کاربرد برآوردگرهای مقاوم در تعیین داده‌های خارج از ردیف؛ مطالعه موردی: داده‌های ژئوشیمیایی منطقه شاه سلیمان علی در استان خراسان جنوبی

حمید گرانیان*[1]، زهرا خواجه میری[2]

۱- استادیار، گروه معدن، دانشگاه صنعتی بیرجند

۲- دانشجوی دکتری، گروه زمین شناسی، دانشگاه شهید باهنر کرمان

(دریافت: اردیبهشت ۱۳۹۵، پذیرش: آذر ۱۳۹۶)

چکیده

شناسایی و تعدیل نمونه‌های خارج از ردیف چند متغیره اولین مرحله برای تحلیل آماری داده‌های اکتشافی محسوب می‌شود. کاهش بُعد داده‌ها به یک بُعد توسط فاصله‌ی نمونه از مرکز داده‌ها و مقایسه آن با یک حد آستانه کلید این کار محسوب می‌شود. در برآوردگرهای مقاوم از ماتریس‌های موقعیت و پراکندگی به جای ماتریس‌های میانگین و واریانس- کواریانس برای محاسبه این فاصله استفاده می‌شود. بنابراین برای مقاوم بودن این فاصله زیر مجموعه‌ی بهینه به جای کل داده‌ها برای محاسبه‌ی این ماتریس‌ها به کار می‌رود. چهار برآوردگر مقاوم *MVE*، *MCD*، *S* و *SD* در این مقاله معرفی گردیده‌اند. سپس از این برآوردگرها برای تعیین نمونه‌های خارج از ردیف ۱۴۶ نمونه‌ی رسوبات آبراهه‌ای منطقه شاه سلیمان علی در استان خراسان جنوبی و برای نتایج آنالیز ۱۸ عنصر استفاده شده است. نتایج محاسبات نشان داده است که روش کلاسیک فاصله ماهالانوبیتس ۷ نمونه و برآوردگرهای مقاوم *MVE*، *MCD*، *S* و *SD* به ترتیب ۲۳، ۳۵، ۲۰ و ۳۴ نمونه را به عنوان داده‌ی پرت معرفی می‌کنند. همچنین آنالیز مولفه‌های اصلی در مد *Q* نشان داده است که نمونه‌های خارج از ردیف با بارهای منفی خود را در مولفه‌ی دوم و سایر نمونه‌ها تقریباً با بارهای مثبت بالا در مولفه‌ی اول خود را نشان می‌دهند. تفکیک جامعه‌ی نمونه‌های خارج از ردیف از سایر نمونه‌ها نیز در نمودار پراکندگی بارهای مولفه‌ی دوم نسبت به مولفه‌ی سوم امکانپذیر است. استفاده از ماتریس‌های موقعیت و پراکندگی به دست آمده از برآوردگرهای مقاوم در روش‌های آمارهای چند متغیره یکی دیگر از کاربردهای پیشنهادی مهم برآوردگرهای مقاوم در تجزیه و تحلیل داده‌های اکتشافی محسوب می‌شوند.

کلید واژه‌ها

برآوردگر مقاوم، داده خارج از ردیف، آمار چند متغیره، داده ژئوشیمیایی، منطقه شاه سلیمان علی

* عهده دار مکاتبات: *h.geranian@birjandut.ac.ir*

۱- مقدمه

داده خارج از ردیف (داده پرت) به داده‌ای اطلاق می‌شود که به طور آشکار از سایر داده‌ها فاصله‌ی معنی‌داری داشته باشد. بنابراین فاصله زیاد نشان دهنده‌ی تشکیل داده‌ی خارج از ردیف توسط مکانیسم متفاوت نسبت به سایر داده‌ها خواهد بود[۱]. خطاهای انسانی در اندازه‌گیری، آماده‌سازی، آنالیز و ثبت یکی از مهم‌ترین این مکانیسم می‌تواند باشد. انتخاب نمونه اریب و تعلق نمونه به جوامع آماری متفاوت از دیگر دلایل ایجاد داده خارج از ردیف در داده‌های اکتشافی محسوب می‌شود. هرگونه پردازش و تفسیر داده‌ها بدون در نظر گرفتن این مسئله می‌تواند باعث ایجاد خطا و اشتباه در نتایج گردد. بنابراین شناسایی داده‌های خارج از ردیف اولین گام در پردازش داده‌ها محسوب می‌شود.

حذف، تعدیل و استفاده از برآوردگرهای مقاوم سه راه‌حل مقابله با داده‌های خارج از ردیف است. شناسایی داده‌های خارج از ردیف گام اول در استفاده از دو روش حذف و تعدیل است. به طور کلی روش‌های شناسایی داده‌های خارج از ردیف را می‌توان به دو گروه ۱- روش‌های بصری بر اساس نمودار (مثل روش نمودار احتمال، نمودار جعبه‌ای و نمودار خی‌دو) و ۲- روش‌های محاسباتی بر اساس تعیین آستانه (مثل روش دورفل، روش فانوپ و روش فرکتالی) تقسیم‌بندی نمود[۲-۴]. شناسایی داده‌های خارج از ردیف تک متغیره می‌تواند توسط این دو روش و یا تلفیق آنها صورت گیرد. درحالی‌که با افزایش بُعد داده‌ها (داده‌های چند متغیره) شناسایی داده‌های خارج از ردیف به دلیل مخفی شدن آنها در میان داده‌های چند متغیره، مشکل‌تر و با خطای بیشتر همراه است. برای حل این مشکل دو راه‌حل استفاده از روش‌های تفکیک جوامع آماری توسط روش‌های کلاسه‌بندی و خوشه‌بندی و روش‌های تبدیل داده‌های چند متغیره به یک بعد پیشنهاد شده است[۵، ۶]. در کلیه این روش‌های، از ماتریس‌های میانگین و واریانس-کواریانس استفاده می‌شود. وجود داده‌های خارج از ردیف باعث ایجاد اریب در تخمین ماتریس میانگین و تورم ماتریس واریانس-کواریانس خواهد شد، که نتیجه آن ایجاد اثر پوششی (Masking Effect) و اثر درون‌آوری (Swamping Effect) در تحلیل داده‌ها به روش‌های آمار چندمتغیره خواهد بود[۷]. بنابراین روش استفاده از برآوردگرهای مقاوم، که بدون نیاز به شناسایی داده‌های خارج از ردیف می‌توان از روش‌های آمار چند متغیره بهره برد، بهترین انتخاب برای این منظور خواهد بود.

برآوردگرهای مقاوم به دو گروه تقسیم می‌شوند. گروه اول برآوردگرهای هستند که به دنبال کمینه کردن مقیاس مقاوم فاصله ماهالانوبیتس هستند. برآوردگر بیضی‌وار با حجم کمینه (MVE)، برآوردگر ماتریس کواریانس با کمترین دترمینال (MCD)، برآوردگرهای S و برآوردگرهای τ از این دسته هستند. گروه دوم برآوردگرهای هستند که با نگاشت داده‌ها کار می‌کنند. برآوردگر استاهل- دانهو (SD) و برآوردگرهای P متعلق به این گروه هستند[۷، ۸]. در این مقاله ضمن معرفی مهم‌ترین و پرکاربردترین برآوردگر مقاوم، کاربرد آنها را بر روی داده‌های ژئوشیمیایی رسوبات آبراهه‌ای در مقیاس ۱/۲۵۰۰۰ در منطقه اکتشافی شاه سلیمان علی بررسی خواهد شد.

۲- روش کلاسیک تعیین داده خارج از ردیف چندمتغیره

اگر در یک دسته داده n تعداد نمونه‌ها و p تعداد متغیرها اندازه‌گیری شده باشد، ماتریس داده‌ها به صورت رابطه (۱) خواهد بود:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \qquad (۱)$$

به منظور شناسایی داده‌های خارج از ردیف چندمتغیره، بایستی بعد داده‌ها را به یک کاهش داده تا توسط نمودار ساده، این کار به آسانی قابل انجام باشد. محاسبه فاصله هر داده از مرکز ابر داده‌ها (فاصله ماهالانوبیتس) یکی از مهم‌ترین روش‌های کاهش بعد داده‌های چندمتغیره محسوب می‌شود، که این فاصله از رابطه‌های (۲) و (۳) به دست می‌آید:

$$MD(x_i) = \sqrt{(x_i - \mu)' \Sigma^{-1} (x_i - \mu)} \qquad (۲)$$
$$i = 1,2,\dots, n$$

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip})' \qquad (۳)$$

که μ ماتریس میانگین و $\sum$ ماتریس واریانس-کواریانس داده‌ها است. اگر فاصله ماهالانوبیتس نمونه‌ای از مقدار $\sqrt{\chi^2_{p,0.975}}$ بیشتر باشد، نمونه خارج از ردیف (داده پرت) خواهد بود. مقدار زیر رادیکال برابر چندک α - ۱ توزیع خی‌دو با درجه آزادی p است. به منظور جلوگیری از ایجاد شرایط تکینه‌گی بهتر است $n\geq 2p$ باشد. اگر $n>5p$ باشد، اثر بعد داده‌ها در محاسبه ماتریس کواریانس به شدت کاهش می‌یابد [۵، ۷]. اگر کلیه نمونه‌هایی را که فاصله آنها از مرکز ابر داده‌های برابر مقدار ثابت c^2 است ($c^2=\sqrt{\chi^2_{p,1-\alpha}}$)، به صورت مجموعه زیر تعریف کنیم، مکان هندسی داده‌ها به صورت یک بیضی‌وار به مرکز μ خواهد بود که نمونه‌های خارج از ردیف در خارج محدوده‌ای این بیضی‌وار قرار می‌گیرند.

