معرفی آنالیز کلاستری ارتباط کامل به عنوان بهترین روش تجزیه و تحلیل کلاستری نوع R

بهزاد تخم‌چی [۱*]، زهره جنگروی [۲]

۱- عضو هیأت علمی دانشگاه صنعتی شاهرود

۲- دانشجوی کارشناسی اکتشاف معدن دانشگاه صنعتی شاهرود

E-mail: tokhmechi@ut.ac.ir

چکیده

یکی از روشهای جدا سازی آنومالی در اکتشافات ژئوشیمیایی استفاده از هاله‌های مرکب است. هاله‌های مرکب در مقایسه با هاله‌های تک عنصری به مراتب بزرگتر هستند، همچنین تعداد متغیرهای مورد بررسی را کاهش داده و هرچند عنصر یا اکسید را در یک گروه قرار می‌دهند که امکان تفسیر آسانتر نتایج مطالعات ژئوشیمیایی و معرفی دقیقتر آنومالیها را فراهم می‌آورد. به منظور طبقه بندی عناصر مورد بررسی در گروهها یا زیرگروهها از تجزیه و تحلیل خوشه‌ای یا کلاستر استفاده می‌شود. بدین ترتیب که ضریب همبستگی بین عناصر مختلف محاسبه شده و بر اساس نزدیکی بین عناصر گروه بندی صورت می‌گیرد. روشهای مختلفی برای محاسبه ضریب همبستگی و گروه‌بندی ارائه شده است. بدیهی است روشی بهینه است که نزدیکترین گروه‌بندی به پاراژنز عناصر را نتیجه داده و در ضمن نمونه‌هایی را به عنوان آنومالی یک عنصر معرفی کند که در صورت عدم گروه‌بندی نیز آنومالی آن عنصر می‌بود. در این مقاله روشهای مختلف محاسبه ضریب همبستگی و گروه‌بندی مرور شده و با توجه به نزدیکی نتایج به نتایج حاصل از تجزیه و تحلیل تک متغیره به روشهای $\overline{X}+NS$ و $P.N$، روش ارتباط کامل (دورترین همسایه) به عنوان دقیقترین روش تجزیه و تحلیل کلاستری نوع R معرفی شده است.

۱- مقدمه

در بیشتر موارد تجزیه و تحلیل داده‌ها در ژئوشیمی اکتشافی تجربی است. دلیل این موضوع نیز نبود مدلهای ریاضی لازم جهت توصیف نوع منبع، چگونگی تفریق، مهاجرت، ته‌نشست و تمرکز عناصر کمیاب در سنگهاست [۱]. علی‌رغم این موضوع، از آنجا که نتایج آنالیز نمونه‌ها به صورت کمی است، امکان تجزیه و تحلیل آماری داده‌ها و جود خواهد داشت.

* تهران صندوق پستی، ۱۳۱۴۵-۱۳۳۷

به هر حال در انجام اکتشافات ژئوشیمیایی، کشف الگوهای ژئوشیمیایی غیر عادی یا آنومالیهای ژئوشیمیایی مد نظر است (هاکس و وب)، که برای این منظور می‌بایست نسبت به شناخت مقدار زمینه عیار عناصر در محیط اقدام کرد [۲]. بدیهی است هر بیش بودی نسبت به مقدار زمینه نشان دهنده یک آنومالی نیست. لذا برای تفکیک آنومالی روشهای آماری مختلفی وجود دارد که به عنوان مثال می‌توان از روشهای $\overline{x} + ns$، حاصلضرب $P.N$، آماره انفصال و آنالیز فاکتوری نام برد. در ضمن طبق نظر بئوس (Beus) و گریگوریان (Grigorian،۱۹۷۷) اگر ترکیبی از مقادیر یک گروه از عناصر معرف به جای مقدار یک عنصر خاص به کار گرفته شوند، هاله‌های ژئوشیمیایی در اطراف توده های کانسار بهتر مشخص می‌شوند[۳]. روشهای مختلفی برای گروه‌بندی وجود دارند که تحت عنوان تجزیه و تحلیل خوشه‌ای یا کلاستر معرفی می‌شوند. به طور خاص از روشهای کلاستری می‌توان به ارتباط بین گروهی (between - groups linkage)، ارتباط درون گروهی(within - groups linkage)، نزدیکترین همسایه (nearest neighbor)، دورترین همسایه (neighbor furthest، ارتباط کامل) خوشه‌بندی متمرکز و خوشه‌بندی میانه اشاره کرد.

