نشریه علمی-پژوهشی "مهندسی معدن"
Iranian Journal of Mining Engineering
(IRJME)

دوره نهم، شماره ۲۵، سال ۱۳۹۳، صفحه ۸۵ تا ۹۶
Vol. 9, No. 25, 2015, pp.85-96

جداسازی داده‌های خارج از رده به روش تک‌متغیره و چند متغیره در داده‌های ژئوشیمی محدوده طلای اپی‌ترمال ساری گونای

صادق کیانپوریان[۱]؛ هوشنگ اسدی هارونی[*۲]؛ سهراب افشاری[۳]؛ مهران فرهمندیان[۳]

۱- کارشناس ارشد گروه ژئوشیمی جهاد دانشگاهی واحد صنعتی اصفهان، s.kianpouryan@ut.ac.ir
۲- استادیار دانشکده مهندسی معدن، دانشگاه صنعتی اصفهان hooshang@cc.iut.ac.ir
۳- عضو هیئت علمی جهاد دانشگاهی واحد صنعتی اصفهان afshari@acecr.ac.ir

(دریافت ۲۸ شهریور ۱۳۹۲ پذیرش ۱۳ بهمن ۱۳۹۳)

چکیده

داده‌های پرت در ژئوشیمی اکتشافی بسیار مهم هستند و می‌توانند اثرات زیادی بر نتایج روش‌های آماری از جمله جداسازی آنومالی از زمینه داشته باشند. بنابراین یکی از اولین مراحل پیش پردازش در تحلیل داده‌های ژئوشیمیایی، تشخیص و تصمیم‌گیری در مورد حذف یا تصحیح آن‌ها است. این داده‌ها را به ۳ روش تک متغیره، دو متغیره و چند متغیره می‌توان شناسایی کرد که هدف از این تحقیق جداسازی آن‌ها با استفاده از روش‌های تک متغیره و چند متغیره است. با این حال، با توجه به این که داده‌های ژئوشیمیایی به صورت ترکیبی هستند قبل از هر تجزیه و تحلیلی باید از یک تبدیل مناسب برای رفع این مشکل استفاده کرد. خانواده تبدیل لگاریتمی ریشه‌ای از مناسب‌ترین تبدیل‌ها برای این کار هستند که از بین آن‌ها تبدیل ریشه‌ای ایزومتریک برای این مطالعه انتخاب شد. پس از اعمال تبدیل ایزومتریک بر روی داده‌ها، روش‌های نمودار جعبه‌ای و فاصله‌ی ماهالانوبیس مقاوم به ترتیب برای تشخیص آن‌ها به روش‌های تک متغیره و چند متغیره انتخاب شدند. در روش تک متغیره از چارک سوم به اضافه یک و نیم برابر فاصله‌ی چارکی به عنوان حد آستانه و در روش چند متغیره از یک حد آستانه‌ی تصحیح شده بر اساس تابع توزیع تجربی مربع فاصله ماهالانوبیس مقاوم و تابع توزیع مربع کای برای این کار استفاده شده است. نتایج این مطالعه نشان می‌دهد که با توجه به ماهیت چند متغیره بودن داده‌های ژئوشیمی اکتشافی، استفاده از روش‌های چند متغیره برای تشخیص داده‌های پرت با دقت بیشتری همراه خواهد بود.

کلمات کلیدی

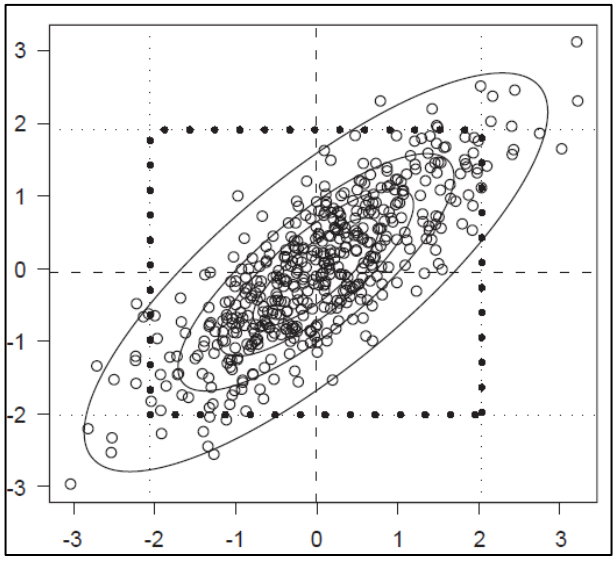
داده‌های پرت، داده‌های ترکیبی، تبدیل لگاریتمی ریشه‌ای ایزومتریک، نمودار جعبه‌ای، فاصله ماهالانوبیس مقاوم

۱- مقدمه

تشخیص داده‌های خارج از رده و ساختارهای داده‌ای غیرمعمول، یکی از مهم‌ترین پیش پردازش‌ها در آنالیز آماری داده‌های ژئوشیمیایی است. روش‌های تشخیص تک متغیره، دو متغیره و چند متغیره را می‌توان برای جداسازی داده‌های پرت به کار برد [۱]. در روش‌های تک متغیره توزیع مشاهدات بررسی شده و داده های خارج از یک حد آستانه به عنوان پرت تلقی می‌شوند. با این حال، کاوش برای شناخت داده‌های خارج از رده به صورت دو و چند متغیره معمولاً بر اساس مکان و توسعه آن‌ها است. بدین ترتیب که هر چه یک نمونه از مرکز داده‌ها فاصله بیشتری داشته باشد، پتانسیل بیشتری برای پرت بودن دارد. به عبارت دیگر، داده‌های خارج از رده فاصله زیادی از مرکز کل مشاهدات دارند [۲]. تعریف یک حد یا حد آستانه برای جداسازی نمونه-های پرت از سایر نمونه‌ها، در کارهای ژئوشیمی بسیار مورد توجه قرار گرفته است. با این وجود، یک روش کاربردی عمومی برای تعیین این حدآستانه تاکنون معرفی نشده است.

داده‌های پرت در ژئوشیمی علاوه بر اشتباه یا خطا، اغلب ناشی از فرآیندهای کانی‌سازی، آلتراسیون و فعالیت‌های انسانی است [۳،۴]. اشتباه یا خطا می‌تواند ناشی از نمونه‌برداری و آماده‌سازی نادرست، خطای روش‌های اندازه‌گیری و اشتباه وارد کردن داده‌ها باشد. در جاهایی که کانی‌سازی اتفاق می‌افتد غلظت بعضی از عناصر زیاد بوده و این امر باعث ایجاد چولگی مثبت در توزیع احتمال داده‌ها می شود، آلتراسیون سبب کاهش غلظت بعضی از عناصر می‌شود. از سوی دیگر آلودگی زیست محیطی ناشی از فعالیت‌های انسانی می‌تواند سبب افزایش غلظت بعضی از عناصر شود.

