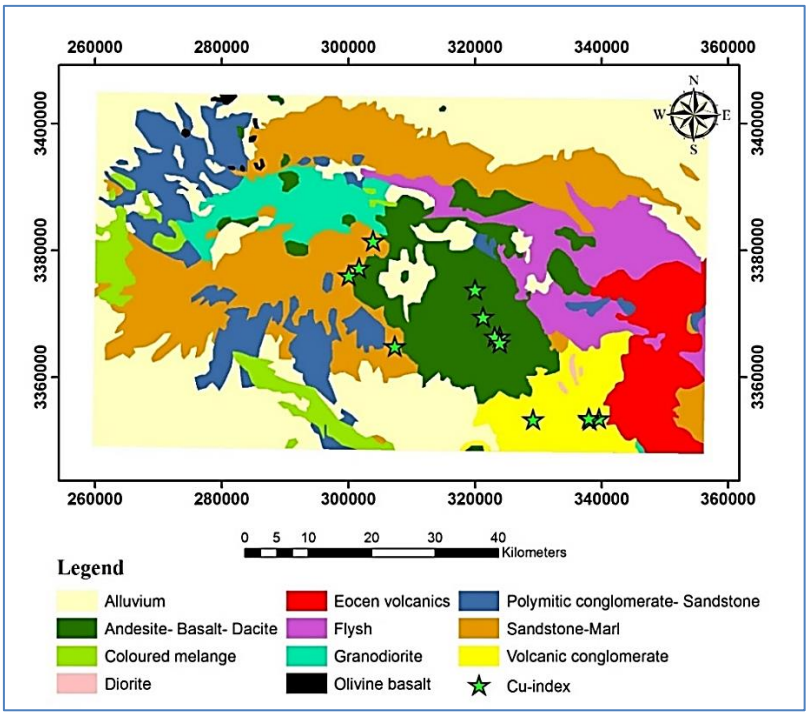
منطقه مورد مطالعه قسمتی از نقشه زمین‌شناسی ۱:۲۵۰۰۰۰ انار بین کرمان و یزد است. این محدوده به وسعت ۴۷۶۴ کیلومترمربع از نظر ساختاری در پهنه ایران مرکزی و در کمان ماگمایی ارومیه- دختر قرار دارد و شامل ترادف ضخیمی از سنگ‌های آتشفشانی و آواری- آتشفشانی ائوسن است که توسط توده‌های گرانیتوئیدی تحت تاثیر قرار گرفته و دگرسانی‌های گرمابی گسترده‌ای را متحمل شده‌اند. رخنمون‌های سنگی غالب در این کمان ماگمایی دارای ترکیب آندزیتی، تراکی آندزیتی، بازالتی تا داسیت و ریوداسیتی است که توسط توده‌های گرانودیوریت، کوارتزدیوریت و مونزونیتی الیگومیوسن و توده‌های نیمه عمیق میکروکوارتز دیوریت- داسیت پورفیری میوپلیوسن قطع شده است[۳۰]. نقشه زمین‌شناسی ساده شده منطقه انار در شکل ۱ نشان داده شده است.

توده‌های میکرو کوارتزدیوریت- داسیت جوان‌تر، میزبان کانی‌سازی پورفیری بوده که ازآن جمله می‌توان کانسارهای مس میدوک، پرکام، چاه فیروزه، ایجو، گودکلواری، سرنو و آبدر را نام برد[۳۱]. فعالیت‌های گرمابی توده‌های مذکور منجر به نفوذ شیرابه‌های سیلیسی درون شکستگی‌های سنگ‌های آتشفشانی ائوسن گردیده که بعضاً با کانی‌سازی مس، سرب، روی، طلا و نقره همراه است [۳۲].

۳- داده‌های ترکیبی

ماهیت ترکیبی یا بسته داده‌های ژئوشیمی، امروزه موضوع مهمی است که باید قبل از هر تحلیل مد نظر قرار گیرد. داده‌های بسته یا ترکیبی، مجموعه‌ای از داده‌های نسبی است که متغیرهای آن مستقل از یکدیگر نبوده و به صورت درصد یا قسمت در میلیون یا به طور کلی جزئی از کل بیان می‌شوند[۱۷].