بدیهی است معرفی یکی از روشهای خاص کلاستر به عنوان روشی که احتمالاً بیشترین توانمندی را در تفکیک آنومالی‌ها دارد، به ژئوشیمیست‌ها کمک خواهد کرد تا با دقت و سرعت بیشتری نسبت به پردازش آماری داده‌های حاصل از آنالیزها اقدام کرده و آنومالی‌ها را گویاتر معرفی کنند. در این مقاله مقایسه‌ای بین روش‌های مختلف خوشه‌بندی صورت گرفته و از نزدیکی آنومالی‌های گروه‌بندی شده به آنومالی‌های تک عنصره که به روشهای آماری $\overline{X} + NS$ و حاصلضرب $P.N$ مشخص شده‌اند استفاده شده و بهترین روش کلاستری معرفی شده است.

۲- معرفی مختصر روشهای خوشه بندی سلسله مراتبی

گروه‌بندی متغیرها با هدف تقلیل ابعاد ماتریس داده‌ها انجام می‌شود. روشهای گروه‌بندی در مواقعی که تعداد گروه‌ها از پیش مشخص‌اند، با حالاتی که از پیش تعداد گروهها مشخص نیستند فرق می‌کنند[۴]. بدیهی است در تحلیل‌های ژئوشیمیایی تعداد گروهها از پیش مشخص نیستند. لذا بهترین روش گروه‌بندی در مورد اطلاعات ژئوشیمیایی، خوشه‌بندی سلسله مراتبی است. در زیر به اختصار روشهای خوشه بندی سلسله مراتبی معرفی می‌شوند. برای این منظور در جدول ۱ اطلاعات فرضی بین چهار نمونه که برای آنها دو عنصر x و y آنالیز شده‌اند آورده شده است. در ادامه نحوه عمل روشهای مختلف خوشه‌بندی بر اساس این مجموعه اطلاعات ساده شرح داده می‌شود[۵].

جدول ۱- مقادیر حاصل از آنالیز دو عنصر برای چهار نمونه

عناصر / نمونه	x	y
۱	+۵	+۳
۲	-۱	+۱
۳	+۱	-۲
۴	-۳	-۲

در *روش ارتباط بین گروهی*، ترکیب خوشه‌ها از کمینه کردن متوسط فاصله بین تمام زوج مشاهداتی که در خوشه‌های مختلف قرار دارند، ایجاد می‌شود. در این روش از کلیه فواصل موجود بین نمونه‌های خوشه‌ها و نه فقط نزدیکترین یا دورترین فاصله‌ها استفاده می‌شود. در اینجا معیار مربع فاصله اقلیدسی بین نمونه‌های موجود در جدول۱، برای مقایسه روشهای خوشه‌بندی محاسبه می‌شوند. ماتریس فاصله، ماتریسی متقارن است. علامت *روی قطر نشان دهنده فاصله هر نقطه با خود است که صفر بوده و در ماتریس گنجانده نمی‌شود.

$$\begin{pmatrix} 1 & * & & \\ 2 & 40 & * & \\ 3 & 41 & 13 & * \\ 4 & 89 & 13 & 16 \end{pmatrix}$$

مثلاً برای دو نمونه ۱و۲ به صورت زیر عمل می‌شود:

$$D_{12}^2 = (5-(-1)^2 + (3-1)^2 = 36+4 = 40$$

$$D_{24}^2 = 13$$

$$D_{24,3}^2 = \frac{D_{23}^2 + D_{34}^2}{2} = \frac{13+16}{2} = 14.5$$

$$D_{24,1}^2 = \frac{D_{21}^2 + D_{14}^2}{2} = \frac{40+89}{2} = 64.5$$

چون ۱۴/۵ کوچکتر از۶۴/۵ است، نمونه ۳ به خوشه ۲۴ جذب می‌شود و خوشه ۲۴۳ تشکیل می‌شود.