در روش‌های تک‌متغیره هر یک از متغیرها به صورت جداگانه بررسی شده و نمونه‌های خارج از رده آن‌ها شناسایی می‌شود. در این روش‌ها معمولاً دامنه توزیع مشاهدات بررسی شده و داده‌های خارج از یک دامنه معین به عنوان پرت تلقی می‌شوند. از مهم‌ترین کارهای انجام شده در این زمینه می‌توان به توکی (۱۹۷۷)، ژانگ و همکاران (۱۹۹۹)، لالور و ژانگ (۲۰۰۱)، چیانگ و همکاران (۲۰۰۳) و ریمن و همکاران (۲۰۰۵) اشاره کرد. اما در آنالیزهای چندمتغیره نه تنها فاصله‌ی یک مشاهده از مرکز داده‌ها بلکه شکل داده‌ها نیز باید در نظر گرفته شود. برای شرح این موضوع، دو متغیر با توضیح نرمال که دارای همبستگی مشخصی هستند توسط فیلزموزر و همکارانش (۲۰۰۵) شبیه‌سازی شده است (شکل ۱). مکان تخمینی مرکزی هر کدام از متغیرها با خط تیره نشان داده شده است. یکی از حد آستانه‌هایی که ژئوشیمیدانان برای جداسازی داده‌های پرت به صورت تک متغیره استفاده می‌کردند، مشخص کردن ۲ درصد بالایی و پایینی داده‌ها به عنوان مقادیر خارج از رده بود. بنابراین اگر مشاهداتی که در این بخش از داده‌ها قرار می‌گیرند را به عنوان خارج از رده در نظر بگیریم، مستطیل مشخص شده با نقاط تیره پر رنگ در شکل ۱ داده‌های پرت را از غیر پرت جدا می‌کند. همانطور که از شکل پیداست، این فرایند بیضوی بودن داده‌های دومتغیره را در نظر نمی‌گیرد و در نتیجه نمی‌تواند روش موثری باشد.



شکل ۱: شبیه سازی داده‌های با توزیع نرمال استاندارد با همبستگی از پیش تعریف شده. خطوط تیره میانگین متغیرها، بیضوی‌ها ۲۵، ۵۰، ۷۵ و ۹۸ درصد توزیع مربع کای و نقاط پررنگ برابر ۲ و ۹۸ درصد برای هر متغیر هستند [۲].

شکل و اندازه داده‌های چند متغیره با استفاده از ماتریس کواریانس مشخص می‌شود. یکی از مهم‌ترین معیارهایی که ماتریس کواریانس را در نظر می‌گیرد به فاصله ماهالانوبیس مشهور است. برای یک نمونه چند متغیره p بعدی، فاصله ماهالانوبیس برای مشاهده i ام از رابطه زیر بدست می‌آید:

$$MD_i = \left((x_i - t)^T C^{-1}(x_i - t)\right)^{1/2} \; for\; i = 1, \dots, n \quad (1)$$

که در اینجا، x_i بردار متغیرها برای مشاهده i ام، t بردار میانگین متغیرها (مرکز ثقل مشاهدات) و C ماتریس کواریانس نمونه-هاست. برای داده‌های نرمال چند متغیره مربع فاصله ماهالانوبیس هم ارز با توزیع مربع کای با p درجه آزادی است (χ_p^2). بنابراین با قرار دادن مربع فاصله ماهالانوبیس برابر یک مقدار ثابت (برای مثال مقداری از χ_p^2) می‌توان بیضوی‌هایی تعریف کرد که دارای فاصله ماهالانوبیس یکسان از مرکز مشاهدات هستند. شکل ۱ این موضوع را برای داده‌های دومتغیره با توضیح نرمال نشان داده است. در این شکل بیضوی-ها برابر ۲۵، ۵۰، ۷۵ و ۹۸ درصد توزیع مربع کای هستند. بنابراین نقاط درون هر بیضی دارای فاصله یکسان از مرکز ثقل داده‌ها هستند. در گذشته گارت (۱۹۸۹) فیلزموزر و همکاران (۲۰۰۵) و فیلزموزر و هارون (۲۰۰۸) از فاصله‌ی ماهالانوبیس برای شناسایی داده‌های پرت به صورت چند متغیره استفاده کرده‌اند.

بنابراین داده‌های خارج از رده را می‌توان مشاهداتی در نظر گرفت که دارای فاصله ماهالانوبیس بزرگی هستند. با این توضیحات، با مشخص کردن یک حد آستانه برای این فاصله (برای مثال ۹۸ درصد توزیع مربع کای)، داده‌های پرت را به آسانی می‌توان جدا کرد. با این وجود پارامترهای فاصله ماهالانوبیس (میانگین متغیرها و ماتریس کواریانس) خود متاثر ار داده‌های پرت هستند و در نتیجه فاصله یاد شده نیز تحت تاثیر داده‌های پرت خواهد بود. برای حل این مسئله باید از فاصله ماهالانوبیس مقاوم استفاده کرد که به داده‌های پرت حساس نیست. یکی از نکات مهم دیگری که باید در نظر گرفته شود این است که قبل از مشخص کردن داده‌های پرت باید از تبدیلی برای برطرف کردن مشکل بسته بودن سیستم داده‌های ژئوشیمیایی استفاده کرد. هدف از این تحقیق، در ابتدا بررسی روش‌های باز کردن سیستم‌های عددی بسته داده‌های ژئوشیمیایی و جداسازی داده‌های پرت به روش تک متغیره و سپس شناسایی داده‌های خارج از رده با استفاده از روش چند متغیره فاصله ماهالانوبیس مقاوم و مقایسه کیفی آن‌ها با روش-های تک متغیره است.

۲- پارامترهای آماری داده‌های خام

در جدول (۱) پارامترهای آماری داده‌های خام حاصل از آنالیز نمونه‌های برداشت شده از محدوده مورد مطالعه نشان داده شده است. با توجه به مقدار کمینه و بیشینه عناصر می‌توان نتیجه گرفت که محدوده داده‌ها بسیار وسیع می‌باشد و در نتیجه آن، احتمال وجود داده‌های خارج از رده در این مجموعه داده بسیار بالا است. علاوه بر این، این داده‌ها از توزیع نرمال پیروی نمی-کنند، بنابراین باید تبدیلی بر روی آن‌ها انجام داد که بتواند آن‌ها را به نرمال نزدیک کند. در ادامه تبدیل‌های متفاوتی برای نرمال کردن و همچنین باز کردن سیستم عددی این داده‌ها معرفی شده است.

جدول ۱: پارامترهای آماری داده‌های خام محدوده مورد مطالعه.