$$\{x_i : (x_i-\mu)'\Sigma^{-1}(x_i-\mu)\} = c^2 \qquad i=1,2,...,n \qquad (۴)$$

فاصله ماهالانوبیتس معمولاً تحت اثر پوششی داده‌های خارج از ردیف قرار دارد، به طوری که نمونه‌های خارج از ردیف فاصله‌ی کمتری از مقدار واقعی به خود می‌گیرد. در نتیجه در محدوده بیضی‌وار قرار گرفته و شناسایی آنها در روش کلاسیک با مشکل همراه می‌شود. از طرف دیگر به علت جابجایی مرکز ابر داده‌ها، فاصله ماهالانوبیتس بعضی از نمونه‌های غیر خارج از ردیف، افزایش یافته و خارج از محدوده بیضی‌وار قرار می‌گیرند. در نتیجه اشتباهاً به عنوان نمونه‌های خارج از ردیف شناسایی خواهند شد (به این پدیده اثر درون‌آوری می‌گویند). لذا به منظور دستیابی به معیار معتبرتری برای شناسایی داده‌های خارج از ردیف، بایستی از برآوردگرهای مقاوم برای تخمین ماتریس‌های میانگین و واریانس- کواریانس استفاده نمود. در ادامه ضمن معرفی این برآوردگرها، اصول کاری هر یک مورد بررسی قرار می‌گیرد.

۳- روش‌های مقاوم تعیین داده خارج از ردیف چندمتغیره

اکثر روش‌های آماری بر فرض نرمال بودن توزیع داده‌ها بنا نهاده شده‌اند، در صورتی که در بسیاری از پروژه‌های اکتشافی با داده‌های غیر نرمال که حاوی مقادیر خارج از ردیف نیز هستند، مواجه می‌باشیم. بنابراین نیاز به برآوردگرهای مقاوم در مقابل داده‌های خارج از ردیف هستیم. فاصله مقاوم در این برآوردگرها از رابطه (۵) به دست می‌آمد:

$$RD(x_i) = \sqrt{(x_i-\hat{\mu}_{raw})'\hat{\Sigma}^{-1}_{raw}(x_i-\hat{\mu}_{raw})} \qquad i=1,2,...,n \qquad (۵)$$

که $\hat{\mu}_{raw}$ ماتریس موقعیت (یا مکان) و $\hat{\Sigma}_{raw}$ ماتریس پراکندگی (یا ماتریس مقیاس) نامیده می‌شود[۹]. تفاوت برآوردگر کلاسیک (رابطه ۲) با برآوردگرهای مقاوم (معادله ۵) تأثیر وزن‌دار نمونه‌ها در تخمین ماتریس‌های موقعیت و پراکندگی است. به طوری که در برآوردگرهای مقاوم داده‌های خارج از ردیف وزن کمتری در محاسبه ماتریس‌ها دارند. بنابراین این برآوردگرها نسبت به وجود این نوع داده‌ها مقاوم خواهند بود[۷، ۱۰]. انتخاب یک برآوردگر مقاوم مناسب در خواص نقطه فروریزش (Breakdown Point)، تابع نفوذ (Influence Function) و خاصیت هم‌وردای نسبی (Affine Equivariant) آن نهفته است.

نقطه فروریزش نشان دهنده‌ی عدم حساسیت یک برآوردگر نسبت به داده‌های خارج از ردیف است. مقدار این نقطه به صورت کسر نسبت تعداد داده‌های خارج از ردیف مؤثر در تابع برآوردگر به تعداد کل داده‌ها محاسبه می‌شود. به طور کلی هر چه نقطه فروریزش یک برآوردگر کمتر باشد، میزان حساسیت آن به داده‌های خارج از ردیف بیشتر خواهد بود. بنابراین به منظور کاهش میزان حساسیت نسبت به داده‌های خارج از ردیف بایستی به دنبال برآوردگرهایی با نقطه فروریزش بالا بود[۱۱].

تابع نفوذ نشان دهنده‌ی میزان اریبی یا نرخ تغییر برآوردگر به اضافه شدن نمونه‌های جدید به داده‌ها است. برآوردگری که نسبت به نقاط خارج از ردیف مقاوم است، دارای تابع نفوذ کراندار است. یعنی برآوردگر مقاوم، برآوردگری است که اجازه نمی‌دهد تغییرات در دسته داده باعث ایجاد تغییرات بزرگ در آن شود[۹].

ویژگی هم‌وردایی نسبی خاصیتی است که باعث می‌شود، ماتریس بهینه زیر مجموعه داده‌های اصلی به لحاظ موقعیت و پراکندگی همان خاصیت داده‌های اصلی را نیز دارا باشد، یعنی آنالیز ماتریس بهینه مستقل از مقیاس اندازه‌گیری متغیرها و همچنین انتقال و چرخش داده‌ها است[۱۲].

برآوردگرهای مقاوم چند متغیره به دو گروه برآوردگرهای دارای خاصیت هم‌وردای نسبی و فاقد آن تقسیم‌بندی

می‌شوند. متأسفانه بیشتر برآوردگرهای دارای خاصیت هم‌وردای نسبی دارای نقطه فروریزش پایینی هستند. بنابراین در انتخاب برآوردگرها بایستی به دو نکته سازگاری و کارایی آنها نیز توجه نمود. در ادامه ضمن ارائه کلیاتی از چند برآوردگر مقاوم به این نکات نیز اشاره خواهد شد.

۳-۱- برآوردگر MVE

این برآوردگر برای اولین بار توسط رسئو پیشنهاد گردیده است[۱۳]. در این روش $\widehat{\mu}_{MVE}$ برآوردگر موقعیت چند متغیره است که نشان دهنده‌ی مرکز بیضی‌وار با حجم مینیم است. این بیضی‌وار حداقل h نمونه از مجموعه‌ی داده‌های اصلی را که بیشتر در مرکز ابر داده‌ها قرار دارند، شامل می-شود. مقدار h توسط کاربر تعیین می‌شود به نحوی که بین $[\frac{n}{2}]+1$ تا n قرار داشته باشد ($[\frac{n}{2}]+1\le h\le n$). به طور معمول برای داشتن نقطه فروریزش بالا $h=[\frac{n+p+1}{2}]$ در نظر گرفته می‌شود[۱۴]. برآورد ماتریس پراکندگی $(\widehat{\Sigma}_{MVE})$ نیز از روی زیر مجموعه hتای نمونه‌های انتخاب شده به دست می‌آید. برای به دست آوردن برآوردگری سازگار، مقدار ماتریس پراکندگی در یک ضریب ثابت ضرب می‌شود. این ضریب از رابطه $\sqrt{\chi^2_{(p,0.5)}}$ محاسبه می‌شود، که مقدار جزر میانه توزیع خی دو است[۱۴]. بنابراین فاصله هر نمونه از مرکز بیضی‌وار از رابطه (۶) به دست خواهد آمد:

$$d_{MVE}(x_i)=\sqrt{(x_i-\widehat{\mu}_{MVE})'\widehat{\Sigma}^{-1}_{MVE}(x_i-\widehat{\mu}_{MVE})}\qquad i=1,2,\ldots,n \qquad (۶)$$

نقطه فروریزش این برآوردگر نیز از رابطه (۷) به دست می‌آید:

$$\varepsilon^*=\frac{\min(n-h+1,h-p)}{n} \qquad (۷)$$

می‌توان نشان داد که، وقتی $n\to\infty$ این برآوردگر دارای بیشترین نقطه فروریزش و برابر ۵۰٪ است[۱۴]. برای تعیین زیرمجموعه بهینه‌ای که دارای این خواص باشند، باید تمام زیرمجموعه ممکن از $p+1$ عضوی به بالا را مورد ارزیابی قرار داد. بنابراین بایستی از میان $\binom{n}{h}$ بیضی‌وار، بیضی‌واری با کمترین حجم را پیدا نمود، که با افزایش n و p تعداد این زیر مجموعه‌ها به شدت افزایش می‌یابد. حل این مشکل به همراه حفظ شرایط سازگاری و کارایی یک از مهم‌ترین نکاتی است که در طول سه دهه گذشته مورد توجه محققین بوده است. در این خصوص راه‌کارهای متفاوتی ارائه شده است که از مهم‌ترین آنها می‌توان به الگوریتم‌های پیشنهادی MINVOL، MVEE، MVCE و FAST-MVE اشاره نمود[۱۵-۱۸].