همین خاصیت داده‌های ترکیبی سبب شده نتوان از روش‌های آماری استاندارد برای تحلیل آنها استفاده نمود. فضای اقلیدسی برای داده‌های ترکیبی مناسب نیست و محدودیت حاصل جمع ثابت این داده‌ها و نسبی بودن آنها دلالت بر هندسه خاصی دارد که در اصطلاح، هندسه آیچیسون[1] نامیده می‌شود[۱۷].



شکل ۱: نقشه زمین‌شناسی ساده‌شده محدوده مورد مطالعه

با استفاده از تبدیل لگاریتم نسبیتی مرکزی[۲] (*clr*)، می‌توان داده‌ها را به فضای اقلیدسی منتقل کرد[۲۲]. فاصله داده‌ها در این هندسه، فاصله آیچیسون[۳] نام دارد که برای دو ترکیب $X=(x_1,\ldots,x_D)$ و $Y=(y_1,\ldots,y_D)$ عبارت است از:

$$d_A(X,Y)=\sqrt{\frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D}\left(\log\frac{x_i}{x_j}-\log\frac{y_i}{y_j}\right)^2} \qquad (۱)$$

خاصیت ایزومتری تبدیل clr یعنی به ازای دو ترکیب X و Y رابطه (۲) بین دو فضای آیچیسون و اقلیدسی آنها برقرار است:

$$d_A(X,Y)=d_E(\mathrm{clr}(x),\mathrm{clr}(Y)) \qquad (۲)$$

تعریف میانگین در این نوع داده‌ها مطابق رابطه (۳) محاسبه می‌شود:

$$\overline{X}=\mathrm{clr}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{clr}(x_i)\right) \qquad (۳)$$

معیارهای مختلفی برای واریانس در داده‌های ترکیبی تعریف شده است. واریانس متریک یا واریانس کل یا واریانس عام یکی از آنهاست که متوسط مربع فاصله از مرکز داده‌ها با درجه آزادی تصحیح شده واریانس داده‌های معمولی است و از رابطه (۴) بدست می‌آید:

$$\mathrm{var}(X)=\frac{1}{n-1}\sum_{i=1}^{n}d_A^2(x_i,\overline{X}) \qquad (۴)$$

در رابطه (۴)، $\overline{X}$ میانگین هندسی ترکیب X است. بنابراین استاندارد کردن داده‌های ترکیبی نیز با روش‌های معمول آماری برای داده‌های دیگر، بسیار متفاوت است. نخست اینکه داده‌های ترکیبی یک مقیاس مشترک بدون بُعد دارند و در نتیجه با فرایند معمول استانداردکردن سبب از دست رفتن اطلاعات مهمی خواهد شد که متغیرها درباره آن بیشترین تغییرپذیری را نشان می‌دهند. دوم اینکه میانگین‌گیری معمولی، مقادیر منفی تولید می‌کند که تفسیر نتایج را با اشتباه همراه خواهد کرد. در نتیجه، استاندارد کردن این داده‌ها با تقسیم بر میانگین به توان عکس مجذور واریانس ترکیبی از رابطه (۵) بدست می‌آید:

$$Z=\frac{1}{\sqrt{\mathrm{var}(X)}}\bullet(X\otimes\overline{X}) \qquad (۵)$$

که در این رابطه، $\bullet$ عملگر توان و $\otimes$ عملگر معکوس در فضای سیمپلکس هستند[۳۳].

۴- دندروگرام داده‌های ترکیبی

به دلیل طبیعت پیچیده داده‌های ژئوشیمی، نتایج بعضی تحلیل‌ها مانند تحلیل خوشه‌ای بسیار وابسته به آماده‌سازی داده‌ها (انتخاب تبدیل صحیح) و الگوریتم خوشه‌بندی است[۱۰]. در مورد مسائلی که با ترکیب اجزاء سروکار دارند (مانند ژئوشیمی) جستجو برای روش‌های کاهش بُعد متناسب با زیرمجموعه‌های ترکیبی به استراتژی جدیدی منجر شده است که ماتریس بالانس‌ها[۴] نامیده می‌شود. این ماتریس‌ها بر اساس یک پارتیشن $D-$جزئی دودوئی ترتیبی[۵](SBP)، داده‌های ترکیبی $D-$بعدی را به گروه‌هایی غیر همپوشان تبدیل می‌کنند. مختصاتی که در نهایت به دست می‌آید، تفسیر ترکیب‌ها را آسان می‌کند و منجر به تجزیه واریانس کل به واریانس‌های جزئی می‌شود که می‌توان آنها را ناشی از تغییرپذیری درون گروهی یا بین گروهی دانست[۳۴]. با استفاده از این خاصیت داده‌های ترکیبی، ابزاری به نام دندروگرام داده‌های ترکیبی توسعه یافته است که نمایش یک SBP یا مبنای ایزومتریک متعامد همراه با خلاصه اطلاعات آماری بالانس‌هاست[۲۶، ۲۹، ۳۵].