	میانگین	میانه	انحراف معیار	کمینه	بیشینه
Au(ppb)	۳۲۲/۵۶	۷۸/۵	۶۶۸/۱۴	۲	۶۰۵۸
Ag(ppm)	۲/۹	۰/۵۲۵	۹/۹۵	۰/۵	۱۱۰/۵۵
As(ppm)	۵۳۳/۹۳	۳۳۸/۵	۶۵۱/۲۱	۳۷	۶۷۲۸
Hg(ppm)	۲/۴۴	۰/۵	۱۰/۶۵	۰/۵	۱۹۴
S(%)	۰/۰۷۸	۰/۰۴	۰/۰۸۹	۰/۰۰۵	۰/۸۸
Sb(ppm)	۲۷۲/۰۷	۱۰۶	۷۷۴/۰۶	۲/۵	۱۱۱۱۵
Tl(ppm)	۲/۶۸	۲/۵	۱/۹۸	۱/۳	۲۵

۳- داده‌های ترکیبی

داده‌های ترکیبی، داده‌های بسته حاوی اطلاعات نسبی می‌باشند که حاصل جمع این گونه داده‌ها ثابت می‌باشد (به عنوان مثال ۱۰۰٪). در بیشتر حالات، این داده‌ها را داده‌های بسته می‌نامند زیرا دارای حاصل جمع ثابت هستند. یک مثال کلاسیک برای

آرایه بسته یا یک سیستم عددی بسته، مجموعه‌ای از داده‌هاست که متغیرهای آن مستقل از یکدیگر نمی‌باشند و به صورت درصد یا قسمت در میلیون بیان می‌شوند [۱۰]. در گذشته، مجموعه داده‌های با حاصل جمع ثابت را داده‌های ترکیبی می‌نامیدند اما در حال حاضر این داده‌ها دارای تعریف وسیع‌تری هستند و مجموعه داده‌هایی که دارای حاصل جمع ثابت نیز نمی‌باشند را شامل می‌گردد. در تعریف جدید، این داده‌ها بخشی از کل هستند که فقط دارای اطلاعات نسبی هستند. فضای اقلیدسی برای داده‌های ترکیبی مناسب نمی‌باشند و محدودیت حاصل جمع ثابت این داده‌ها دلالت بر هندسه خاصی را دارد که در اصطلاح هندسه اتکینسون در محیط ساده شده نامیده می‌شود [۱۱،۱۰].

داده‌های ترکیبی دارای خواص مهم و خاصی هستند که سبب شده نتوان از روش‌های آماری استاندارد استفاده نمود. روش‌های آماری استاندارد برای استفاده جهت داده‌های آزاد که در بازه منفی تا مثبت بی‌نهایت تغییر می‌نمایند، طراحی شده‌اند [۱۲]. داده‌های ترکیبی همیشه مثبت می‌باشند و هنگامی که به شکل بسته هستند فقط در بازه ۰ تا ۱۰۰ یا هر ثابت دیگری تغییر می‌کنند. روش‌های مختلفی برای تبدیل داده‌های ترکیبی معرفی شده‌اند که تبدیلات خانواده لگاریتم ریشه‌ای معروف‌ترین آن‌ها هستند [۱۰]. ۳ تبدیل لگاریتم ریشه‌ای برای باز کردن سیستم‌های عددی بسته معرفی شده‌اند. اولین نوع از آن‌ها تبدیل لگاریتمی ریشه‌ای جمع‌پذیر است که توسط اتکینسون (۱۹۸۶) معرفی شد. برای یک نمونه ترکیبی x با D بعد این تبدیل به صورت زیر تعریف می‌شود:

$$alr(x) = \left(ln\frac{x_1}{x_j}, \dots, ln\frac{x_{j-1}}{x_j}, ln\frac{x_{j+1}}{x_j}, \dots, ln\frac{x_D}{x_j} \right) \quad (2)$$

در اینجا j یکی از متغیرهای مجموعه D بعدی است. با توجه به این رابطه، یکی از متغیرها با اندیس j باید به عنوان تقسیم کننده یا مقسوم علیه انتخاب شود. بنابراین یکی از مشکلات این روش انتخاب متغیر تقسیم کننده است، به عبارت دیگر این تبدیل وابسته به طرز تفکر شخص است. زیرا با انتخاب متغیرهای مختلف، نتایج متفاوتی به دست می‌آید. برای حل این مشکل تبدیل لگاریتمی ریشه‌ای مرکزی توسط اتکینسون (۱۹۸۶) برای نمونه ترکیبی D بعدی x به صورت زیر تعریف شد:

$$\mathrm{clr}(x) = (y_1, \dots, y_D) = \left(\ln\frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \dots, \ln\frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right) \quad (3)$$

در تبدیل لگاریتم ریشه‌ای مرکزی، هر کدام از متغیرها بر میانگین هندسی کل متغیرها تقسیم می‌شوند، بنابراین تفسیر نتایج حاصل از این تبدیل آسان است. با این وجود، با توجه به رابطه ۳ می‌توان نتیجه گرفت که $\sum_{i=1}^{D} y_i = 0$. بنابراین ماتریس خروجی حاصل از داده‌ها یک ماتریس منفرد خواهد بود و برای بسیاری از کارهای چند متغیره نمی‌تواند مورد استفاده قرار گیرد. برای حل این مشکل، ایگوزکو و همکاران (۲۰۰۳) تبدیل لگاریتم ریشه‌ای ایزومتریک را پیشنهاد دادند که به صورت زیر است:

$$z = (z_1, \dots, z_D)',\ z_i = \sqrt{\frac{D-i}{D-i+1}} \ln\frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \text{for i} = 1, \dots, D-1 \quad (4)$$

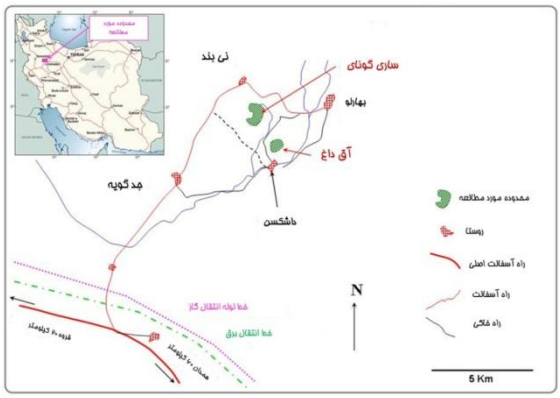
خروجی حاصل از این تبدیل به صورت منفرد نیست، اما به دلیل کاهش بعد داده‌ها تفسیر آن‌ها را تا حدی غیرممکن می‌سازد. بنابراین، در مواردی که نیاز هست نتایج به صورت گرافیکی و غیره برای متغیرها تفسیر شوند معمولا داده‌های تبدیل یافته به ایزومتریک را با تبدیل ریشه‌ای مرکزی به ابعاد داده‌های اصلی برمی‌گردانند [۹].

داده‌های این مطالعه مربوط به کانی‌سازی طلای اپی‌ترمال ساری گونای می‌باشد. به منظور بررسی ژئوشیمیایی در محدوده ساری‌گونای، اقدام به نمونه‌برداری از محیط خاک در منطقه بر روی شبکه منظم شد. نمونه‌برداری در منطقه‌ای به مساحت تقریبی چهار کیلومترمربع انجام شد که در مجموع ۱۲۰۰ نمونه خاک برداشت شد. چگالی شبکه نمونه‌برداری با توجه به شواهد کانی‌سازی در سطح متغیر است و نمونه‌برداری با شبکه‌ای به ابعاد ۱۰۰×۱۰۰ و ۲۵×۱۰۰ متر انجام شده است. آنالیز نمونه‌ها برای تعیین غلظت ۴۷ عنصر انجام شد اما نتایج عناصر

Au، As، Sb، S، Hg، Tl و Mo (که ارتباط نزدیکتری با کانی-سازی داشتند) در محدوده‌ی تپه ساری گونای (شامل ۶۲۰ نمونه) برای این مطالعه مورد استفاده قرار گرفتند. با توجه به توضیحات داده شده، قبل از انجام آنالیزهای مورد نیاز برای شناختن داده‌های پرت، تبدیل ایزومتریک بر روی داده‌های مورد مطالعه انجام شد و نتایج حاصل از آن‌ها برای ادامه کار مورد استفاده قرار گرفت.