۳-۲- برآوردگر MCD

این برآوردگر نیز برای اولین بار توسط روسئو برای محاسبه تابع رگرسیون به روش حداقل توان دوم میانه ارائه شده است [۱۳]. در این برآوردگر، هدف یافتن h داده از n داده اولیه است که دترمینان ماتریس کواریانس آنها کمترین مقدار ممکن را داشته باشد. بنابراین اگر $\widehat{\mu}_{MCD}$ ماتریس موقعیت این زیر دسته داده بهینه و $\widehat{\Sigma}_{MCD}$ ماتریس پراکندگی آن باشد، فاصله مقاوم هر داده از مرکز ابر داده‌ها از رابطه (۸) به دست خواهد آمد[۹].

$$d_{MCD}(x_i)=\sqrt{(x_i-\widehat{\mu}_{MCD})'\widehat{\Sigma}^{-1}_{MCD}(x_i-\widehat{\mu}_{MCD})}\qquad i=1,2,\ldots,n \qquad (۸)$$

به منظور جلوگیری از ایجاد شرایط تکینه‌گی برای ماتریس پراکندگی، تعداد داده‌های زیر دسته بهینه بایستی از تعداد متغیره‌ها بیشتر باشد ($h>p$). بنابراین این تعداد به صورت $[\frac{n+p+1}{2}]\le h\le n$ انتخاب می‌شود. اگر تعداد داده‌های خارج از ردیف کمتر از ۲۵ ٪ داده‌ها را تشکیل داده باشد، بهترین انتخاب برای h با حفظ نقطه فروریزش و کارایی آماری مناسب، مقدار $h=0.75n$ است[۱۹]. بیشترین مقدار نقطه فروریزش این برآوردگر نیز از رابطه (۹) به دست خواهد آمد:

$$\varepsilon^*=\frac{\min(n-h+1,h-k(X_n))}{n} \qquad (۹)$$

که $k(X_n)$ حداکثر تعداد نمونه‌ها، از مجموعه داده‌های اصلی است که بر روی ابر صفحه $\mathbb{R}^p$ قرار می‌گیرد[۹]. برای تخمین زیر دسته داده‌ی بهینه، الگوریتم‌های زیادی تاکنون ارائه شده است که مهم‌ترین آنها می‌توان به الگوریتم‌های Det-MCD و FAST-MCD اشاره نمود[۱۲، ۱۹]. در هنگام محاسبه ماتریس موقعیت و پراکندگی مقاوم بایستی به فاکتورهای عامل سازگاری، عامل تصحیح نمونه متناهی و برآوردگرهای مکانی و پراکندگی وزن‌دار نیز توجه نمود.

۳-۳- برآوردگر SD

این برآوردگر به طور جداگانه توسط استاهل و دانهو به ترتیب در سال‌های ۱۹۸۱ و ۱۹۸۲ ارائه شده است. در این برآوردگر با نگاشت یک نمونه چند متغیره به یک فضای تک متغیره، برای هر نمونه یک مقدار پرتی (Outlyingness) تعریف می‌شود[۲۰]. اگر μ و σ به ترتیب نشان دهنده‌ی پارامترهای موقعیت و پراکندگی تک متغیره باشند که دارای خاصیت هم‌وردای مقیاس باشند، مقدار پرتی هر نمونه از رابطه (۱۰) به دست می‌آید:

$$r(x_i : X) = \sup_{a \in S_p} \left| \frac{x_i' a - \mu(Xa)}{\sigma(Xa)} \right| \qquad (۱۰)$$

که $S_p = \{a \in R^p : \|a\| = 1\}$ و $r(x_i; X)$ مقدار پرتی هر نمونه است که به صورت r_i نمایش داده می‌شود[۲۱]. بنابراین اگر $\hat{\mu}_{SD}$ و $\hat{\Sigma}_{SD}$ به ترتیب ماتریس موقعیت و پراکندگی برآوردگر SD باشد:

$$\hat{\mu}_{SD} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

$$\hat{\Sigma}_{SD} = \frac{\sum_{i=1}^{n} w_i (x_i - \hat{\mu}_{SD})(x_i - \hat{\mu}_{SD})'}{\sum_{i=1}^{n} w_i} \qquad (۱۱)$$

$$w_i = w(r_i)$$

$$\text{and } w: R^+ \to R^+ \qquad (۱۲)$$

که w_i تابع وزن نمونه است ($w_i = \{0 , 1\}$). به طوری که نمونه‌های خارج از ردیف مقدار وزن کم و نمونه‌های دیگر وزن بیشتری را خواهند گرفت. برای محاسبه تابع وزن w_i روش‌های مختلفی همچون تابع وزنی گوسی، نمایی و هابر ارائه شده است، که تفاوت معنی‌دار در استفاده از نوع این توابع وجود ندارد[۲۲]. مقدار وزن در تابع وزنی هابر از رابطه (۱۳) به دست می‌آید[۲۳]:

$$w(r_i) = I_{(r \le c)} + (\frac{c}{r})^2 I_{(r > c)} \qquad (۱۳)$$

در رابطه فوق c مقدار آستانه است. در انتخاب آستانه بایستی تعادلی بین مقاومی و کارایی برآوردگر صورت گیرد. مقدار آستانه کم باعث کاهش وزن نمونه‌های می‌شود که کمی خارج از ردیف هستند. در صورتی که انتخاب آستانه زیاد باعث کم وزن شدن نمونه‌های با مقدار خارج از ردیفی زیاد می‌شود. برای حفظ مقاوم بودن برآوردگر در داده‌ها با ابعاد زیاد می‌توان از مقدار آستانه کم استفاده شود. بنابراین مقدار $c = \min(\sqrt{\chi^2_{p,0.50}}, 4)$ برای آستانه پیشنهاد شده است[۷].

به منظور افزایش نقطه فروریزش در برآوردگر SD می‌توان از مقدار میانه و انحراف مطلق از میانه به جای آماره‌های موقعیت و پراکندگی در رابطه ۱۰ استفاده نمود. همچنین استفاده از ماتریس کواریانس با کمترین دترمینال داده‌هایی که دارای مقدار پرتی کمتری هستند، باعث افزایش کارایی این برآوردگر خواهد شد[۲۱، ۲۴].

۳-۴- برآوردگر S

برآوردگرهای S اولین بار توسط روسئو و یوهای در سال ۱۹۸۴ ارائه گردید که سپس توسط دیویس تکمیل شد[۲۵]. در این برآوردگرها ماتریس‌های موقعیت و پراکندگی $(\hat{\mu}_S, \hat{\Sigma}_S)$ به نحوی تخمین زده می‌شود که دترمینال ماتریس پراکندگی تحت شرایط زیر کمینه باشد:

$$\frac{1}{n} \sum_{i=1}^{n} \rho(\sqrt{(x_i - \hat{\mu}_S)' \hat{\Sigma}_S^{-1} (x_i - \hat{\mu}_S)}) = b \qquad (۱۴)$$

که $\hat{\mu}_S$ برداری متعلق به R^p و $\hat{\Sigma}_S$ یک ماتریس متقارن معین مثبت با ابعاد $p \times p$ است. به منظور به دست آوردن برآوردگرهای با نقطه فروریزش مثبت، تابع ρ باید دارای شرایط زیر باشد[۲۶]:

۱- تابع ρ اطراف صفر متقارن بوده و دارای مشتق مرتبه دوم پیوسته باشد.

۲- تابع ρ در بازه $[0, C_0]$ اکیداً صعودی (برای $C_0 > 0$)، در بازه $[C_0, \infty]$ ثابت و $\rho(0) = 0$ باشد.

ρ از تابع دو وزنی توکی (Tukey's biweight) به صورت زیر محاسبه می‌شود.

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2C_0^2} + \frac{x^6}{6C_0^4} & , |x| \le C_0 \\ \frac{C_0^2}{6} & , |x| > C_0 \end{cases} \qquad (۱۵)$$

مقدار ثابت b نیز از رابطه $E_{F_0}[\rho(\|z\|)]$ محاسبه می‌شود، که $F_0 = N(0, I_p)$ برای اطمینان از سازگاری در مدل نرمال قرار دارد. در این حالت مقدار فروریزش برآوردگر S برابر $b/\rho(C_0)$ است. تحت شرایط مدل نرمال، مقدار b می‌تواند از رابطه (۱۶) به دست آید[۲۶]:

$$b = \frac{p}{2}\chi^2_{p+2}(C_0^2) - \frac{p(p+2)}{2C_0^2}\chi^2_{p+4}(C_0^2) + \frac{p(p+2)(p+4)}{6C_0^2}\chi^2_{p+6}(C_0^2) + \frac{C_0^2}{6}(1-\chi^2_p(C_0^2)) \quad (16)$$

که χ^2_v تابع cdf توزیع خی‌دو با درجه آزادی v است. مقدار ثابت C_0 با توجه به انتخاب نقطه فروریزش از جدول ۱ برآورد می‌گردد.

جدول ۱: رابطه بین مقدار ثابت C_0 و نقطه فروریزش

C_o	نقطه فروریزش (درصد)
۱/۵۴۷	۵۰
۱/۷۵۶	۴۵
۱/۹۸۸	۴۰
۲/۲۵۱	۳۵
۲/۵۶۰	۳۰
۲/۹۳۷	۲۵
۳/۴۳۰	۲۰
۴/۰۹۶	۱۵
۵/۱۸۳	۱۰

تاکنون سه الگوریتم مختلف برای محاسبه ماتریس‌های موقعیت و پراکندگی برآوردگر S ارائه شده است که از آن جمله می‌توان به الگوریتم بر پایه زیرنمونه‌گیری، الگوریتم Fast-S و الگوریتم Det-S اشاره نمود[۲۶، ۲۷].

