مجله پژوهش‌های حفاظت آب و خاک
جلد هفدهم، شماره اول، ۱۳۸۹
www.gau.ac.ir/journals

تشخیص داده‌های پرت در روش منشأیابی رسوب

شاهرخ حکیم‌خانی[۱] و *احمد علیجان‌پور[۲]

[۱]استادیار گروه مرتع و آبخیزداری، دانشگاه ارومیه، [۲]استادیار گروه جنگلداری، دانشگاه ارومیه

تاریخ دریافت: ۸۸/۷/۲۹ تاریخ پذیرش: ۸۹/۲/۱۸

چکیده

اولین و مهم‌ترین مرحله روش منشأیابی رسوب، انتخاب ترکیب مناسبی از ردیاب‌ها است که قادر به جداسازی منابع رسوب باشند. انتخاب ترکیب یادشده با استفاده از تجزیه تابع تشخیص انجام می‌شود. وجود داده‌های پرت بر انتخاب ترکیب مناسبی از متغیرها اثر گذاشته و ممکن است مانع انتخاب متغیرهای مهم شده و توان جداسازی یا درصد طبقه‌بندی صحیح تجزیه تابع تشخیص را کاهش دهد. بنابراین داده‌های یاد شده باید شناسایی، و در صورت وجود شواهد کافی دال بر پرت بودن، نسبت به تصحیح یا حذف آنها اقدام شود. در این پژوهش، وجود داده‌های پرت به روش‌های مختلف در بین عناصر ژئوشیمیایی، آلی و رادیواکتیو نمونه‌های خاک جمع‌آوری شده از حوضه قره‌آغاج واقع در شهرستان ماکو بررسی شد. براساس ٤ روش تشخیص یک متغیره داده‌های پرت، هیچ نمونه‌ای از نظر زیادی از عناصر پرت نیستند. روش‌های میانه به اضافه یا منفی میانه انحراف‌های تمام داده‌ها از میانه و نمودار جعبه‌ای کارایی بهتری نسبت به روش‌های آزمون گراب و میانگین به اضافه یا منهای ۳ برابر انحراف معیار در تشخیص داده‌های پرت نشان دادند. روش‌های چندمتغیره تشخیص داده‌های پرت یعنی مربع فاصله ماهالانوبیس، نمودارهای جعبه‌ای مربع فاصله ماهالانوبیس به تفکیک منابع رسوب، نمودار چندک- چندک مربع فاصله ماهالانوبیس در مقابل توزیع مربع کای و روش تجزیه به مؤلفه‌های اصلی نیز نشان دادند که هیچ نمونه‌ای پرت محسوب نمی‌شوند. بنابراین براساس

* مسئول مکاتبه: aalijanpour@yahoo.com

هر دو گروه روش، شواهد و دلایل کافی مبنی بر پرت بودن هیچ نمونه‌ای وجود ندارد. مزایای رویکرد اتخاذ شده در این پژوهش سادگی و قابل انجام بودن روش‌های مورد استفاده با نرم‌افزارهای آماری موجود و بهره‌گیری از روش‌های مختلف می‌باشند و داده‌ای پرت محسوب می‌شود که از نظر چند روش تأیید شده باشد.

واژه‌های کلیدی: تشخیص داده‌های پرت، حوضه قره‌آغاج، نمودار جعبه‌ای، فاصله ماهالانوبیس، ردیاب‌ها

مقدمه

کسب اطلاعات از منابع رسوب، تعیین سهم و اهمیت نسبی آنها در تولید رسوب به‌منظور اجرای طرح‌های حفاظت خاک و کنترل رسوب ضروری است. به‌دلیل وجود مشکلات نمونه‌گیری چه از بعد مکانی و چه از بعد زمانی و نیز تنگناهای اجرایی و داشتن نیاز به زمان و هزینه زیاد در کاربرد روش‌های سنتی (کالینز و والینگ، ۲۰۰٤؛ لاگران و کمپبل، ۱۹۹۵)، روش انگشت‌نگاری، ردیابی یا به‌عبارتی منشأیابی[1] به‌عنوان روشی جایگزین و مناسب، کاربرد روزافزونی در تعیین سهم و اهمیت نسبی منابع رسوب در تولید رسوب پیدا کرده است (والینگ، ۲۰۰۵). در این روش، از خصوصیات فیزیکی، ژئوشیمیایی و آلی رسوب و منابع رسوب برای تعیین منابع اصلی رسوب و اهمیت نسبی آنها استفاده می‌شود (والینگ و همکاران، ۲۰۰۸؛ کالینز و والینگ، ۲۰۰٤). این روش بر این فرض استوار است که منابع مختلف رسوب با استفاده از تعدادی از خصوصیات شیمیایی (عناصر ژئوشیمیایی و نادر)، فیزیکی (بافت و اندازه ذرات) و آلی (کربن، نیتروژن و فسفر آلی) قابل شناسایی بوده و با مقایسه این خصوصیات با همان خصوصیات در نمونه‌های رسوب می‌توان سهم و اهمیت نسبی منابع رسوب در تولید رسوب را به‌دست آورد.