۴- موقعیت جغرافیایی و زمین‌شناسی محدوده مورد مطالعه

محدوده طلای ساری‌گونای در شمال‌غربی ایران، جنوب‌شرقی استان کردستان و ۶۰ کیلومتری شهر همدان واقع شده است. فعالیت‌های اکتشافی در منطقه بر روی دو تپه با نام‌های ساری-گونای و آق‌داغ که به فاصله یک کیلومتر از یکدیگر قرار گرفته‌اند، متمرکز شده است. بیشترین ارتفاع در منطقه در حدود ۲۲۰۰ متر از سطح دریا می‌باشد. به لحاظ آب و هوایی این منطقه نیمه خشک، همراه با ریزش باران به خصوص در فصل بهار و برف زودرس در زمستان می‌باشد. منطقه مورد مطالعه با مساحتی حدود ۱/۶ کیلومترمربع، تپه ساری‌گونای را در بر می‌گیرد. در شکل (۲) موقعیت محدوده مورد مطالعه نشان داده شده است.

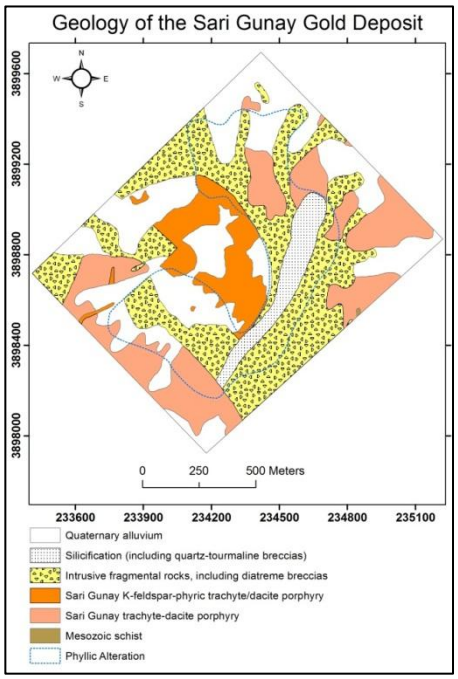


شکل ۲- موقعیت جغرافیایی منطقه مورد مطالعه.

از لحاظ زمین‌شناسی محدوده ساری‌گونای در جنوب شرقی کمربند قروه- بیجار (ولکانیک‌های با روند شمال غربی- جنوب شرقی) واقع شده است. کمربند قروه- بیجار بین دو کمربند اصلی سنندج- سیرجان و ارومیه- دختر واقع است. این کمربند هم‌راستا با دو کمربند اصلی و البته به نسبت کوچک‌تر از آن‌ها می‌باشد. ولکانیک‌های موجود در این کمربند اغلب از آندزیت و داسیت‌های ائوسن تا میوسن و بازالت‌های پلیستوسن می‌باشند. مجموعه‌ای از ساختارهای بزرگ با روند شمال غربی- جنوب شرقی و مجموعه‌ای از ساختارهای کوچک متقاطع با روند جنوب غربی- شمال شرقی در این کمربند مشاهده می‌شود. تقاطع این دو مجموعه ساختاری را می‌توان مؤثر بر توسعه ولکانیزم و کانی-سازی در منطقه دانست [۱۵].

نقشه زمین‌شناسی محدوده اکتشافی ساری گونای در شکل (۳) نشان داده شده است. آندزیت پورفیری به طور غالب در شمال سیستم ساری‌گونای وجود دارد. این واحد اغلب به صورت آندزیت پورفیری دگرسان نشده یا دگرسان شده ضعیف وجود دارد. داسیت پورفیری یکی از فراوان‌ترین و احتمالاً قدیمی‌ترین واحد سنگی ولکانیکی در ساری‌گونای می‌باشد. این واحد بیشتر در دامنه جنوب شرقی تپه ساری‌گونای وجود دارد، البته در دیگر بخش‌های تپه ساری‌گونای نیز مشاهده می‌شود. تغییرات اندک میزان فنوکریست‌های کوارتز در این واحد سنگی که حدود ۱۰ درصد حجمی گزارش شده است، منجر به تشخیص واحد سنگی دیگری شده است. فنوکریست‌های کوارتز موجود به نام کوارتزهای چشمی نام‌گذاری شده‌اند و این واحد سنگی که با نام (کوارتز – داسیت پورفیری) شناخته می‌شود، در دامنه جنوب شرقی ساری‌گونای دیده می‌شود. نوع دیگر سنگ در این گروه، داسیت – آندزیت پورفیری می‌باشد که به دلیل داشتن مقدار برابر کانی بیوتیت و هورنبلند با واحد داسیت پورفیری تفاوت دارد. دو توده حلقه‌ای شکل برش‌ها و توف‌های دودکشی/دیاترمی در دو ناحیه، داسیت‌های پورفیری را قطع کرده‌اند. واحدی که شامل خرده‌سنگ‌های قدیمی و نادر در منطقه، و همچنین کریستال‌های شکسته و گرد شده پلاژیوکلاز و بیوتیت باشد با نام توف کریستالی داسیتی معرفی می‌شود. وقتی که واحد توف کریستالی شامل مقدار مشهود (نه به طور غالب) از خرده سنگ‌های معلق در ماتریکس کریستالی باشد، با نام توف

لیتیک داسیتی معرفی می‌شود. این دو واحد سنگی بخش اعظم توده حلقه‌ای شکل دودکش توف/پیروکلاستیک در ساری گونای را تشکیل می‌دهند. اگر سنگ‌های خرد شده دارای مقداری قابل توجه از خرده سنگ‌ها و بافت برشی باشند، به نام توف برشی لیتیکی خوانده می‌شوند. واحد سنگی برش هیدروترمالی هر دو واحد سنگی داسیت پورفیری و توف‌های آتشفشانی / دیاترمی را قطع می‌کند؛ بنابراین به لحاظ زمانی پس از آن دو تشکیل شده است [۱۵].

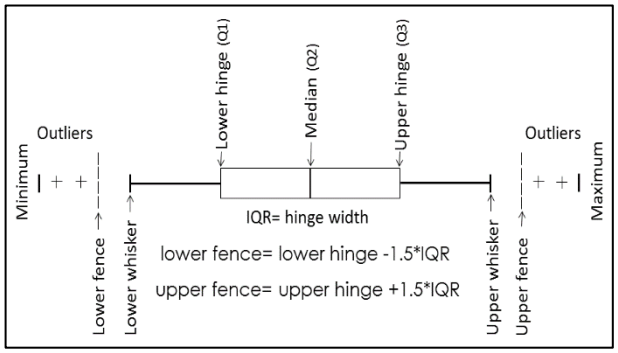


شکل ۳: نقشه زمین شناسی کانسار طلای اپی ترمال ساری گونای.