دندروگرام داده‌های ترکیبی یک پارتیشن‌بندی سلسله مراتبی است که می‌تواند اطلاعات متغیرها را حتی زمانی که بردارهای ترکیبی اجزاء زیاد دارند خلاصه کند. معمولاً تعیین ماتریس بالانس‌های SBP به صورت پیش فرض و توسط نرم‌افزار انجام می‌شود. این کار با استفاده از آنالیز مولفه‌های اصلی روی داده‌های ترکیبی و تعیین جهات اصلی صورت می‌گیرد. اگر همبستگی و ارتباط بین اجزاء از قبل معلوم باشد (با توجه به اطلاعات زمین‌شناسی و تجربه) در این صورت ساخت ماتریس بالانس می‌تواند توسط شخص انجام شود تا تفسیر نتایج بهبود یابد. هرکدام از خطوط عمودی معرف بالانس‌ها و مقدار واریانس همان بالانس و نقطه تماس با خطوط افقی معرف میانگین آن بالانس است[۳۴].

در این پژوهش در الگوریتم خوشه‌بندی، به منظور کاهش تاثیر منفی فاصله اقلیدسی در دندروگرام ترکیبی از معیار وارد[۶] در تحلیل سلسله مراتبی استفاده شده که به کمترین واریانس نیز مشهور است[۵].

۵- خوشه بندی فازی داده های ترکیبی

یکی از روش‌های موثر خوشه‌بندی، روش تحلیل خوشه‌ای فازی است که با استفاده از یک عملگر فازی سبب انعطاف‌پذیری بیشتر در تقسیم‌بندی نمونه‌ها و تعیین خوشه‌ها می‌شود[۳۶]. در خوشه‌بندی فازی با الگوریتم *fanny* [۳۷]، هدف کمینه کردن تابع هدف (۶) است:

$$\sum_{v=1}^{k}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}u_{iv}^{r}u_{jv}^{r}d(i,j)}{2\sum_{j=1}^{n}u_{jv}^{r}} \qquad (۶)$$

که در آن، n تعداد نمونه‌ها، k تعداد خوشه‌ها، r درجه فازی‌شدگی و $d(i,j)$ معیار ناشباهت[۷] بین نمونه i و j است. u_{iv} درجه عضویت نمونه i به خوشه u است که عددی بزرگتر یا مساوی صفر است. فاصله دو نمونه یا دو متغیر می‌تواند معیار ناشباهت قرارگیرد. r عددی بزرگتر از یک است. انتخاب r در تعیین بهینه خوشه‌ها می‌تواند موثر باشد. هرچه r به سمت یک نزدیکتر باشد، خوشه‌بندی غیرفازی‌تر و هرچه به دو نزدیک شود فازی‌شدگی بیشتر می‌شود. الگوریتم با مقداردهی تصادفی به درجه عضویت‌ها شروع شده و به کمک روش‌های محاسبات عددی تا زمانی تکرار می‌شود که تابع هدف به کمترین مقدار، همگرا شود. مزیت این روش نسبت به دیگر روش‌های خوشه‌بندی فازی این است که (۱) می‌توان ماتریس شباهت را برای آن تعریف کرد، (۲) نسبت به فرض خوشه‌بندی کروی مقاوم‌تر است و (۳) معیار گرافیکی جدیدی برای پراش بین خوشه‌ای به نام سیلوئت[۷] ارائه می‌دهد[۳۷]. در این پژوهش، جهت امکان مقایسه با دندروگرام از درجه عضویت ۱/۲ استفاده شده است تا به خوشه‌بندی سخت نزدیکتر باشد.