در نهایت به منظور شناسایی نقاط خارج از ردیف به کمک برآوردگرهای مقاوم باید مراحل زیر را انجام داد:

- محاسبه فاصله مقاوم به کمک یکی از روش‌های ذکر شده (به همراه رعایت نکات فوق).
- اگر نمونه‌ای به زیرمجموعه بهینه تعلق داشته باشد، فاصله مقاوم آن با صدک ۹۷/۵ درصد توزیع خی‌دو یعنی $\sqrt{\chi^2_{p,0.975}}$ مقایسه می‌شود.
- اگر نمونه‌ای به زیر مجموعه بهینه تعلق نداشته باشد، فاصله مقاوم آن با صدک $m\text{-}p+1$ توزیع F یعنی $F_{p,m-p+1}$ مقایسه می‌شود.

اگر فاصله مقاوم محاسبه شده برای نمونه‌ای از مقدار توزیع خی‌دو یا F به دست آمده از جدول بیشتر باشد، نمونه یک داده خارج از ردیف محسوب خواهد شد. در غیر این صورت نمونه به داده‌های اصلی تعلق خواهد داشت.

۴- معرفی منطقه مطالعاتی و داده‌های ژئوشیمیایی

محدوده مطالعاتی شاه سلیمان علی با مساحت ۴۰ کیلومترمربع در ۱۲۰ کیلومتری جنوب غربی شهر بیرجند در استان خراسان جنوبی و در برگه ۱:۱۰۰۰۰۰ زمین‌شناسی مختاران قرار دارد. بر اساس تقسیمات ساختمانی- رسوبی ایران، منطقه شاه سلیمان علی در خاور خرد قاره ایران مرکزی و در خاور مرکز بلوک لوت و در نزدیکی محل اتصال این بلوک با حوضه فلیش کرتاسه نهبندان- خاش یا زون زمین درز سیستان واقع است[۲۸]. بر اساس نقشه‌های زمین‌شناسی ۱:۲۵۰۰۰۰ بیرجند و ۱:۱۰۰۰۰۰ مختاران، منطقه دارای مجموعه‌ای از سنگ‌های ولکانیکی، آذرآواری و توف مارنی است[۲۹]. سن این سنگ‌ها را به کرتاسه انتهایی نسبت داده‌اند که به وسیله سنگ‌های آتشفشانی- رسوبی ترشیری پوشیده شده‌اند. سیمای غالب منطقه‌ی مطالعاتی را نهشته‌های آذرآواری و سنگ‌های آتشفشانی نیمه عمیق تشکیل می‌دهد. بر اساس مطالعات ۱:۲۰۰۰۰ زمین‌شناسی، محدوده شاه سلیمان علی شامل واحدهای سنگی زیر است (شکل ۱):

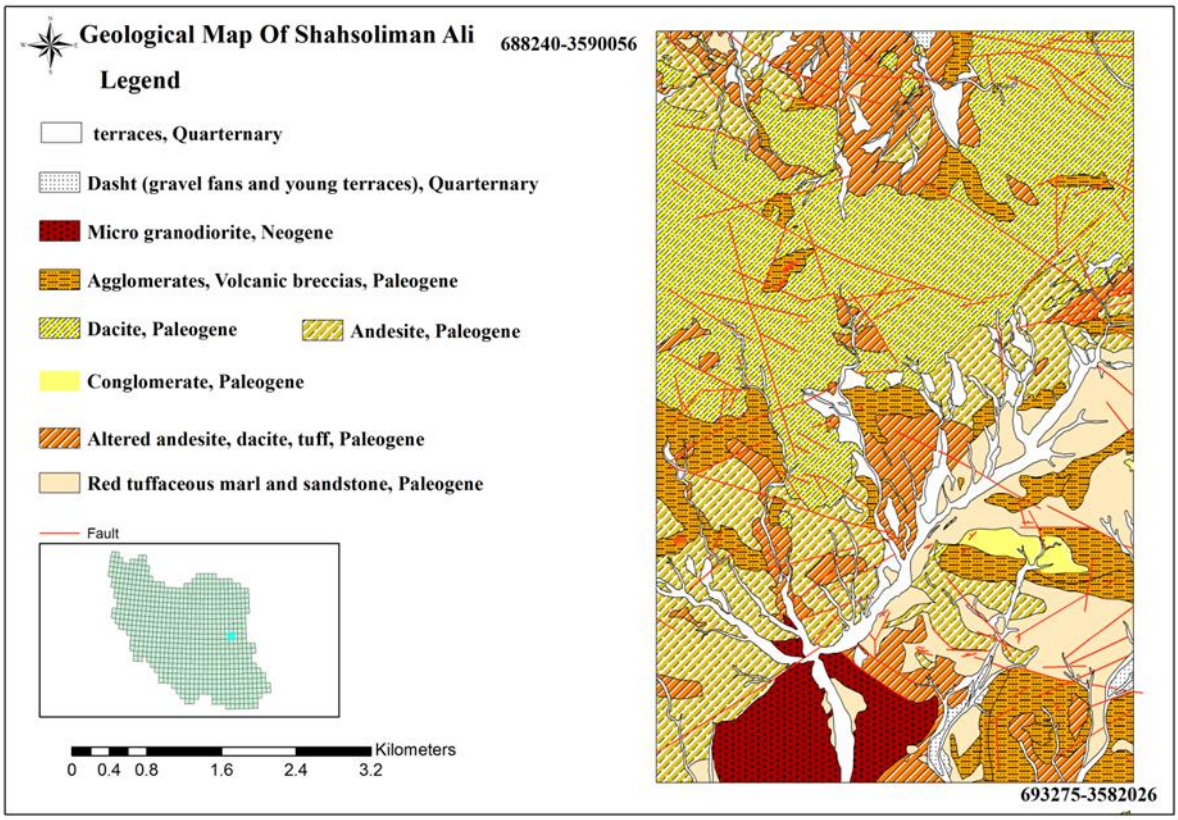
۱- کنگلومرا، مارن توفی و ماسه سنگی قرمز با سن پالئوژن تحتانی که عمدتاً در بخش جنوب شرقی منطقه قرار گرفته‌اند.

۲- سنگ‌های آذرین شامل برش آتشفشانی، آگلومرا، آندزیت، داسیت، توف دگرسان شده به سن ائوسن و الیگوسن که بخش اعظمی از منطقه را تحت پوشش قرار داده‌اند.

۳- میکروگرانودیوریت با سن نئوژن در بخش‌های جنوبی منطقه مطالعاتی.

بر اساس مطالعات صورت گرفته توسط کریم پور و همکاران در منطقه کوه شاه سلیمان علی، فعالیت‌های ماگمایی از نوع توده‌های نفوذی و نیمه عمیق متعدد با

نفوذهای تلسکوپی در یکدیگر و در واحدهای ولکانیکی است که دارای سن ترشیری (پس از نئوژن) است[۳۰]. این توده‌ها طیف سنگی بین دیوریت تا مونزوگرانیت دارند. منطقه از لحاظ تکتونیکی نیز شامل: ۱- گسل‌های اولیه یا گسل‌های در مقیاس ناحیه‌ای با روند N50E، ۲- گسل‌های امتداد لغز راستگرد با روند N50W که همراه با گسل‌های R و R' (گسل‌های شیب لغز معکوس و امتداد لغز چپگرد با روند S50W) پهنه‌های برشی را به وجود آورده است و ۳- گسل‌های نرمال تقریباً شمالی- جنوبی با روند N7W که این گسل‌ها بسیار جوان بوده و سبب جابجایی‌های به شکل پایین افتادگی و بالا آمدگی در واحدهای زمین‌شناسی و حتی آبرفت‌ها شده است.