۵- تشخیص داده‌های پرت به روش تک متغیره

روش های مختلفی برای بررسی یک متغیره داده‌های پرت وجود دارند که می توان آن‌ها را به ۲ گروه دامنه و آزمون‌های آماری تقسیم کرد. در روش‌های دامنه توزیع مشاهدات بررسی شده و داده های خارج از یک دامنه معین به عنوان داده پرت تلقی می-شوند. مهم‌ترین موضوع در این ارتباط تعیین دامنه یاد شده برای مشخص کردن داده‌های پرت است. در روش های سنتی داده-های بزرگ تر از میانگین به اضافه ۳ برابر انحراف معیار و کوچک تر از میانگین منهای ۳ برابر انحراف معیار خارج از رده محسوب می‌شدند (هایر و همکاران، ۱۹۹۸). چون این روش تحت تاثیر داده‌های پرت است از این رو روش‌های دیگری از جمله میانه به اضافه منهای انحراف مطلق از میانه (MAD) و نمودار جعبه ای معرفی شدند که تحت تاثیر داده های پرت قرار نمی‌گیرند. ریمن و همکاران (۲۰۰۵) نشان دادند که از بین این دو روش، نمودار جعبه ای در مواردی که درصد داده های پرت کمتر از ۱۰ درصد و روش دیگر در مواردی که درصد این داده ها بیشتر از ۱۵ درصد باشد بیشترین کارایی را دارند. بنابراین در این تحقیق از نمودار جعبه ای برای شناسایی داده‌های پرت استفاده شده است.

نمودار جعبه‌ای مقادیر داده‌ها را به ۴ قسمت مساوی (با توجه به میانه داده‌ها) تقسیم می‌کند (شکل ۴). این نمودار ۵ آماره شامل مینیمم، چارک اول (Q_1)، چارک دوم یا میانه، چارک سوم (Q_3) و ماکزیمم را نشان می‌دهد. طول مستطیل در این نمودار را که اختلاف بین چارک های اول و سوم است، فاصله‌ی چارکی می‌نامند (IQR) (توکی، ۱۹۷۷). در نمودار جعبه‌ای که از آن برای تشخیص داده های پرت استفاده می‌شود، داده‌های بزرگ‌تر از Q_3+1.5IQR و کوچک‌تر از Q_1-1.5IQR به عنوان داده‌های پرت محسوب می‌شوند. در این مطالعه از این معادله‌ها برای تعیین داده‌های پرت استفاده شده است.



شکل ۴: نمودار جعبه ای توکی و پارامترهای مربوط به آن.

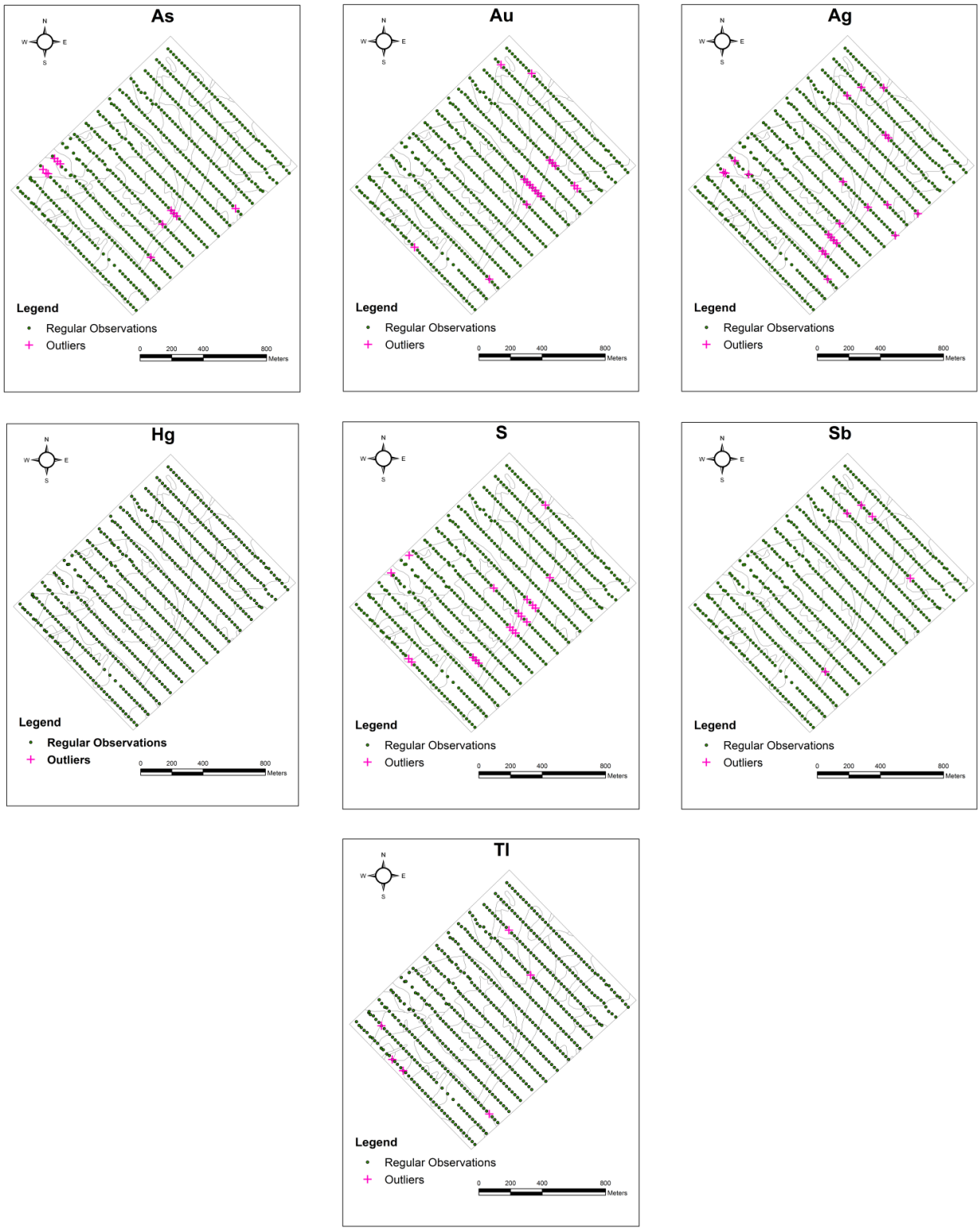
شکل ۵ نتایج حاصل از اعمال روش نمودار جعبه‌ای بر روی متغیرهای مورد بررسی جهت شناسایی داده‌های پرت را نشان می‌دهد. در این شکل داده‌های خارج از رده با علامت + صورتی رنگ مشخص شده‌اند. بیشترین داده‌های پرت مربوط به عنصر گوگرد (۲۲ نمونه) و کمترین مربوط به عنصر جیوه (بدون نمونه پرت) است. یکی از نکات مورد بحث در نمودارهای مربوط به این

متغیرها، ارتباط ضعیف نمونه‌های به‌دست آمده است. به عبارت دیگر، نمونه‌های خارج از رده در هر کدام از عناصر متفاوت از این نتیجه نشان دهنده‌ی این موضوع است که روش‌های شناسایی داده‌های پرت به روش تک متغیره تنها در مواردی مناسب هستند که سایر تجزیه و تحلیل‌ها نیز به صورت تک متغیره انجام شود. یکی دیگر از نکات جالب توجه در این نمودارها این است که نمونه‌های پرت مربوط به عناصر گوگرد و نقره، طلا و ارسنیک در حوالی تپه‌ی ساری گونای (مکان اصلی کانی‌سازی طلای اپی ترمال در منطقه) قرار دارند که با توجه به ماهیت زمین‌شناسی این محدوده می‌تواند درست باشد. به عبارت دیگر، این احتمال وجود دارد که نمونه‌برداری از این محدوده به صورت تصادفی برداشته نشده باشد.