متداول‌ترین معیار مورد استفاده برای شباهت، فاصله اقلیدسی است. در این تحقیق، به دلیل اینکه فاصله آیچیسون بین داده‌های تبدیل یافته با clr همان فاصله اقلیدسی در فضای کلاسیک است (رابطه ۲)، فاصله آیچیسون به عنوان معیار شباهت برای خوشه‌بندی متغیرها استفاده شد.

در صورتی که تعداد متغیرها در داده‌های ترکیبی زیاد باشد، کاربرد روش روی داده‌های تبدیل یافته با لگاریتم ساده، ممکن است نتایجی مشابه با فاصله آیچیسون روی

داده‌های بسته خام تولید کند. اما خوشه‌بندی فازی روی داده‌های ترکیبی خام (بسته) با معیار فاصله اقلیدسی (به جز در حالت خاص تمرکز داده‌ها در مرکز سیمپلکس)، مشکلات آمار کلاسیک در تفسیر داده‌های ترکیبی (بدلیل اثرات متقابل متغیرها بر هم) را خواهد داشت. بنابراین اهمیت استفاده از یک چهارچوب نظری صحیح برای خوشه‌بندی این نوع داده‌ها را نمی‌توان از نظر دور داشت[۳۸].

نکته مهم در تمام روش‌های خوشه‌بندی، تعیین تعداد بهینه خوشه‌ها است، که معیارهای مختلفی نیز برای آن ارائه شده است. معیار مورد استفاده در این پژوهش بر اساس بیشترین مقدار سیلوئت[۸] میزان پراش بین خوشه‌ای نشان داده می‌شود. مقدار شاخص اعتبارسنجی سیلوئت بین ۱- تا ۱ تغییر می‌کند. درصورتی که این شاخص برابر یک باشد، نمایانگر این است که خوشه‌بندی به درستی صورت گرفته است. اگر مقدار شاخص نزدیک صفر باشد، این بدین معنی است که نمونه را می‌توان به یک خوشه نزدیکتر نسبت داد و نحوه قرارگیری نمونه از دو خوشه، به یک اندازه دور است. در صورتی که این شاخص برابر ۱- باشد، این بدین معنی است که خوشه‌بندی به درستی صورت نپذیرفته است[۳۹]. این معیار در این پژوهش در حالت‌های a، b و c به ترتیب برای ۳، ۴، ۵ خوشه (داده‌های ترکیبی) مورد آزمون قرار گرفت و در مناسب‌ترین حالت، یعنی b (با بیشترین سیلوئت) ۴ خوشه به دست آمد. جدول (۱) نشان دهنده سیلوئت هر خوشه در ۳ حالت مذکور و متوسط کل آنها است.

جدول ۱: جدول مقادیر متوسط سیلوئت در الگوریتم *fanny* در سه حالت مختلف

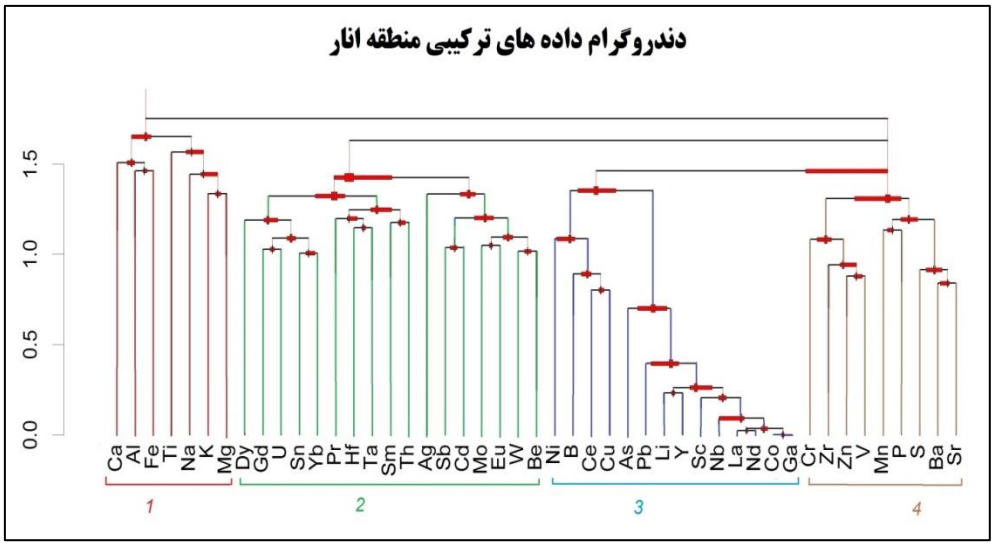
	۱	۲	۳	۴	۵	متوسط
a	۰/۴۰۹۸	۰/۶۳۷۸	۰/۷۷۸۷			۰/۵۵۲۱
b	۰/۶۱۵۹	۰/۵۰۰۸	۰/۵۴۱۹	۰/۶۹۵۲		۰/۵۷۱۹
c	۰/۵۹۰۹	۰/۴۰۴۴	۰/۵۱۰۲	۰/۵۲۷۴	۰/۶۹۵۲	۰/۵۴۷۶