$$D^2_{243,1} = \frac{D^2_{21} + D^2_{41} + D^2_{31}}{3} = \frac{40+41+89}{3} = 56.666$$

در *روش ارتباط درون گروهی*، سعی می‌شود مشاهدات در خوشه‌ها چنان تعیین شوند که معدل فاصله نمونه‌های داخل خوشه‌ها از یکدیگر به کمترین مقدار برسد:

$$D^2_{24} = 13 \qquad D^2_{24,3} = \frac{13+13+16}{3} = 14$$

$$D^2_{243,1} = \frac{D^2_{21} + D^2_{31} + D^2_{41} + D^2_{23} + D^2_{24} + D^2_{34}}{6} = 35.333$$

در *روش نزدیکترین همسایه*، کمترین فاصله هر مشاهده نسبت به مشاهدات موجود به دست می‌آید و سپس این عمل به ازای کلیه مشاهدات باقی مانده و نسبت به کلیه عناصر خوشه‌های موجود تکرار می‌شود.

$$D^2_{24} = 13$$

خوشه دوم

$$D^2_{23} = 13, D^2_{21} = 40, D^2_{31} = 41 \Rightarrow D^2_{23} \langle D^2_{21} \langle D^2_{31} \Rightarrow 2,3$$

$$D^2_{23,1} = D^2_{21} = 40,\ D^2_{31} = 41, D^2_{41} = 89 \Rightarrow Min = 40$$

بنابراین اگر فاصله حداکثر ۴۰ فرض شود، هر چهار نمونه در یک خوشه قرار می‌گیرند، اما اگر فاصله عددی مانند ۲۰ فرض شود، نمونه‌های ۲و۳و۴ در یک خوشه و نمونه ۱ در یک خوشه مجزا قرار می‌گیرد.

در *روش دورترین همسایه یا ارتباط کامل*، فاصله بین دو خوشه بر حسب فاصله بین دورترین نمونه‌ها محاسبه می‌شود.

$$D^2_{24} = 13 \quad D^2_{23} = 13 \quad D^2_{21} = 40$$
$$D^2_{34} = 16 \quad D^2_{13} = 41$$
$$D^2_{14} = 89 \rightarrow MAX\ \ D_{1j} \quad \forall j = 2,3,4$$

در *روش خوشه بندی متمرکز*، فاصله بین دو خوشه، فاصله بین میانگین‌های آنهاست. فاصله‌ای که خوشه‌ها با یکدیگر ترکیب می‌شوند از مرحله‌ای به مرحله دیگر کاهش می‌یابد.

$$D^2_{24} = 13 \Rightarrow \qquad \overline{X}_{2,4} = (-2,-0.5)$$
$$D^2_{24,3} = (-2-1)^2 + (-2+0.5)^2 = 11.25$$
$$\overline{X}_{24,3} = (\frac{-1+1-3}{3}, \frac{+1-2-2}{3}) = (-1,-1)$$

$$D^2_{243,1} = (-1-5)^2 + (-1-3)^2 = 36+16 = 52$$

در *روش خوشه بندی میانه*، به دو خوشه‌ای که ترکیب می‌شوند، وزنهای یکسانی صرف نظر از تعداد نمونه‌های آنها داده می‌شود. این عمل سبب می‌شود گروههای کوچک نسبت به سایر گروههای دیگر اثر مشابهی در ساختن خوشه‌های بزرگتر داشته باشند.

$$D^2_{24} = 13 \quad \overline{X}_{2,4} = (-2,-0.5)$$
$$D^2_{24,8} = 11.25$$
$$\overline{X}_{24,3} = (\frac{-2+(+1)}{2}, \frac{-0.5-2}{2}), (-0.5,-1.25)$$
$$D^2{}_{243,1} = (-0.5-5)^2 + (-1.25-3)^2 = 48.3125$$

۳- جوامع مطالعه شده

برای مطالعه توانمندی روشهای مختلف خوشه‌بندی از اطلاعات آنالیز XRF مربوط به دو منطقه استفاده شده است. یک مجموعه اطلاعات مربوط به طرح اکتشاف نیمه تفصیلی تنگستن در منطقه اشنویه آذربایجان‌غربی است که مشتمل بر آنالیز XRF ۳۰ عنصری ۵۰ نمونه است. در این مجموعه، اطلاعات عیار ۱۹ عنصر یا ترکیب اکسیدی در خصوص بعضی نمونه‌ها به صورت سنسورد بوده که مقادیر سنسورد به روش بیشترین درستنمایی تخمین زده شده‌اند [۶].