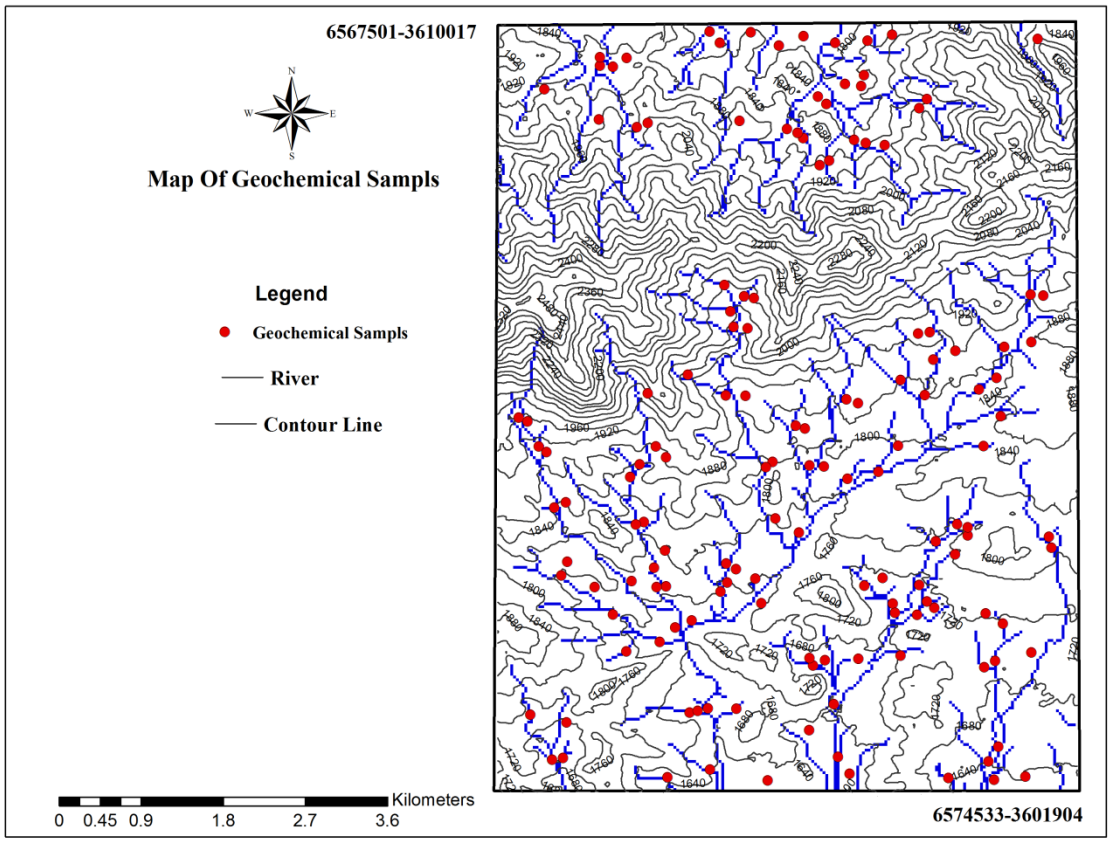
شکل ۱: موقعیت منطقه مطالعاتی به همراه نقشه زمین‌شناسی محدوده شاه سلیمان علی

بر اساس مطالعات عبدی و همکاران کانی‌سازی در منطقه کو ه شاه، در جایگاه تکتونیکی مناسب (حاشیه بلوک لوت) و مرتبط با گستره‌ای از نفوذهای متعدد و تلسکوپی توده‌های نیمه عمیق حد واسط و دگرسانی‌های مرتبط با آنها شکل گرفته است. دگرسانی‌های این منطقه شامل دگرسانی کوارتز- سرسیت- پیریت (لیمونیت)، پروپلیتیک، آرژیلیک، سیلیسی- آرژیلیک، آرژیلیک پیشرفته، گوسان، برش هیدروترمالی و سیلیسی است. در این منطقه کانی‌سازی به‌صورت سولفیدی (پیریت و به مقدار کمتر کالکوپیریت) و اکسیدهای آهن، به اشکال پراکنده در متن سنگ، رگچه‌ای، استوک‌ورک لیمونیت- ژاروسیت، برش هیدروترمالی، رگه‌های موازی کوارتز- کلسیت- لیمونیت (پیریت) صفحه‌ای و تنوره‌های برشی هیدروترمال- نفوذی مشاهده می‌شود[۳۱].

حضور، گسترش و شدت میزان اکسیدهای آهن پراکنده و رگچه‌های متعدد کوارتز- اکسید آهن (لیمونیت+ ژاروسیت ±گوتیت) و لیمونیت+ کلسیت+ کوارتز، حاکی از قرارگیری زون‌های سولفیدی با عیار بالا در معرض هوازدگی (عمده کانی‌سازی سولفیدی به گوتیت، هماتیت و ژاروسیت اکسیده شده است) و اسیدشویی است. زون‌بندی‌های آلتراسیون از نظر تنوع، گسترش، شکل و خصوصیات کانی‌سازی در این منطقه، مشابه با سیستم‌های مس- طلای پورفیری و اپی‌ترمال با سولفیداسیون بالا است[۳۱].

از محدوده مطالعاتی ۱۴۶ نمونه‌ی ژئوشیمیایی از بستر رسوبات آبراهه‌ها برداشت شده است (شکل ۲). آنالیز

شیمیایی برای ۴۴ عنصر به روش ICP-MS و طلا به روش Fire Assay بر روی بخش ۱۰۰-۴۰ مش نمونه‌ها صورت گرفته است. به منظور بررسی صحت آنالیزهای شیمیایی ۱۲ نمونه تکراری نیز برداشت شده است. بررسی صحت آنالیزها به روش تامسون- هاوارث، نشان دهنده‌ی انحراف معیار نسبی (RSD) کمتر از ۱۰ درصد برای کلیه عناصر است. پس از جایگزینی مقادیر داده‌های سنسورد با نصف حد تشخیص، ۱۸ عنصر مرتبط با کانی‌سازی‌ها و زون‌های آلتراسیون جهت تشخیص نمونه‌های خارج از ردیف به روش‌های ذکر شده انتخاب شده‌اند. لازم به ذکر است، که برآوردگرهای مقاوم محدودیتی به جهت تعداد متغیرها ندارند. ولی به جهت تفسیر بهتر نتایج و جلوگیری از ایجاد وضعیت تکینه‌گی ماتریس واریانس- کواریانس، این تعداد عنصر انتخاب شده‌اند.



شکل ۲: موقعیت نمونه‌های برداشت شده از رسوبات آبراهه‌ای در منطقه شاه سلیمان علی

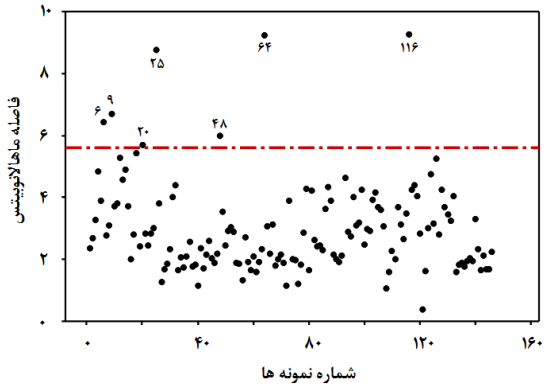
در جدول ۲ پارامترهای آمار توصیفی داده‌ها ژئوشیمیایی نشان داده شده است. پارامتر انحراف معیار یا انحرف معیار نسبی بالا می‌تواند به طور غیر مستقیم نشان دهنده‌ی وجود داده خارج از ردیف باشد. بنابراین احتمال وجود داده خارج از ردیف تک متغیره در عناصر طلا، مس، نیکل و آنتیموان بالا خواهد بود. همچنین پارامترهای چولگی و کشیدگی داده‌ها نشان دهنده‌ی غیر نرمال بودن توزیع اکثر عناصر به استثناء کلسیم است. توزیع داده‌های عناصری از قبیل طلا، کروم، مولیبدن، آنتیموان و روی به دلیل چولگی بیشتر از ۲ به شدت از توزیع نرمال فاصله دارند. این نکته نشان دهنده‌ی چند جامعه‌ای بودن توزیع آنها و در نتیجه وجود داده خارج از ردیف تک متغیره در توزیع آنها است. احتمال وجود داده‌ی خارج از ردیف در توزیع عناصر آرسنیک، آهن، منیزیم، نیکل و تیتانیوم، که دارای چولگی بین ۱ تا ۲ هستند نیز در رتبه بعدی قرار دارد. بنابراین از اطلاعات جدول ۲ می‌توان نتیجه گرفت که احتمال وجود داده‌های خارج از ردیف چند متغیره بسیار بالا است. برای این منظور، در ادامه از روش‌های کلاسیک و مقاوم برای تشخیص این نوع داده‌ها استفاده خواهد شد.