در مطالعات امروزی از روش منشأیابی مرکب و کمی استفاده می‌شود. به این شکل که در مرحله اول، ترکیبی مناسب از ردیاب‌ها که قادر به جداسازی منابع رسوب (نظیر واحدهای سنگ‌شناسی، کاربری‌های اراضی و انواع فرسایش) باشند انتخاب می‌شود. در مرحله دوم، ترکیب یادشده برای تعیین سهم نسبی هر یک از منابع رسوب با استفاده از مدل‌های ترکیبی چندمتغیره به‌کار می‌رود.

1- Fingerprinting Techniques

همان‌طور که گفته شد اولین و مهم‌ترین مرحله روش منشأیابی، انتخاب ترکیب مناسبی از ردیاب‌ها است که قادر به جداسازی منابع رسوب باشند. انتخاب ترکیب یادشده با استفاده از تجزیه تابع تشخیص[1] گام به گام انجام می‌شود (والینگ و همکاران، ۱۹۹۹؛ والینگ و همکاران، ۲۰۰۸؛ کالینز و والینگ، ۲۰۰۷). وجود داده‌های پرت[2] بر انتخاب ترکیب مناسبی از متغیرها اثر گذاشته و ممکن است مانع انتخاب متغیرهای مهم شود (ویگاند و همکاران، ۲۰۰۹) و در نتیجه توان جداسازی یا درصد طبقه‌بندی صحیح تجزیه تابع تشخیص را کاهش دهد. یکی از پیش‌فرض‌های اصلی روش تجزیه تابع تشخیص و سایر روش‌های چندمتغیره، نرمال چندمتغیره[3] است (هایر و همکاران، ۱۹۹۸). در تحلیل‌های چندمتغیره علاوه‌بر این‌که هریک از متغیرها باید از توزیع نرمال تبعیت کنند، ترکیب متغیرها نیز باید از توزیع نرمال (نرمال چندمتغیره) پیروی کند. از سوی دیگر بیشتر عناصر ژئوشیمیایی که از مهم‌ترین ردیاب‌ها در منشأیابی محسوب می‌شوند از توزیع نرمال تبعیت نمی‌کنند (ریمن و فیلزموزر، ۲۰۰۰؛ ژانگ و همکاران، ۲۰۰۸). ولی اگر تبعیت نکردن مجموعه متغیرها از فرض نرمال ناشی از وجود داده‌های پرت نبوده و به داده‌های حد مربوط باشد نتایج روش‌های چندمتغیره معتبر خواهد بود (تاباکنیک و فیدل، ۱۹۹٦). بنابراین داده‌های پرت از اهمیت زیادی برخوردارند و می‌توانند اثر زیادی بر نتایج روش‌های آماری از جمله صحت طبقه‌بندی تجزیه تابع تشخیص داشته باشند و باید قبل از هر چیز، شناسایی شده و تأثیر آنها بررسی شود و در صورت داشتن اثر منفی و وجود شواهد کافی دال بر پرت بودن، نسبت به تصحیح یا حذف آنها اقدام شود (هایر و همکاران، ۱۹۹۸). داده‌های پرت مشاهداتی هستند که از نظر خصوصیات مورد نظر متفاوت از مشاهدات دیگر می‌باشند. در تعریف ساده می‌توانیم بگوییم داده پرت مشاهده‌ای است که در فاصله دورتری از سایر داده‌ها قرار می‌گیرد و مقدار آن بسیار بزرگ یا بسیار کوچک می‌باشد.