مجموعه دوم اطلاعاتی مربوط به طرح پتانسیل‌یابی مواد معدنی منطقه گزل‌بلاغ (شاهین دژ) واقع در آذربایجان‌غربی است که مشتمل بر آنالیز XRF ۳۵ عنصری ۲۴ نمونه است . در این مورد نیز عیار بعضی عناصر در بعضی نمونه‌ها به صورت سنسورد بوده‌اند که مقادیر سنسورد به روش بیشترین درستنمایی تخمین زده شده‌اند [۷].

۴- مقایسه روشهای مختلف کلاستری

همانگونه که در معرفی روشهای مختلف خوشه‌بندی ملاحظه شد، اساس تقسیم‌بندی‌های ذکر شده بر مبنای ضریب همبستگی عناصر است که با استفاده از آمار چند متغیره صورت می‌گیرد. برای بررسی اینکه

کدام روش‌های خوشه‌بندی پاراژنز عناصر را بهتر مدل می‌کنند و خطای کمتری دارند، اقدام به ترسیم دندوگرام همبستگی عناصر به تفکیک برای مناطق اشنویه و شاهین دژ گردید.

در دندوگرام رسم شده به روش خوشه‌بندی متمرکز، حداکثر ۴ عنصر یا اکسید در یک جامعه قرار گرفتند. به این ترتیب تعداد متغیرهایی که در نهایت به دست می‌آید زیاد بوده و عملاً نظر بئوس و گریگوریان در خصوص مطالعه ترکیبهای بزرگی از عناصر به جای ترکیبهای تک عنصری ارضاء نمی‌شود. لذا روش خوشه‌بندی متمرکز روش مناسبی نخواهد بود. روشهای خوشه‌بندی نزدیکترین همسایه، خوشه‌بندی میانه و ارتباط بین گروهی در جداسازی جوامع و نشان دادن ارتباط عناصر به مراتب بهتر از روش خوشه‌بندی متمرکز عمل کردند، اما در نهایت دندوگرامهای رسم شده، پاراژنز بین عناصر را به خوبی مدل نمی‌کنند، به گونه‌ای که از نظر زمین‌شناسی اقتصادی امکان تفسیر معنی‌دار آنومالی‌های گروهی معرفی شده وجود ندارد.