جدول ۲: پارامترهای آماری داده‌های ژئوشیمیایی منطقه مطالعاتی

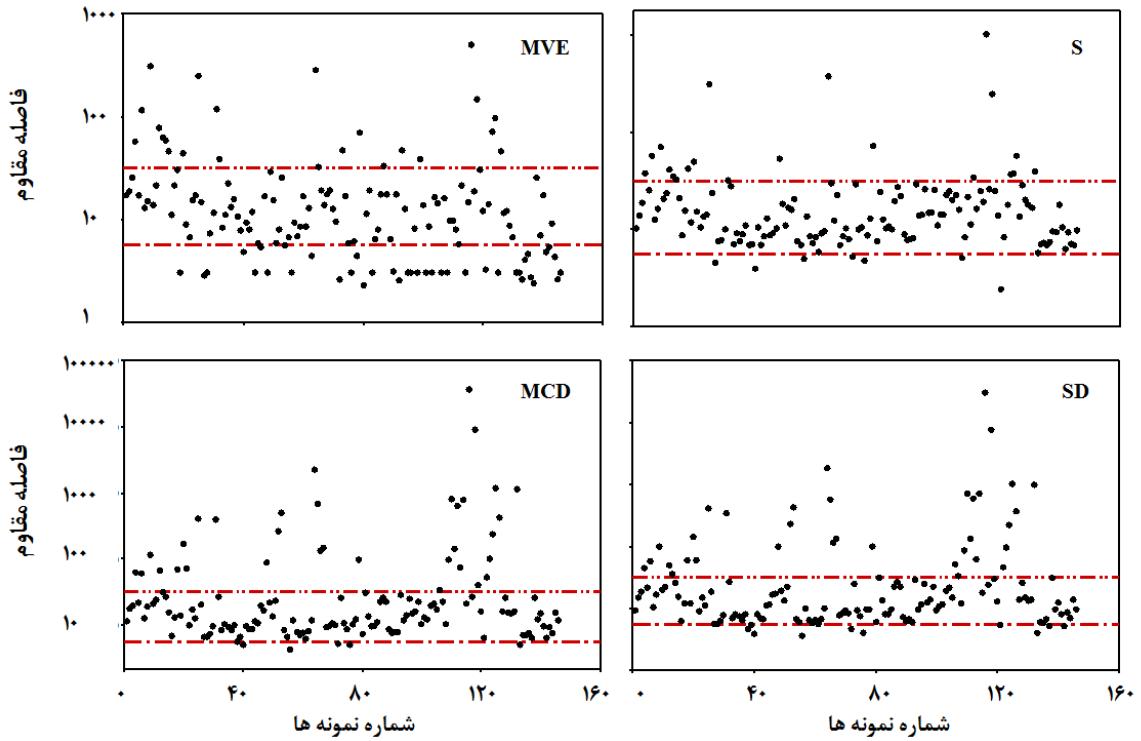
عناصر (واحد اندازه گیری)	کمینه	بیشینه	میانگین	انحراف معیار	چولگی	کشیدگی	MAD
Ag (ppm)	۰/۰۱	۱/۳۵	۰/۵۷	۰/۳	۰/۷	-۰/۰۳	۰/۱۹
As (ppm)	۲/۶	۵۳/۶	۱۶/۴	۹/۹	۱/۳	۱/۷	۵/۲
Au (ppb)	۰/۵	۱۸	۱/۸	۱/۷	۶/۱	۵۰/۹	۱
Ba (ppm)	۲۳۲	۷۹۲	۴۹۷	۱۳۵	-۰/۱	-۰/۶	۸۱
Ca (%)	۳/۰۲	۶/۷	۴/۸	۰/۷	۰/۰۷	-۰/۶	۰/۶
Co (ppm)	۸/۱	۳۳/۲	۱۸/۴	۵/۴	۰/۷	-۰/۱	۳/۰۵
Cr (ppm)	۹	۱۶۹۰	۱۱۶/۹	۱۹۵/۱	۴/۶	۳۰/۲	۱۶/۵
Cu (ppm)	۲۰/۵	۱۰۰	۴۸/۳	۱۲/۵	۰/۵	۱/۳	۷/۵
Fe (%)	۲/۹	۱۲	۵/۸	۱/۶	۱/۰	۱/۶	۱/۰۲
K (%)	۰/۶	۳/۸	۱/۹	۰/۵	۰/۳	۰/۷	۰/۳
Mg (%)	۰/۳	۶/۲	۱/۸	۱/۱	۱/۳	۱/۸	۰/۵
Mn (ppm)	۵۳۷	۱۸۴۰	۱۱۲۹	۲۲۴	۰/۳	۰/۲	۱۳۸
Mo (ppm)	۰/۴	۴/۶	۱/۱	۰/۵	۳/۳	۱۶/۴	۱
Ni (ppm)	۸	۳۷۵	۶۷/۶	۷۹/۱	۱/۹	۲/۸	۱۱
Pb (ppm)	۶/۴	۴۱/۲	۲۳	۵/۱	۰/۴	۱/۵	۲/۹
Sb (ppm)	۰/۳	۲۱/۳	۲/۴	۳/۳	۳	۱۰	۰/۵
Ti (ppm)	۲۴۲۰	۹۴۶۰	۴۴۲۳	۱۲۰۴	۱/۳	۲/۹	۶۲۵
Zn (ppm)	۲۶/۸	۳۱۴	۹۰/۷	۳۲/۱	۲/۸	۱۵/۶	۱۵/۸

۵- تعیین داده‌های خارج از ردیف چند متغیره

در این مقاله، به منظور محاسبه‌ی فاصله هر نمونه از مرکز داده‌ها، نرم افزار متلب به کار رفته است. در مرحله اول از روش کلاسیک برای تعیین فاصله ماهالانوبیتس نمونه‌ها استفاده شده است. برای این منظور فاصله هر نمونه از مرکز داده‌ها با استفاده از فرمول ۲ برآورد شده، که نتایج آن در شکل ۳ ارائه شده است. برای تعیین نمونه‌های خارج از ردیف فاصله به دست آمده با مقدار $\sqrt{\chi^2_{18,0.975}} = 5.61$ مقایسه می‌شود (خط نقطه نشان داده شده در شکل ۳). سه نمونه (نمونه‌های شماره ۲۵، ۶۴ و ۱۱۶) به طور معنی‌دار از خط آستانه فاصله دارند. همچنین ۴ نمونه‌ی دیگر (نمونه‌های شماره ۶، ۹، ۲۰، ۲۵ و ۴۸) تا حدودی با این خط فاصله دارند. بنابراین ۷ نمونه از ۱۴۶ نمونه (حدود ۵ درصد نمونه‌ها) به روش کلاسیک، خارج از ردیف محسوب می‌شوند (شکل ۳). در مرحله دوم از چهار برآوردگر مقاومی که در بالا معرفی شده‌اند برای تعیین نمونه‌های خارج از ردیف استفاده شده است. به منظور افزایش کارایی برآوردگرها از نقطه فروریزش ۰/۲۵ و سطح اعتماد ۹۷/۵ درصد نیز در محاسبات استفاده شده است. نتایج محاسبه فاصله مقاوم برای هر نمونه در هر روش، در شکل ۴ آورده شده است. برای تعیین خارج از ردیف بودن نمونه‌های متعلق به زیر مجموعه بهینه، از حد آستانه $\sqrt{\chi^2_{18,0.975}} = 5.61$ (خط نقطه‌های نشان داده شده در شکل ۴) و برای سایر نمونه‌ها از حد آستانه $F_{18,2,0.975} = 31.46$ (خط دو نقطه‌های نشان داده شده در شکل ۴) استفاده می‌شود.



شکل ۳: نمودار فاصله ماهالانوبیتس داده‌های منطقه مطالعاتی به همراه خط آستانه تشخیص نمونه‌ی خارج از ردیف

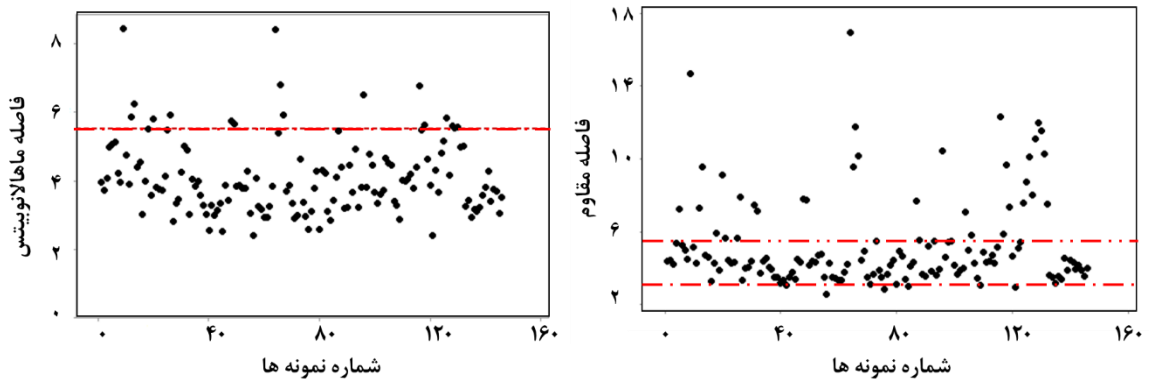


شکل ۴: نمودار فاصله مقاوم نمونه‌ها به چهار روش MVE، MCD، S و SD (در این شکل‌ها خط نقطه برابر فاصله ۵/۶۱ و خط دو نقطه برابر فاصله ۳۱/۴۶ است)

محاسبات انجام با استفاده از الگوریتم Fast نشان داده است که زیر مجموعه بهینه در روش‌های MVE و MCD از ۱۱۴ نمونه (h=۱۱۴) که کمترین فاصله را با مرکز داده‌ها دارند، تشکیل شده است. در روش MVE نمونه‌های شماره‌ی ۴، ۶، ۹، ۱۲، ۱۳، ۱۴، ۱۵، ۲۰، ۲۵، ۳۱، ۳۲، ۶۴، ۶۵، ۷۳، ۷۹، ۸۷، ۹۳، ۹۹، ۱۱۶، ۱۱۸، ۱۲۳، ۱۲۴ و ۱۲۶ خارج از ردیف محسوب می‌شود. یعنی از ۱۴۶ نمونه، ۲۳ نمونه (حدوداً ۱۵/۸٪ از نمونه‌ها) در این روش داده‌ی پرت در نظر گرفته شده است. در حالیکه در روش MCD نمونه‌های شماره ۴، ۶، ۹، ۱۲، ۱۳، ۱۸، ۲۰، ۲۱، ۲۵، ۳۱، ۴۸، ۵۲، ۵۳، ۶۴، ۶۵، ۶۶، ۶۷، ۷۹، ۸۱، ۱۰۶، ۱۰۹، ۱۱۰، ۱۱۱، ۱۱۲، ۱۱۳، ۱۱۴، ۱۱۶، ۱۱۸، ۱۱۹، ۱۲۲، ۱۲۳، ۱۲۴، ۱۲۵، ۱۲۶ و ۱۳۲ یعنی ۳۵ نمونه (حدوداً ۲۴٪ از نمونه‌ها) خارج از ردیف به دست آمده‌اند. این مطلب نشان می‌دهد که تعداد نمونه‌های خارج از ردیفی که در الگوریتم Fast-MCD به دست می‌آید از الگوریتم Fast-MVE بیشتر خواهد بود. در برآوردگر مقاوم Fast-S، ۲۰ نمونه (حدوداً ۱۳/۷ ٪ از نمونه-ها) یعنی نمونه‌ها شماره ۴، ۶، ۹، ۱۲، ۱۳، ۱۴، ۱۸، ۲۰، ۲۵، ۳۱، ۴۸، ۶۴، ۷۹، ۱۱۲، ۱۱۶، ۱۱۸، ۱۲۴، ۱۲۵، ۱۲۶ و ۱۳۲ خارج از ردیف هستند. همچنین در برآوردگر مقاوم SD نمونه‌ها شماره ۴، ۶، ۹، ۱۲، ۱۳، ۱۸، ۲۰، ۲۱، ۲۵، ۳۱، ۴۸، ۵۲، ۵۳، ۶۴، ۶۵، ۶۶، ۶۷، ۷۹، ۱۰۶، ۱۰۷، ۱۰۹، ۱۱۰، ۱۱۱، ۱۱۲، ۱۱۳، ۱۱۴، ۱۱۶، ۱۱۸، ۱۲۲، ۱۲۳، ۱۲۴، ۱۲۵، ۱۲۶ و ۱۳۲ یعنی ۳۴ نمونه (حدوداً ۲۳/۲٪ از نمونه‌ها) خارج از ردیف محسوب می‌شوند.