۶- بحث و بررسی

برای بررسی خوشه‌بندی متغیرها، از داده‌های رسوبات آبراهه‌ای برگه ۱:۲۵۰،۰۰۰ انار استفاده شد. نمونه‌ها از جزء ۱۰۰- مش رسوبات آبراهه‌ای در محدوده‌ای به وسعت ۴۷۶۴ کیلومترمربع برداشت شده است. کیفیت آنالیزها با برداشت نمونه‌های تکراری و کنترل دقت آزمایشگاه مربوطه تایید شد[۳۰]. پس از اصلاح و آماده سازی داده‌ها که شامل جایگزینی مقادیر سنسورد و مفقود با الگوریتم $IrEM$[۹] برای داده‌های ترکیبی و حذف Bi از متغیرها به دلیل کیفیت نامناسب است، ماتریسی شامل ۶۶۱ نمونه و ۴۷ متغیر به دست آمد. محاسبات جایگزینی به کمک کدنویسی در محیط بسته *zCompositions* نرم‌افزار R انجام شد[۴۰].

سپس دندروگرام داده‌های ترکیبی با SBP پیش فرض و به کمک بسته *compositions* در R محاسبه و ترسیم شد که در شکل (۲) نشان داده شده است[۴۱]. بر اساس نتیجه حاصل از دندروگرام، تعداد ۴ خوشه قابل تفکیک است که با چهار رنگ نشان داده شده‌اند.

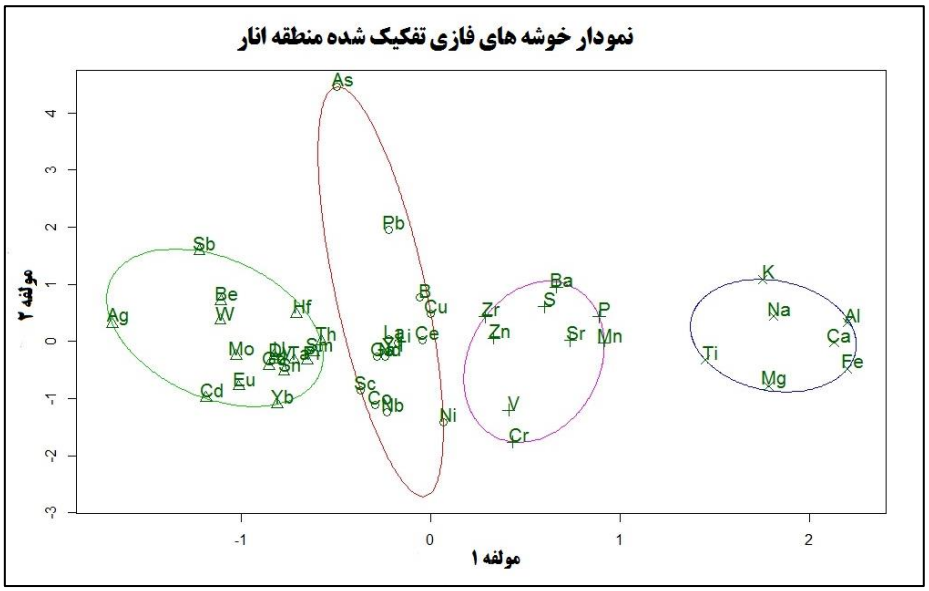
در مرحله بعد، الگوریتم *fanny* روی داده‌های بازشده با clr انجام شد که متغیرهای خوشه‌های فازی تفکیک شده توسط الگوریتم *fanny* در جدول ۲ و شکل ۳ نشان داده شده است. بر اساس نتایج حاصل که کاملا منطبق بر دندروگرام است، خوشه (۱) نشان دهنده مجموعه عناصر نادر خاکی و کانی‌سازی مس و طلا در منطقه است. خوشه (۲) اغلب شامل عناصر رادیو اکتیو و نادر خاکی است که می‌تواند بدلیل وجود اندیس‌های REE پلاسری مروست در غرب منطقه مورد مطالعه و احتمال پراکندگی در آبراهه‌های منطقه باشد. خوشه (۳) و خوشه (۴) عناصر سنگ ساز بوده و عمدتاً ارتباط با کمپلکس افیولیتی را نشان می‌دهد. ضمن اینکه نحوه نمونه‌گیری نیز می‌تواند روی نتایج خوشه‌بندی تأثیر بگذارد.