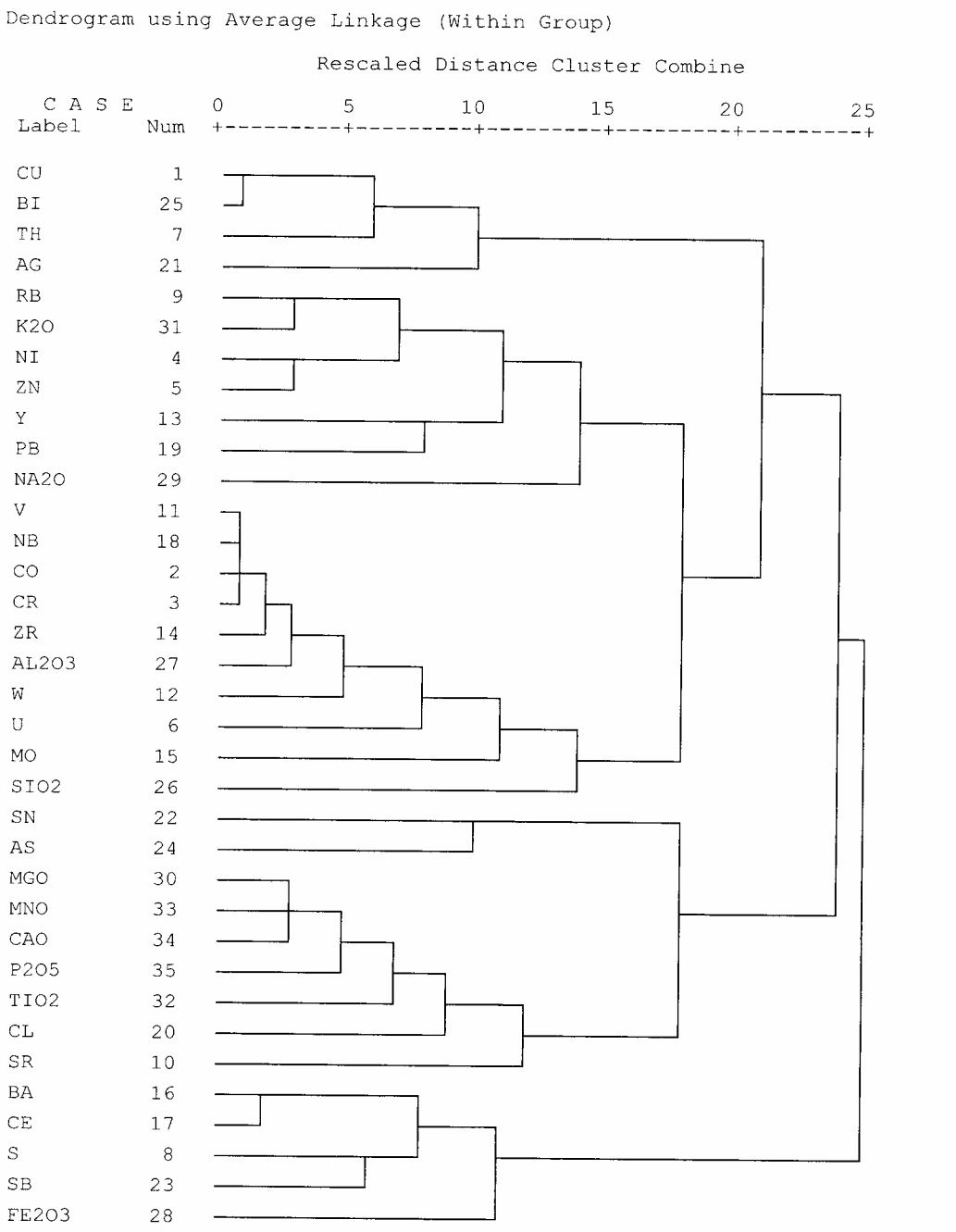
روشهای خوشه‌بندی ارتباط درون گروهی و ارتباط کامل (دورترین همسایه) در مقایسه با کلیه روشهای دیگر ارتباط عناصر را بهتر نشان داده و در ضمن گروه‌بندی آنها با پاراژنزهای عنصری نیز همخوانی معنی‌داری برقرار ساخت. به عنوان نمونه در شکل‌های ۱و۲ دندوگرامهای ترسیم شده بر روی مقادیر شاخص غنی‌شدگی داده‌های منطقه شاهین‌دژ به روشهای خوشه‌بندی ارتباط کامل (دورترین همسایه) و ارتباط درون گروهی آورده شده است. از آنجایی که این روشهای کلاستری، بهینه به نظر می‌رسند، بررسی‌های دقیقتر در خصوص آنها صورت می‌گیرد.