به منظور ارزیابی تأثیر سیستم عددی باز و یا بسته بودن داده‌ها بر روی تعیین نمونه‌های خارج از ردیف، در مرحله بعدی از روش تبدیل لگاریتمی میان مرکز (Centered Log-Ratio (clr) Transformation) برای باز کردن سیستم عددی داده‌ها استفاده شده است. عدم حذف هیچ یک متغیر از مهم‌ترین مزیت این روش تبدیل لگاریتمی محسوب می‌شود[۳۲]. در شکل ۵ (سمت چپ) مقادیر فاصله ماهالونوبیتس نمونه‌ها در سیستم عددی باز نشان داده شده است. مطابق شکل، نمونه‌های ۹، ۱۲، ۱۳، ۱۸، ۲۰، ۲۵، ۲۶، ۴۸، ۴۹، ۶۴، ۶۶، ۶۷، ۹۶، ۱۱۶، ۱۱۸، ۱۲۶، ۱۲۸ و ۱۳۰ به عنوان نمونه‌ی خارج از ردیف محسوب می‌شوند. این نکته نشان می‌دهد که باز کردن داده‌ها در روش کلاسیک باعث شناسایی بهتر نمونه‌های خارج از ردیف می‌شود. در شکل ۵ (سمت راست) مقادیر فاصله مقاوم نمونه‌ها با سیستم عددی باز و به کمک برآوردگر مقاوم MCD آورده شده است (به

دلیل شناسایی بیشترین تعداد نمونه‌ی خارج از ردیف در مرحله قبل، این برآوردگر مقاوم انتخاب شده است). نمونه‌های خارج از ردیف به دست آمده در این شکل جز در نمونه‌های ۵، ۳۲، ۸۷، ۸۸، ۹۶، ۱۲۷، ۱۲۸، ۱۲۹، ۱۳۰ و ۱۳۱ با شکل ۴ تقریباً یکسان است. این موضوع نشان دهنده‌ی تأثیر کم باز کردن داده‌ها بر روی عملکرد برآوردگرهای مقاوم دارد.



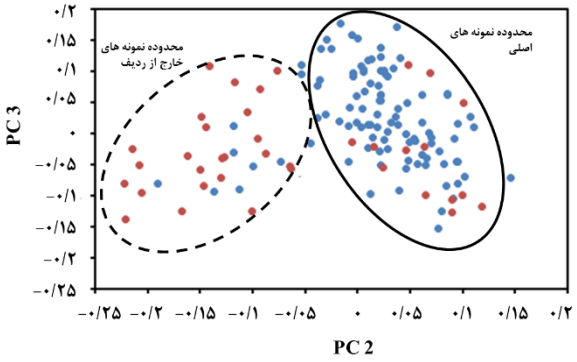
شکل ۵: نمودار فاصله ماهالانوبیتس (شکل سمت چپ) و فاصله مقاوم به روش MCD (شکل سمت راست) برای داده‌ها در سیستم عددی باز (مقادیر عددی خطوط مزر تعیین نمونه خارج از ردیف همانند شکل ۴ است)

از ۱۴۶ نمونه‌ی مورد بررسی در این مقاله، ۱۰۳ نمونه توسط هیچ کدام از برآوردگرهای مقاوم داده خارج از ردیف شناخته نشده‌اند. در حالیکه ۱۴ نمونه توسط هر چهار برآوردگر، بعنوان نمونه‌ی خارج از ردیف شناسایی شده است، که نمونه‌های خارج از ردیف شناسایی شده توسط روش کلاسیک نیز در این گروه قرار دارند. در کل ۴۳ نمونه با روش‌های برآوردگرهای مقاوم به عنوان نمونه‌ی خارج از ردیف در نظر گرفته شده‌اند. به منظور اعتبار سنجی عملکرد برآوردگرهای مقاوم، روش آنالیز مولفه‌های اصلی در مد Q بر روی نمونه‌ها انجام شده است. نتایج مقادیر ویژه که نشان دهنده‌ی شدت تغییرپذیری قابل توجیه توسط هر مولفه‌ی اصلی است، در جدول ۳ نشان داده شده است. این نتایج نشان می‌دهد که ۳ مولفه‌ی اصلی اول قادر است بیش از ۹۹ درصد تغییرپذیری را توجیه نماید. بنابراین ابعاد داده‌ها از ۱۴۶ بعد به ۳ بعد کاهش یافته است.

جدول ۳- مقادیر ویژه به همراه شدت تغییرپذیری هر مولفه

مولفه اصلی	PC 1	PC 2	PC 3
مقدار ویژه	۱۳۹/۰۴	۳/۶۵	۲/۵۱
نسبت تغییرپذیری	۰/۹۵۲	۰/۰۲۵	۰/۰۱۷
درصد تغییرپذیری تجمعی	۹۵/۲	۹۷/۷	۹۹/۴

نتایج تجزیه و تحلیل بردارهای ویژه نشان داده است که نمونه‌های خارج از ردیف با بارهای منفی خود را در مولفه‌ی اصلی دوم و سایر نمونه‌ها (نمونه‌های متعلق به زیر مجموعه بهینه) خود را در مولفه‌های اصلی اول و سوم نشان می‌دهند. اکثر نمونه‌های زیر مجموعه‌ی بهینه دارای بارها مثبت بالا در مولفه‌ی اول هستند. در مرحله بعد، نمودار پراکندگی بردارهای ویژه این مولفه‌ها نسبت به هم، دو به دو ترسیم شده است که در شکل ۶ یکی از این نمودارها نشان داده شده است. در این شکل نمونه‌هایی که توسط برآوردگرهای مقاوم خارج از ردیف تشخیص داده شده‌اند، با دایره قرمز و سایر نمونه‌ها با دایره آبی ترسیم شده‌اند.



شکل ۶: نمودار پراکندگی بردارهای ویژه مولفه اصلی دوم نسبت به مولفه اصلی سوم

مطابق شکل دو دسته داده در این نمودار به خوبی قابل تفکیک و شناسایی هستند. این نمودار نشان دهنده‌ی عملکرد مناسب برآوردگرهای مقاوم در شناسایی داده‌های خارج از ردیف محسوب می‌شود. بنابراین به طور کلی می‌توان گفت که هر چهار برآوردگر مقاوم قادر به شناسایی داده‌های خارج از ردیف هستند. در صورتی که تعداد نمونه‌ها زیاد باشد

و یا احتمال وجود داده‌ی پرت در مجموعه داده‌ها بالا باشد، می‌توان از برآوردگرهای MCD و SD استفاده کرد و برعکس اگر تعداد نمونه‌ها کم باشد و یا احتمال وجود داده‌ی پرت در مجموعه داده‌ها پایین باشد، می‌توان از برآوردگرهای MVE و S استفاده نمود.

۶- نتیجه‌گیری

شناسایی و تعدیل داده‌های خارج از ردیف یکی از چالش‌های مهم در پردازش اولیه داده‌های اکتشافی محسوب می‌شود. افزایش بعد داده‌ها نیز باعث پیچیده شدن بیشتر این مشکل خواهد شد. همچنین استفاده از روش‌های آمار چند متغیره نیز بدون تعدیل داده‌های پرت میسر نمی‌باشد. این نکات نشان دهنده‌ی اهمیت کاربرد برآوردگرهای مقاوم در شناسایی داده‌های خارج از ردیف در مباحث تحلیل داده‌های اکتشافی چند متغیره خواهد بود. در برآوردگرهای مقاوم با کاهش بعد داده‌ها به یک بعد، که این بعد نشان دهنده‌ی فاصله مقاوم هر نمونه از مرکز ابر داده‌ها است، و استفاده از یک حد آستانه امکان شناسایی داده‌ی خارج از ردیف میسر خواهد شد. به منظور محاسبه‌ی فاصله مقاوم هر نمونه از بردار ستونی موقعیت و ماتریس پراکندگی بخشی از داده‌ها استفاده می‌شود. تعیین این بخش از داده‌ها که تحت عنوان زیر مجموعه بهینه از آن نام برده می‌شود توسط روش‌های متفاوتی امکانپذیر است. در این مقاله چهار برآوردگر مقاوم MVE، MCD، S و SD برای این منظور معرفی گردیده‌اند.