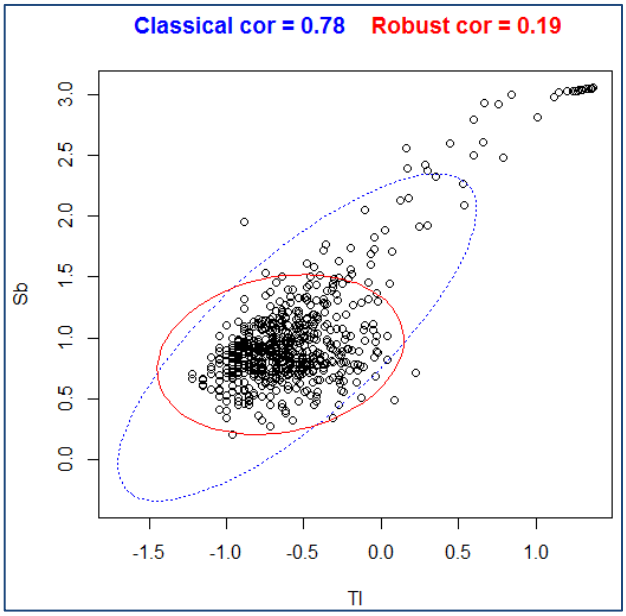
۶- فاصله ماهالانوبیس مقاوم

همان‌طور که در مقدمه قبل هم گفته شد، فاصله ماهالانوبیس نسبت به داده‌های پرت حساس است. بنابراین برای محاسبه‌ی پارامترهای این فاصله (میانگین و ماتریس کواریانس) باید از یک روش مقاوم استفاده کرد تا اثر داده‌های خارج از رده را حذف کند. تاکنون تعداد زیادی تخمین‌گر مقاوم معرفی شده‌اند که در بین آن‌ها "کمترین دترمینان ماتریس کواریانس" به دلیل سرعت بالای الگوریتم آن بیشتر از سایر موارد مورد استفاده قرار گرفته است. در این روش تعداد h مشاهده (معمولاً برابر ۷۵ درصد تعداد کل مشاهدات) را طوری پیدا می‌کنند که ماتریس کواریانس آن دارای کم‌ترین دترمینان باشد. در ادامه بردار میانگین و ماتریس کواریانس متغیرها از h مشاهده یاد شده برآورد شده و در نتیجه فاصله ماهالانوبیس حاصل حساسیت کمتری نسبت به داده‌های پرت خواهد داشت [۱۶].

در شکل ۶ نمودار Tl و Sb نشان داده شده که میانگین و ماتریس کواریانس آن‌ها با روش غیرمقاوم محاسبه شده و با توجه به حد آستانه ۹۸ درصد توزیع مربع کای بیضوی مربوطه رسم شده است. نمونه‌هایی که خارج از بیضوی قرار گرفته‌اند عضوهای خارج از رده این جامعه‌ی دو متغیره هستند. با محاسبه میانگین و ماتریس کواریانس با روش مقاوم و همچنین تعیین حد آستانه بیضوی پر رنگ در این شکل رسم شده و نمونه‌های پرت مشخص شده‌اند. همان‌طور که از شکل مشخص است وقتی پارامترهای فاصله ماهالانوبیس با روش مقاوم محاسبه شده‌اند تعداد مشاهدات بیشتری به عنوان خارج از رده خود را نشان داده‌اند. علاوه بر این در حالی که همبستگی پیرسون در حالت اول برابر ۰/۷۸ است این همبستگی بر اساس روش کمترین دترمینان ماتریس کواریانس برابر ۰/۱۹ است. مقدار همبستگی بالا بر اساس روش اول به این دلیل است که تعداد کمی از نمونه‌های مربوط به هر دو عنصر دارای غلظت غیرمعمول بالایی هستند.



شکل ۵: نقشه‌های تک متغیره نشان دهنده‌ی نمونه‌های پرت و نمونه‌های عادی، شناسایی شده با استفاده از نمودار جعبه‌ای



شکل ۶: نمودار پراکندگی عناصر Sb و Tl پس از تبدیل لگاریتمی ایزومتریک همراه با کواریانس محاسبه شده با استفاده از روش معمولی و کمترین دترمینان ماتریس کواریانس

۷- داده‌های پرت و داده‌های آنومال

داده‌های خارج از رده مشاهداتی هستند که مربوط به یک یا چند توزیع متفاوت هستند، در حالی که داده‌های آنومال اگر چه فاصله زیادی از مرکز مشاهدات دارند ولی متعلق به توزیع یکسان هستند. بنابراین در پردازش‌های آماری باید این دو نوع جامعه از هم تفکیک شوند. در پاراگراف‌های قبل مقدار حد آستانه ثابت (۹۸٪ توزیع مربع کای) برای جداسازی داده‌های پرت معرفی شد ولی به نظر می‌رسد که حد آستانه ثابت به دلایلی از واقعیت کمی دور باشد. اولین دلیل این است که اگر داده‌ها واقعاً متعلق به یک جامعه نرمال باشند آنگاه به دلیل اینکه هیچ نمونه‌ای با توزیع متفاوتی وجود ندارد بنابراین حد آستانه بی‌نهایت خواهد بود. دلیل دوم این است که هیچ توضیح قانع کننده‌ای که چرا حد آستانه ثابت برای همه داده‌ها مناسب است، وجود ندارد. و سوم اینکه حد آستانه باید با توجه به تعداد نمونه‌ها قابل تغییر باشد [۲].

بهترین حد آستانه‌ای که می‌توان برای تعیین داده‌های خارج از ردیف انتخاب کرد حدی است که با توجه به اندازه داده‌ها تغییر پیدا کند [۲]. گرت (۱۹۸۹) از نمودار مربع کای برای نیل به این هدف استفاده کرد (با پلات کردن مربع فاصله‌ی ماهالانوبیس در مقابل درصدهای مختلف این توزیع). در این صورت نمونه‌های با مقادیر بالا حذف می‌شدند تا جایی که سایر نمونه‌ها از یک خط راست پیروی می‌کردند. با انجام این روش، نمونه‌های حذف شده را به عنوان مشاهدات پرت در نظر می‌گرفتند و حد آستانه برابر پایین‌ترین مقدار نمونه‌های حذف شده در نظر گرفته می‌شد. مشکل این روش این بود که به صورت دستی باید انجام می‌گرفت. بنابراین مدت زمان زیادی برای انجام این آنالیز احتیاج بود (بویژه وقتی تعداد نمونه‌ها بالا بود). در ادامه روشی بهینه‌تر که توسط فیلزموزر و همکارانش (۲۰۰۵) معرفی شد، شرح داده خواهد شد.