۵- مقایسه نتایج به دست آمده از روشهای ارتباط کامل و ارتباط درون گروهی

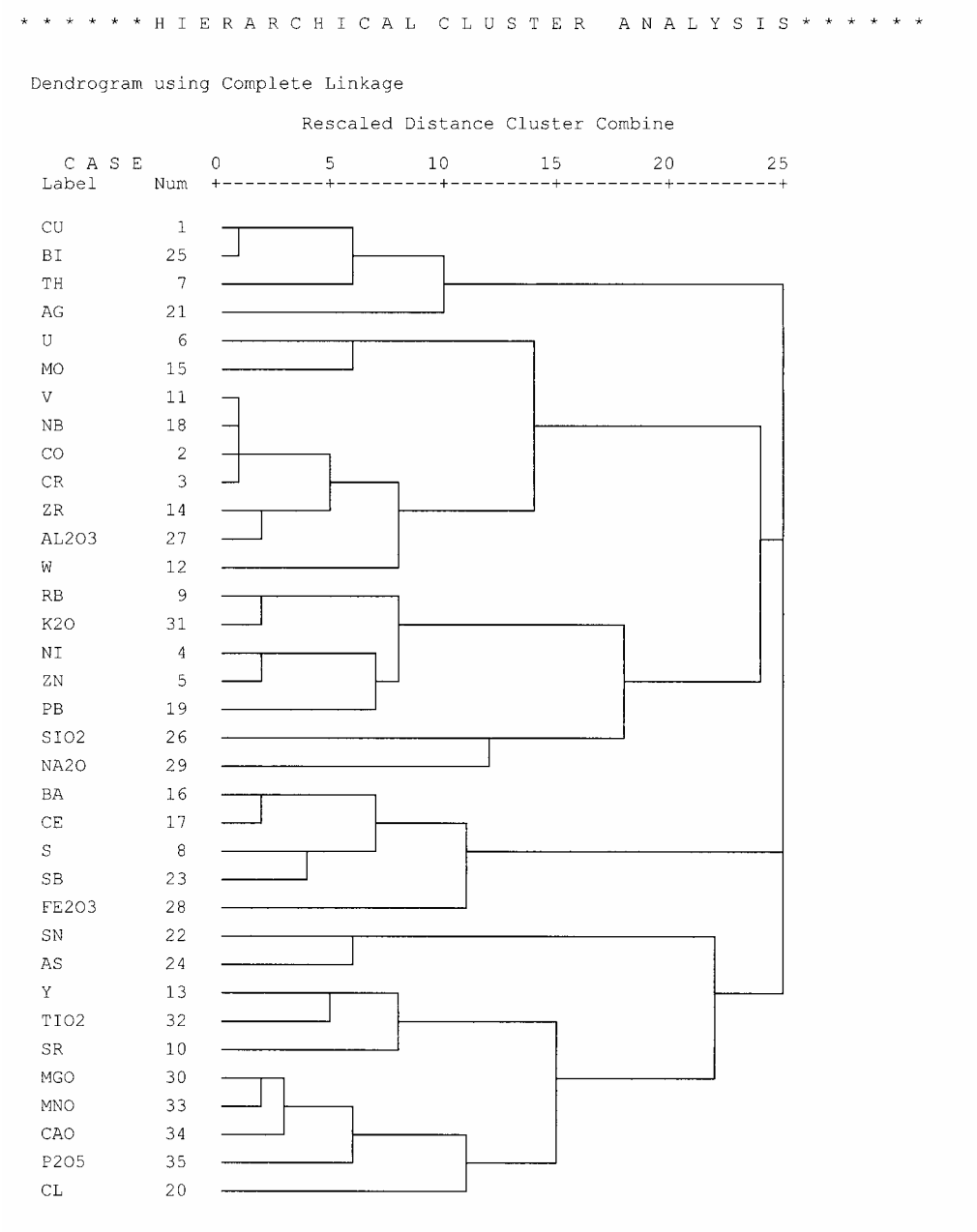
برای گزینش بهترین روش کلاستری در جداسازی گروهها شاخص غنی‌شدگی برای مجموعه داده‌های مناطق شاهین‌دژ و اشنویه محاسبه شده است. در وهله اول آنومالی عناصر و ترکیبات به روشهای آماری $\overline{X}+ns$ و $P.N$ برای تک تک عناصر و ترکیبات دو منطقه محاسبه شده و نتایج به عنوان معیار صحت گروه‌بندی لحاظ شده‌اند. در مرحله دوم آنالیز کلاستر ارتباط کامل و ارتباط درون گروهی برای داده‌های شاخص غنی‌شدگی هر دو منطقه محاسبه گردید. به عنوان مثال در جدول ۲ نتایج آنالیز کلاستر ارتباط درون گروهی بر روی داده‌های منطقه شاهین‌دژ و در جدول ۳ نتایج آنالیز کلاستر ارتباط کامل بر روی داده‌های منطقه اشنویه آورده شده است.