از ۱۴۶ نمونه‌ی رسوبات آبراهه‌ای در منطقه اکتشافی شاه سلیمان علی در استان خراسان جنوبی و نتایج آنالیز ۱۸ عنصر برای هر نمونه استفاده شد. پارامترهای آمار توصیفی این متغیرها نشان دهنده‌ی وجود داده‌ی خارج از ردیف در مجموعه‌ی داده‌ها است. روش کلاسیک فاصله ماهالانوبیتس نشان می‌دهد که فقط ۷ نمونه دارای مقادیر خارج از ردیف می‌باشند. در حالیکه برآوردگرهای مقاوم MVE، MCD، S و SD به ترتیب ۲۳، ۳۵، ۲۰ و ۳۴ نمونه را به عنوان داده‌ی پرت معرفی می‌کنند. از نقطه فروریزش ۰/۲۵ و سطح اعتماد ۹۷/۵٪ برای افزایش کارایی برآوردگرها در محاسبات استفاده گردید. همچنین برای تعیین خارج از ردیف بودن نمونه‌های متعلق به زیر مجموعه‌ی بهینه از توزیع خی‌دو و حد آستانه ۵/۶۱ و سایر نمونه‌ها از توزیع F با حد آستانه ۳۱/۴۶ نیز بهره‌برداری شد. آنالیز مولفه‌های اصلی در مد Q نشان داد که نمونه‌های خارج از ردیف خود را در مولفه دوم و با بارهای منفی و سایر نمونه‌های در مولفه اول و سوم نشان می‌دهند. همچنین تفکیک جامعه‌ی نمونه‌های خارج از ردیف نسبت به سایر نمونه‌ها در نمودار پراکندگی بارهای PC2 بر حسب PC3 قابل نمایش است.

از برآوردگرهای مقاوم علاوه بر تعیین نمونه‌های خارج از ردیف در شناسایی داده‌های آنومال (یعنی آنومالی‌های ژئوشیمیایی مرکب) نیز می‌توان استفاده کرد. استفاده از ماتریس‌های موقعیت و پراکندگی به دست آمده از برآوردگرهای مقاوم در آمار چند متغیره، یکی دیگر از کاربردهای مهم این برآوردگرها محسوب می‌شود. ماتریس‌های موقعیت و پراکندگی می‌توانند به ترتیب به عنوان ماتریس‌های میانگین و واریانس- کواریانس در روش‌های آماری چند متغیره به کار روند. در این صورت روش‌های آماری بدون نیاز به حذف داده‌های خارج از ردیف نسبت به آنها مقاوم خواهند شد. روش‌های آنالیز مولفه‌های اصلی مقاوم (Robust PCA)، آنالیز فاکتوری مقاوم (Robust FA)، رگرسیون چند متغیره مقاوم (Robust MR)، آنالیز تمایز مقاوم (Robust DA)، آنالیزهای طبقه‌بندی مقاوم (Robust Classification Analysis) و آنالیزهای خوشه‌بندی مقاوم (Robust Clustering Analysis) از جمله این روش‌ها هستند.

سپاسگزاری

از سازمان صنعت، معدن و تجارت استان خراسان جنوبی به دلیل استفاده از داده‌های ژئوشیمیایی و زمین‌شناسی منطقه مطالعاتی تشکر و قدردانی می‌شود.

مراجع

[1] Hawkins, D.M. (1980). Identification of Outliers. Volume 13 of Monographs on statistics and applied probability, Chapman and Hall.

[2] Wellmer, F.W. (1998). Statistical Evaluations in Exploration for Mineral Deposit, Translated by D. Large, Springer-Verlag, Berlin Heidelberg.

[3] Filzmoser, P., Garrett, R.G., and C., Reimann (2005). Multivariate outlier detection in exploration geochemistry, Com. & Geosci. 31, 579–587.

[4] Zhang, R., Zhou, M., Gong, X., He, X., Qian, W., Qin, S., and A., Zhou (2015). Detecting anomaly in data streams by fractal model. World Wide Web 18(5), 1419-1441.

[5] Aggarwal, C.C. (2013). Outlier Analysis, Springer, New York.

[6] Santos-Pereira, C.M., and A.M., Pires (2002). Detection of outliers in multivariate data: A method based on clustering and robust estimation, In Härdle, W., Rőnz, B., (eds), Compstat, Physica-Verlag Heidelberg, 291-296.

[7] Maronna, R.A., Martin, R.D., and V.J., Yohai (2006). Robust Statistics: Theory and Methods, John Wiley & Sons.

[8] Maronna, R. A., and R., Zamar (2002). Robust estimation of location and dispersion for high-Dimensional datasets. Technom. 44, 307-317.

[9] Hubert, M., and M., Debruyne (2010). Minimum covariance determinant. WIREs Comp. Stat. 2, 36-43.

[10] Huber, P.J., and E.M., Ronchetti (2009). Robust Statistics 2nd Edition, Wiley & Sons.

[11] Davies, P.L., and U., Gather (2007). The breakdown point – Examples and counterexamples. Stat. Jour. 5(1), 1–17.

[12] Hubert, M., J. Rousseeuw, P.J., and T., Verdonck (2012). A deterministic algorithm for robust location and scatter. Jour. Comput. & Grap. Stat. 21(3), 618–637.

[13] Rousseeuw, P.J. (1984). Least median of squares regression. J. Am. Stat. Assoc. 79, 871–880.

[14] Aelst, S.V., and P.J., Rousseeuw (2009). Minimum volume ellipsoid. WIREs Comp. Stat. 1, 71-82.

[15] Rousseeuw, P.J., and A.M., Leroy (1987). Robust Regression and Outlier Detection. John Wiley and Sons, New York, NY, USA.

[16] Kumar, P., and E.A., Yildirim (2005). Minimum-Volume Enclosing Ellipsoids and Core Sets. Jour. Optim. Theo. & Appl. 126(1), 1-21.

[17] Sun, P., and R.M., Freund (2004). Computation of minimum volume covering ellipsoids, Opera. Rese. 52, 690–706.

[18] Ahipaşaoğlu, S.D. (2014). Fast Algorithms for the Minimum Volume Estimator. Jour.of Glob. Optim. 62(2), 351-370.

[19] Rousseeuw, P. J., and K., Van Driessen (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. Techn. 41, 212-223.

[20] Zuo, Y., and S., Lai (2011). Exact computation of bivariate projection depth and the Stahel–Donoho estimator. Compu. Stat. & Data Analy. 55, 1173–1179.

[21] Van Aelst, S., and E.W., Vandervieren (2011). A Stahel–Donoho estimator based on huberized outlyingness. Compu. Stat. & Data Anal. 56, 531–542.

[22] Gervini, D. (2002). The influence function of the Stahel–Donoho estimator of multivariate location and scatter. Stat. & Prob. Letters 60, 425–435.

[23] Maronna, R.A., and V.J., Yohai (1995). The behavior of the Stahel–Donoho robust multivariate estimator. J. Amer. Statist. Assoc. 90, 329–341.

[24] Debruyne, M., and M., Hubert (2009). The influence function of the Stahel_Donoho covariance estimator of smallest outlyingness. Stat. & Prob. Letters 79, 275_282.

[25] Davies, L., (1987). Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. Annals of Stat. 15, 1269–1292.

[26] Hubert, M., Rousseeuw, P., Vanpaemel, D., and T., Verdonck (2015). The DetS and DetMM estimators for multivariate location and scatter. Comput. Stat. & Data Anal. 81, 64-75.

[27] Salibian-Barrera, M., and V., Yohai (2006). A fast algorithm for S-regression estimates. Jour. of Comput. & Graph. Stat. 15, 414–427.

[28] Aghanabati, A., (2004). "Geology of Iran." Geological Survey of Iran, (586 p.). (In Persian).

[29] Eftekharnejad, J., (1990). "1:250000 Geology Map of Birjand", Geological Survey of Iran. (In Persian).

[30] Abdi, M., Karimpour, M.H. and A. Najafi (2010). "Geology, alteration Geology, alteration and mineralization potential of Kuh-Shah Region, South Khorasan." The First Congress of Economic Geology of Iran. (In Persian).

[31] Abdi, M., and M.H. Karimpour (2012). "Geology, alteration, mineralization, petrogenesis, geochronology, geochemistry and airborne geophysics of Kuh Shah prospecting area, SW Birjand." Journal of Ecomonic Geology 4(1): 77-107. (In Persian).

[32] Roshani Rodsari, P., Mokhtari, A.R., and S.H. Tabatabaei (2012). Investigation on Geochemical Association of Elements in Open and Closed Data System (Case Study: kuh-e Panj Copper Deposit (Kerman)). Journal of Analytical and Numerical Methods in Mining Engineering 2(4), 46-51. (In Persian).