شکل ۲: دندروگرام داده‌های ترکیبی با روش جدایش سلسله مراتبی

جدول ۲:خوشه‌های فازی تفکیک شده توسط الگوریتم *fanny*

شماره خوشه	عناصر تفکیک شده در هر خوشه
۱	*La, Y, Nd, Li, Nb, Ga, Co, Pb, Ce, B, Sc, Cu, Ni, As*
۲	*Eu, Mo, Be, W, Cd, Gd, Sb, U, Dy, Yb, Sn, Ta, Ag, Hf, Pr, Sm, Th*
۳	*Sr, Ba, S, P, Mn, V, Cr, Zn, Zr*
۴	*Ca, Al, Fe, Na, Mg, K, Ti*

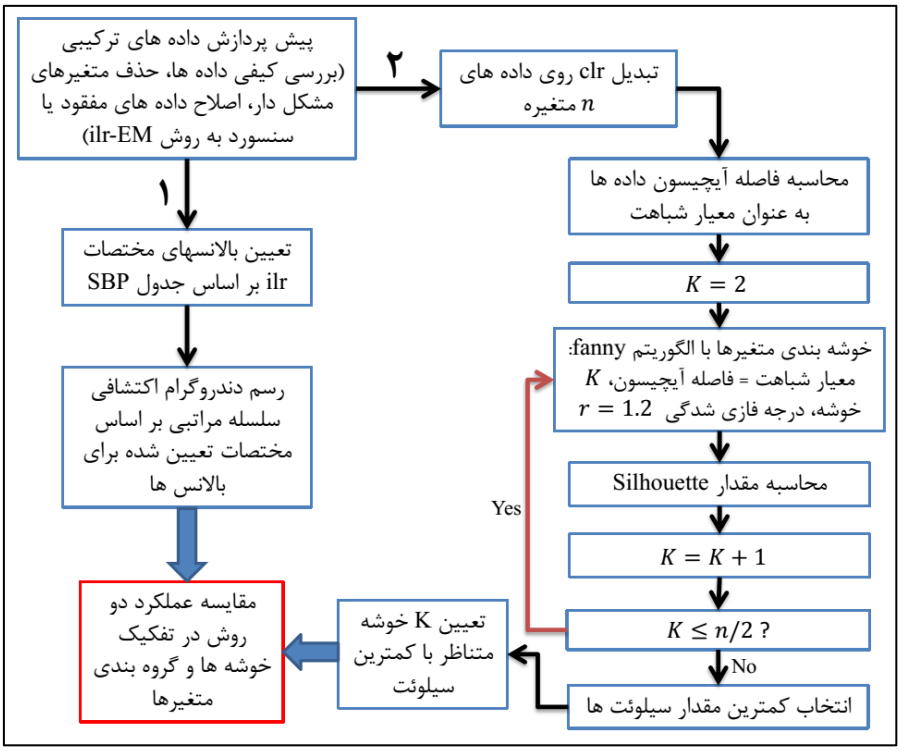


شکل ۳: متغیرهای خوشه‌بندی شده توسط الگوریتم *fanny*

۷- نتیجه گیری

روش‌های خوشه‌بندی متفاوتی با توجه به نوع داده مورد بررسی توسعه یافته‌اند. استفاده از این روش‌ها روی متغیرهای مجموعه داده‌ها منجر به کاهش بُعد و تفکیک آنها به گروه‌های همسان می‌شود. در اکتشافات ژئوشیمیایی رسوبات آبراهه‌ای روش خوشه‌بندی فازی از جمله مهمترین روش‌های طبقه‌بندی نظارت نشده است که در صورت اجرای صحیح و انتخاب دقیق پارامترهای مربوطه از جمله