آنومالی متغیرهای گروهی حاصل از آنالیز خوشه‌بندی به روشهای $\overline{X}+ns$ و $P.N$ تعیین شده‌اند. میزان مطابقت آنومالی‌های معرفی شده عناصر و ترکیبات در حالت غیر گروه‌بندی شده مقایسه شده و به درصد بیان شده است. به عنوان مثال متغیر شماره ۱ (Varl) حاصل از آنالیز کلاستر به روش ارتباط درون گروهی برای داده‌های منطقه شاهین‌دژ شامل عناصر مس (Cu)، بیسموت (Bi)، نقره (Ag) وتانتالیم (Th)

است (جدول ۲). به هر دو روش $\overline{X}+ns$ و $P.N$ نمونه R_4 به عنوان آنومالی احتمالی این گروه معرفی شده است، در صورتیکه در مطالعه تک متغیره ۴ عنصر مذکور به روشهای آماری $\overline{X}+ns$ و $P.N$ درمورد عنصر مس نمونه R_4 آنومالی قطعی، در مورد نقره نمونه R_4 آنومالی احتمالی و در مورد بیسموت نمونه R_4 آنومالی ممکن بوده است. در مورد تانتالیم نمونه R_4 آنومالی نبوده و اصولاً تانتالیم آنومالی نداده است. عناصر مس، بیسموت و نقره نیز جز نمونه R_4 آنومالی دیگری نداده‌اند. در این حالت در نظر گرفته شده است که نتایج تعیین آنومالی گروه خوشه‌بندی شده به روش ارتباط درون گروهی با واقعیت ۳ از ۴ (۷۵ درصد) انطباق دارد. مشابه عملیات مذکور برای تمامی گروهها و عناصر صورت گرفته و نتایج نهایی برای منطقه شاهین‌دژ در جدول ۴ و برای منطقه اشنویه در جدول ۵ آورده شده است. منظور از میانگین درصد مطابقت آنومالی‌ها در جداول ۴ و ۵ ، در واقع میانگین حسابی مطابقت گروهها است. ذکر این نکته نیز ضروری است که تأثیر گروههای تک عنصره از محاسبات حذف شده است. زیرا بدیهی است که در این مواقع انطباق آنومالی گروه و عنصر صددرصد خواهد بود.



شکل ۱- خوشه‌بندی به روش ارتباط کامل (دورترین همسایه) بر روی شاخص غنی‌شدگی داده‌های منطقه شاهین‌دژ



شکل ۲- خوشه‌بندی به روش ارتباط درون گروهی بر روی شاخص غنی‌شدگی داده‌های منطقه شاهین دژ

جدول ۲- نتایج آنالیز کلاستر ارتباط درون گروهی برروی شاخص غنی شدگی داده‌های منطقه شاهین‌دژ

ردیف	متغیر	ترکیبات و عناصر
۱	Var 1	$Th + Ag + Bi + Cu$
۲	Var2	$Mo + U + W + Al_2O_3 + Zr + Cr + Co + Nb + V$
۳	Var3	$Y + Pb + Zn + Ni + k_2O + Rb$
۴	Var4	Na_2O
۵	Var5	SiO_2
۶	Var6	$As + Sn$
۷	Var7	$Fe_2O_3 + Sb + S + Ce + Ba$
۸	Var8	$Cl + Sr + TiO_2 + p_2O_5 + CaO + MnO + MgO$

جدول ۳- نتایج آنالیز کلاستر ارتباط کامل بر روی شاخص غنی‌شدگی داده‌های منطقه اشنویه

ردیف	متغیر	ترکیبات و عناصر
۱	Fac1	$Sc + Sr + Al_2O_3 + Cr + Zr + TiO_2$
۲	Fac2	$Zn + Ni + S + Fe_2O_3 + Co + MnO + MgO + Cu + V$
۳	Fac3	$CaO + Ba$
۴	Fac4	$Cl + SiO_2 + La + Nb + Y + Co + Rb + k_2O$
۵	Fac5	Na_2O
۶	Fac6	Ce
۷	Fac7	P_2O_5
۸	Fac8	F
۹	Fac9	W