۸- حد آستانه تصحیح شده

نمودار مربع کای برای تصویر کردن انحراف توزیع داده‌ها از نرمال بسیار مفید است. در ادامه از این اصل برای تعیین حد آستانه استفاده شده است. اگر $G_n(u)$ نشان دهنده‌ی تابع توزیع تجربی مربع فاصله ماهالانوبیس مقاوم و $G(u)$ تابع توزیع مربع کای با p درجه آزادی باشد، آنگاه برای داده‌های چند متغیره با توزیع نرمال G_n به سمت G همگرا می‌شود. بنابراین با مقایسه دنباله‌های G_n و G می‌توان داده‌های پرت را جدا کرد (فیلزموزر و همکاران، ۲۰۰۵). دنباله‌ها با $\delta=\chi^2_{p;1-\alpha}$ برای یک α مشخص کوچک (برای مثال ۰/۰۲) به صورت زیر تعریف می‌شود:

$$p_n(\delta) = \sup\left(G(u) - G_n(u)\right)^+ \quad u \geq \delta \qquad (5)$$

در اینجا علامت "+" اختلاف مثبت را نشان می‌دهد و $p_n(\delta)$ انحراف توزیع تجربی از توزیع نظری را فقط در دنباله‌ها اندازه‌گیری می‌کند که می‌تواند به عنوان معیاری برای داده‌های پرت در نظر گرفته شود [۲]. جروینی (۲۰۰۳) از این ایده به عنوان یک مرحله‌ی وزن‌دهی برای محاسبه میانگین و ماتریس کواریانس چند متغیره به روش مقاوم استفاده کرد. در این معادله $p_n(\delta)$ به طور مستقیم برای تفکیک داده‌های پرت استفاده نمی‌شود، بلکه همان‌طور که گفته شد باید معیاری برای تفکیک داده‌های آنومال از داده‌های پرت تعریف شود. بنابراین، یک مقدار

بحرانی (p_{crit}) معرفی شد تا به جداسازی این دو نوع داده کمک کند. سپس اندازه‌گیری داده‌های پرت به صورت زیر تعریف شد [۲]:

$$\alpha_n(\delta) = \begin{cases} 0 & if\ \ p_n(\delta) \le p_{crit}(\delta, n, p) \\ p_n(\delta) & if\ \ p_n(\delta) > p_{crit}(\delta, n, p) \end{cases} \quad (6)$$

در نهایت حد آستانه تصحیح شده به صورت زیر تعریف می‌شود:

$$c_n(\delta) = G_n^{-1}(1 - \alpha_n(\delta)) \quad (7)$$

مقدار بحرانی برای تشخیص داده‌های آنومال و پرت هم با استفاده از شبیه‌سازی توسط فیلزموزر و همکاران (۲۰۰۵) به طور تقریبی به صورت زیر تعریف شد:

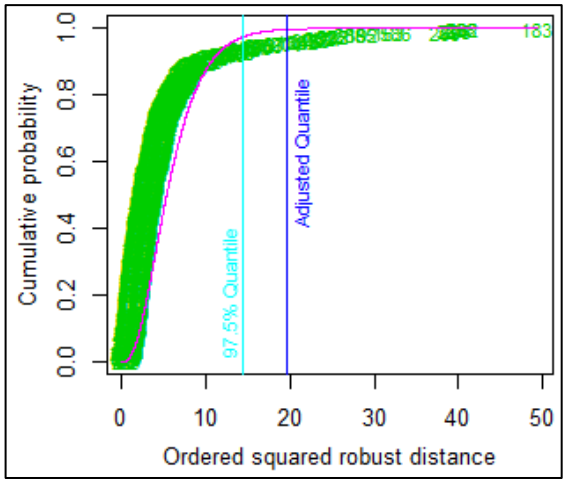
$$p_{crit}(\delta, n, p) = \frac{0.24 - 0.003p}{\sqrt{n}} \quad for \quad \delta = \chi^2_{p,0.975} \text{ and } p \le 10 \quad (8)$$

و

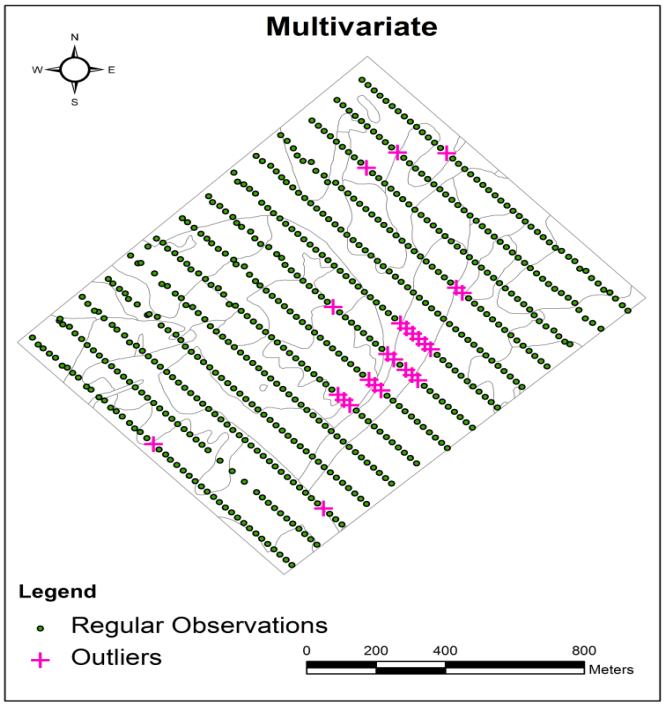
$$p_{crit}(\delta, n, p) = \frac{0.252 - 0.0018p}{\sqrt{n}} \quad for \quad \delta = \chi^2_{p,0.975} \text{ and } p > 10 \quad (9)$$

در شکل ۷ فرایند شناسایی داده‌های خارج از رده با استفاده از حد آستانه تصحیح شده نشان داده شده است. مربع فاصله‌ی ماهالانوبیس مقاوم محاسبه شده و تابع توزیع تجربی آن‌ها با عددهای سبز رنگ نشان داده شده‌اند، منحنی صورتی رنگ نیز تابع توزیع مربع کای با ۶ درجه آزادی است. بر طبق معادله‌ی ۸ مقدار بحرانی برابر ۰/۰۰۹ و با توجه به معادله ۷ مقدار حد آستانه تصحیح شده نهایی برای تشخیص نمونه‌های پرت برابر ۱۹/۱۸ بدست آمد که با خط آبی رنگ در نمودار نشان داده شده است.