تعداد خوشه، نتایج قابل قبولی را ارائه می‌دهد. انطباق زمین‌شناسی، مهمترین معیار در انتخاب تعداد خوشه مناسب و در پی آن صحت خوشه‌بندی است. در پژوهش حاضر، خوشه‌بندی داده‌های رسوبات آبراهه‌ای برگه انار با نگرشی ویژه به ماهیت ترکیبی بودن داده‌های ژئوشیمیایی انجام گرفت. روند انجام کار به صورت خلاصه در شکل ۴ نشان داده شده است. بدین منظور در مرحله اول، با محاسبه دندروگرام اکتشافی داده‌های ترکیبی، با توجه به معیار وارد و روش کمترین مربعات، تعداد (۴) خوشه، مناسب تشخیص داده شد که با واحدهای لیتولوژی و ساختار زمین‌شناسی منطقه انطباق بسیار خوبی نشان می‌دهد. در مرحله دوم، خوشه‌بندی فازی با به کارگیری الگوریتم *fanny* بر روی داده‌های تبدیل یافته با تبدیل *clr* انجام شد که نتایج یکسانی با روش دندروگرام ترکیبی حاصل شد. انطباق بسیار خوب الگوریتم مذکور و دندروگرام داده‌های ترکیبی نشان دهنده جدایش خوب در فضای سیمپلکس است.



شکل ۴: فلوچارت روند خوشه‌بندی توسط دو روش و مقایسه آنها

هدف کلی از این پژوهش، مطالعه همزمان دو روش معمول و کلاسیک خوشه‌بندی فازی و دندروگرام اکتشافی، نه با داده‌های خام یا نرمال لگاریتمی بلکه با روابط موجود بین داده‌های ترکیبی و در نهایت مقایسه نتایج بود که نشان داد در صورت کاربرد تبدیل مناسب و تعریف دقیق *SBP* یا مختصات جدید در فضای سیمپلکس می‌توان خوشه‌بندی قابل قبولی روی داده‌های بسته ژئوشیمی انجام داد.

مراجع

[1] Hassani Pak, A. A., Sharafuddin, M. (1384). Exploration data Analysis. University of Tehran Press. in persian

[2] Carranza, E. J. M. (2008). Geochemical anomaly and mineral prospectivity mapping in GIS (Vol. 11): Elsevier.

[3] Reimann, C., Filzmoser, P., & Garrett, R. G. (2002). Factor analysis applied to regional geochemical data: problems and possibilities. Applied geochemistry, 17(3), 185-206.

[4] Bezdek J.C., Ehrlich R.R., Full W., (1984). FCM: the fuzzy c-means clustering algorithm. Comput. Geosci., 10:191-203.

[5] Gordon, A. D.; (1999); “Classification”, 2nd Edition, Chapman and Hall, Boca Raton.

[6] Bochang, Y., & Xuejing, X. (1985). Fuzzy cluster analysis in geochemical exploration. Journal of Geochemical Exploration, 23(3), 281-291.

[7] Grekousis, G., & Thomas, H. (2012). Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The Fuzzy C-Means and Gustafson–Kessel methods. Applied Geography, 34, 125-136.

[8] De Carvalho, F. D. A., & Tenório, C. P. (2010). Fuzzy K-means clustering algorithms for interval-valued data based on adaptive quadratic distances. Fuzzy Sets and Systems, 161(23), 2978-2999.

[9] Ziaii, M., Pouyan, A. A., & Ziaei, M. (2009). Neuro-fuzzy modelling in mining geochemistry: Identification of geochemical anomalies. Journal of Geochemical Exploration, 100(1), 25-36.

[10] Templ, M., Filzmoser, P., & Reimann, C. (2008). Cluster analysis applied to regional geochemical data: problems and possibilities. Applied Geochemistry, 23(8), 2198-2213.

[11] Reimann, C., Filzmoser, P., & Garrett, R. G. (2005). Background and threshold: critical comparison of methods of determination. Science of the Total Environment, 346(1), 1-16.

[12] Chork, C., & Govett, G. (1985). Comparison of interpretations of geochemical soil data by some multivariate statistical methods, Key Anacon, NB, Canada. Journal of Geochemical Exploration, 23(3), 213-242.

[13] Basilevsky, A. (1994). Statistical factor analysis and related methods: theory and applications. Wiley series in probability and mathematical statistics, 737.