جدول ۴- میانگین درصد مطابقت آنومالی‌های منطقه شاهین‌دژ

روش خوشه بندی	روش آماری	میانگین درصد مطابقت آنومالی ها
ارتباط کامل (دورترین همسایه)	$\overline{X}+NS$	۶۹/۱۸
	حاصلضرب $P.N$	۷۲/۱۱
ارتباط درون گروهی	$\overline{X}+NS$	۶۴/۷۶
	حاصلضرب $P.N$	۷۱/۷۷

جدول ۵ - میانگین درصد مطابقت آنومالی‌های منطقه اشنویه

روش خوشه بندی	روش آماری	میانگین درصد مطابقت آنومالی ها
ارتباط کامل (دورترین همسایه)	$\overline{X}+NS$	۶۸/۰۶
	حاصلضرب $P.N$	۶۸/۰۶
ارتباط درون گروهی	$\overline{X}+NS$	۶۳/۱۸
	حاصلضرب $P.N$	۶۳/۱۸

۶- نتیجه گیری و پیشنهادات

همانگونه که در جداول ۴ و ۵ ملاحظه می‌شود، درصد انطباق آنومالی‌ها در روش خوشه‌بندی ارتباط کامل با مطالعه آماری تک متغیره عناصر حدود ۷۰ درصد است. این در حالی است که درصد انطباق مذکور برای روش خوشه‌بندی ارتباط درون گروهی در حدود ۶۵ درصد است. بنابراین بدیهی است که از بین روشهای مختلف خوشه‌بندی، روش ارتباط کامل (دورترین همسایه) به عنوان گزینه برتر انتخاب می‌شود. البته همانگونه که ملاحظه شد، درصد انطباق روش‌های خوشه‌بندی، تنها در دو منطقه اکتشافی مطالعه شده است. لذا پیشنهاد می‌گردد برای رسیدن به یک نتیجه جامع، مطالعات مذکور در مناطق دیگر نیز صورت گیرد، تا در صورت تأیید نتایج مطالعات فعلی راهکاری مفید برای پردازش چند متغیره داده‌های ژئوشیمیایی ارائه گردد.

۷- تقدیر و تشکر

اطلاعات خام نتایج آنالیزها و مشخصات لیتولوژی‌های مورد نمونه‌برداری از سازمان صنایع و معادن استان آذربایجان غربی و شرکت مهندسین مشاور زمین کاوگستر گرفته شده که بدینوسیله از ایشان تشکر می‌گردد.

۸- منابع و مآخذ به ترتیب استفاده در متن

[۱] حسنی پاک، علی اصغر، (۱۳۷۰)، اصول اکتشافات ژئو شیمیایی (مواد معدنی)، انتشارات دانشگاه تهران، چاپ دوم.

[۲] Howkes, H.E ; and Webb , J.S. , (1962), 'Geochemistry in mineral exploration ''
New York and Evanston: Harper and Row publishing,

[۳] Beus, A.A.; and Grigorian, S.V. , (1962), 'Geochemical exploration methods for mineral deposits ', Trans. by R.T. Schneider, ed. by A.A. Levinson, Illiois: Applied publishing.

[۴] منتظری ، علیرضا، (۱۳۷۷)، کتاب آموزش SPSS در محیط ویندوز، کانون نشر علوم، خرداد.

[۵] شرکت آمار پردازن، (۱۳۷۷)، راهنمای کاربران SPSS6.0 تحت ویندوز ، مرکز فرهنگی انتشارات حامی.

[۶] شرکت زمین کاو گستر،(۱۳۷۸)، طرح اکتشافات نیمه تفصیلی تنگستن (فاز یک) در منطقه دوربه شمال شهرستان اشنویه.

[۷] شرکت زمین کاو گستر، (۱۳۸۰)، طرح پتانسیل یابی موادمعدنی منطقه گوزل بلاغ شاهین‌دژ.