برای داشتن دید بهتر نسبت به محدوده‌هایی که دارای نمونه‌های خارج از رده هستند، نمونه‌ها بر روی نقشه زمین-شناسی محدوده نشان داده شده‌اند (شکل ۸). در این نقشه نمادین، نمونه‌های خارج از رده با رنگ صورتی نشان داده شده‌اند که بیشترین تمرکز آن‌ها در تپه ساری گونای (محل اصلی کانی-سازی) است. با این حال در قسمت شمال و جنوب محدوده‌ی مورد مطالعه نیز نمونه‌هایی به عنوان خارج از رده شناسایی شده‌اند که تقریباً سنگ میزبان تمامی آن‌ها رسوبات عهد حاضر هستند و برای مشخص شدن صحت آن‌ها نیاز به بررسی‌های بیشتری وجود دارد. با مقایسه این نقشه با نقشه‌های مربوط به عناصر تک متغیره می‌توان مشاهده کرد که در نقاطی که چند عنصر خارج از رده نشان داده‌اند، در نقشه‌ی چند متغیره نیز نقاط مربوط به آن‌ها به عنوان خارج از رده شناسایی شده است. با این وجود نقشه چند متغیره داده‌های پرت بدست آمده از روش فاصله‌ی ماهالانوبیس همپوشانی بالایی با بیشتر نقشه‌های تک متغیره ندارد. بنابراین، از آنجا که ماهیت داده‌های ژئوشیمیایی به صورت چند متغیره است در مواردی که نیاز به تحلیل‌های چند متغیره است (برای مثال تحلیل مولفه‌های اصلی یا آنالیزهای طبقه‌بندی) باید از روش‌های چند متغیره به ویژه روش استفاده شده در این پژوهش به دلیل تئوری بسیار قوی آن برای جداسازی داده‌های خارج از رده استفاده کرد.



شکل ۷: شناسایی داده‌های پرت با استفاده از حد آستانه تصحیح شده، خط آبی رنگ این حد آستانه را نشان می‌دهد.



شکل ۸: نقشه چند متغیره نشان دهنده‌ی نمونه‌های خارج از رده و نمونه‌های معمولی، شناسایی شده با استفاده از فاصله‌ی ماهالانوبیس مقاوم بر اساس حد آستانه تصحیح شده

۹- نتیجه‌گیری

داده‌های پرت داده‌هایی هستند که عضو توزیع اصلی جامع نبوده و جزء یک یا چند توزیع متفاوت هستند و حذف یا تصحیح آن‌ها یک از اساسی‌ترین و اولین مراحل پردازش در ژئوشیمی اکتشافی است. این داده‌ها را با استفاده از ۳ روش تک متغیره، دو متغیره و چند متغیره می‌توان شناسایی کرد که هدف این مطالعه تفکیک آن‌ها به روش‌های تک متغیره و چند متغیره بود. از بین روش‌های تک متغیره روش نمودار جعبه‌ای با توجه به مقاوم بودن آن در مقابل داده‌های پرت استفاده شد که در بیشتر متغیرها تمرکز نمونه‌های پرت بر روی تپه ساری گونای بود. برای تشخیص داده‌های پرت به روش چند متغیره از فاصله‌ی ماهالانوبیس مقاوم با یک حد آستانه تصحیح شده بر اساس تابع توزیع تجربی مربع فاصله ماهالانوبیس مقاوم و تابع توزیع مربع کای استفاده شد. از این روش به دلیل تئوری ریاضیاتی بسیار قوی آن در جداسازی داده‌های خارج از رده استفاده شد. نتایج حاصل از دو روش این موضوع را نشان می‌دهد که در مواردی که آنالیزهای چند متغیره مورد نیاز می‌باشد باید از روش‌های چند متغیره با پایه تئوری قوی از جمله ماهالانوبیس مقاوم برای تفکیک داده‌های پرت از غیر پرت استفاده کرد.

۱۰- تشکر و قدردانی

نویسندگان این مقاله از شرکت درساپردازه جهت همکاری در تهیه داده‌های استفاده شده در این تحقیق قدردانی می‌کند.

منابع

[1] Hair, J.F., Andersen, R.E., Tatham, R.L., and Black, W.C.; 1998; *Multivariate Data Analysis*, Prentice Hall, Upper Saddle River, New Jersey.

[2] Filzmoser, P., Garrett, R.G., and Reimann, C.; 2005; "*Multivariate outlier detection in exploration geochemistry*"; Computers and Geosciences, 31: 579-587.

[3] Lalor, G.C., and Zhang, C.; 2001; "*Multivariate outlier detection and remediation in geochemical databases*"; The Science of the Total Environment, 281: 99-109.

[4] Reimann, C., Filzmoser, P., and Garrett, R.G.; 2005; "*Background and threshold: critical comparison of methods of determination*"; Science of the Total Environment, 346: 1-16.

[5] Tukey J.; 1977; *Exploratory data analysis*, Reading, Massachusetts: Addison-Wesley, p. 506.

[6] Zhang, C.S., Wong, P.M., and Selinus, O.; 1999; "*A comparison of outlier detection methods: exemplified with an environmental geochemical dataset*"; In: Proceeding of the 6th International Conference on Neural Information Processing, Perth, Australia, P 183-187.

[7] Chiang, L.H., Pell, R.J., and Seasholtz, M.B.; 2003; "*Exploring process data with the use of robust outlier detection algorithms*"; J. Process Control, 13: 437-449.

[8] Garrett, R.G.; 1989; "*The chi-square plot: A tool for multivariate outlier recognition*"; Journal Geochemical Exploration, 32: 319-341.

[9] Filzmoser P, and Hron K.; 2008; "*Outlier detection for compositional data using robust methods*"; Mathematical Geoscience, 40:233–48.

[10] Filzmoser, P., Hron, and K., Reimann, C.; 2009; "*Univariate statistical analysis of environmental (compositional) data: problems and possibilities*"; Science of the Total Environment 407, 6100–6108.

[11] Filzmoser, P., Hron, K., and Reimann, C.; 2010; "*The Bivariate Statistical Analysis of Environmental (Compositional) Data*"; Science of The Total Environment408, p.p.4230-4238.

[12] Carranza, E.J.M.; 2011; "*Analysis and Mapping of Geochemical Anomalies Using Logratio- Transformation Stream Sediment Data with Censored Values*"; Journal of Geochemical Exploration 110, p.p.167-185.

[13] Aitchison, J.; 1986; *The statistical analysis of compositional data*, London, UK: Chapman and Hall, p. 416.

[14] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C.; 2003; "*Isometric logratio transformations for compositional data analysis*"; Mathematical Geology 35, 279–300.

[15] Wilkinson, L. Damien, 2005. *Geology and mineralization of the Sari Gunay gold deposits*, Kurdistan province Iran, Rio-Tinto Ltd technical report.

[16] Rousseeuw, P.J., and Van Driessen, K.; 1999; "*A fast algorithm for the minimum covariance determinant estimator*"; Technometrics, 41: 212-223.

[17] Gervini, D., 2003; "*A robust and efficient adaptive reweighted estimator of multivariate location and scatte*r"; Journal of Multivariate Analysis 84, 116–144.