[14] Chork, C., & Salminen, R. (1993). Interpreting exploration geochemical data from Outokumpu, Finland: a MVE-robust factor analysis. Journal of Geochemical Exploration, 48(1), 1-20.

[15] Treiblmaier, H., & Filzmoser, P. (2010). Exploratory factor analysis revisited: How robust methods support the detection of hidden multivariate data structures in IS research. Information & management, 47(4), 197-207.

[16] Carranza, E. J. M. (2011). Analysis and mapping of geochemical anomalies using logratio-transformed stream sediment data with censored values. Journal of Geochemical Exploration, 110(2), 167-185.

[17] Filzmoser, P., Hron, K., & Reimann, C. (2009). Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Science of the Total Environment, 407(23), 6100-6108.

[18] Wenlei.W., Zhao, J., & Cheng, Q. (2013). Fault trace-oriented singularity mapping technique to characterize anisotropic geochemical signatures in Gejiu mineral district, China. Journal of Geochemical Exploration, 134, 27-37.

[19] Aitchison, J. (1981). A new approach to null correlations of proportions. Journal of the International Association for Mathematical Geology, 13(2), 175-189.

[20] Aitchison, J. (1983). Principal component analysis of compositional data. Biometrika, 70(1), 57-65.

[21] Aitchison, J. (1984). The statistical analysis of geochemical compositions. Journal of the International Association for Mathematical Geology, 16(6), 531-564.

[22] Aitchison, J. (1986). The statistical analysis of compositional data (Vol. 25): Chapman & Hall.

[23] Aitchison, J. (1999). Logratios and natural laws in compositional data analysis. Mathematical Geology, 31(5), 563-580.

[24] Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J., & Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. Mathematical Geology, 32(3), 271-275.

[25] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. Mathematical Geology, 35(3), 279-300.

[26] Egozcue, J. J., & Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. Mathematical Geology, 37(7), 795-828.

[27] Buccianti, A., & Pawlowsky-Glahn, V. (2005). New perspectives on water chemistry and compositional data analysis. Mathematical Geology, 37(7), 703-727.

[28] Buccianti, A., Mateu-Figueras, G., & Pawlowsky-Glahn, V. (2006). Compositional data analysis in the geosciences: from theory to practice.

[29] Thió-Henestrosa, S., & Martín-Fernández, J. (2005). Dealing with compositional data: the freeware CoDaPack. Mathematical Geology, 37(7), 773-793.

[30] Geological survey of Iran,. (2016). Final report of BLEG geochemical exploration in Anar and Yazd 1:250,000 geological sheets. Tehran. in persian

[31] Anar 1:250,000 geological quadrangle map (1981). Geological Survey of Iran. in persian

[32] Aghanabati, A. (2004). Geology of Iran. Geological Survey of Iran. in persian

[33] Van den Boogaart, K. G., & Tolosana-Delgado, R. (2013). Analyzing compositional data with R. Berlin: Springer.

[34] Egozcue, J. J., & Pawlowsky-Glahn, V. (2011). Basic concepts and procedures. Compositional Data Analysis: Theory and Applications, 12-28.

[35] Pawlowsky-Glahn, V., & Egozcue, J. J. (2006). Compositional data and their analysis: an introduction. Geological Society, London, Special Publications, 264(1), 1-10.

[36] Hassani Pak, A. A. (1384). Principles of Geochemical Explorations. University of Tehran Press. in persian

[37] Kaufman, L. and Rousseeuw, P.J. (1990) Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

[38] Palarea-Albaladejo, J., Martín-Fernández, J. A., & Soto, J. A. (2012). Dealing with distances and transformations for fuzzy C-means clustering of compositional data. Journal of classification, 29(2), 144-169.

[39] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

[40] Palarea-Albaladejo, J., & Martin-Fernandez, J. A. (2014). zCompositions-package.

[41] Van den Boogaart, K. G., Tolosana, R., & Bren, M. (2011). compositions: Compositional Data Analysis. R package version 1.10-2. URL http://CRAN. R-project. org/package5compositions.

1- Aitchison Geometry
2- Centered Log-ratio transform
3- Aitchison Distance
4- Balances
5- Sequential Binary Partition
6- Ward
7- Dissimilarity
8- Silhouette
9- Logratio Expectation Maximization