داده‌های پرت در ژئوشیمی علاوه‌بر اشتباه یا خطا، اغلب ناشی از فرآیندهای کانی‌سازی، آلتراسیون و فعالیت‌های انسانی است (لالور و ژانگ، ۲۰۰۱؛ ریمن و همکاران، ۲۰۰۵). 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

1- Discriminant Function Analysis
2- Outliers
3- Multivariate Normality

چولگی مثبت در توزیع احتمال داده‌ها می‌شود، 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.

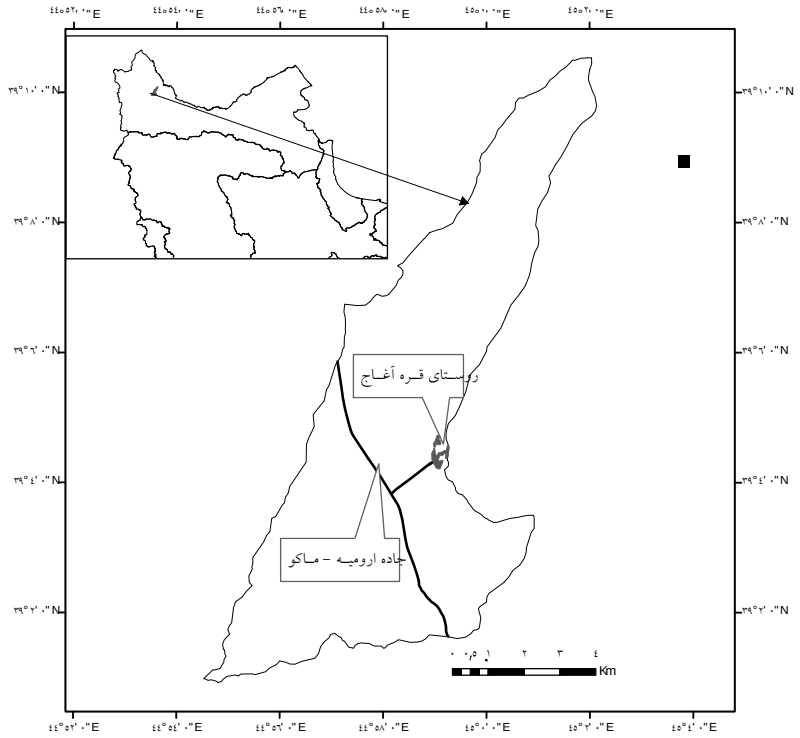
با این‌که تاکنون روش‌های زیادی برای تشخیص داده‌های پرت ارایه شده‌اند ولی هیچ‌کدام از آنها مقبولیت جهانی ندارند (ریمن و همکاران، ۲۰۰۵). در ضمن بسیاری از آنها نیاز به محاسبات زیاد و زمان‌بر داشته و قابل محاسبه با نرم‌افزارهای آماری موجود نیستند. بنابراین برای اطمینان و بررسی بیشتر می‌توان از چند روش تشخیص داده‌های پرت استفاده کرد. هدف این پژوهش، بررسی امکان وجود داده‌های پرت به روش‌های مختلف در بین عناصر ژئوشیمیایی، آلی و رادیواکتیو نمونه‌های خاک جمع‌آوری شده از حوضه قره‌آغاج واقع در شهرستان ماکو از استان آذربایجان می‌باشد.

مواد و روش‌ها

مشخصات حوضه مورد مطالعه: منطقه مورد مطالعه از حوزه‌های آب‌خیز مشرف به سد ارس و در جنوب بخش پلدشت از توابع شهرستان ماکو در استان آذربایجان غربی قرار داشته و به قره‌آغاج معروف است. این حوضه در محدوده ٤٤ درجه، ٥٤ دقیقه و ۳۱ ثانیه تا ٤٥ درجه، ۲ دقیقه و ۸ ثانیه طول شرقی ۳۹ درجه و ۵۵ ثانیه تا ۳۹ درجه، ۱۰ دقیقه و ۳۲ ثانیه عرض شمالی واقع شده است. مساحت حوزه تقریباً ٦٩٥٦ هکتار و شیب متوسط آن تقریباً ۹ درصد می‌باشد. کاربری‌های اصلی در منطقه شامل مرتع و زراعت دیم می‌باشند. متوسط بارندگی سالانه منطقه ۲۲۳ میلی‌متر است. مطالعات زمین‌شناسی نشان می‌دهد که واحدهای سنگ‌شناسی حوضه قره‌آغاج مربوط به دوران سوم زمین‌شناسی (دوره‌های ترشیری و کواترنر) بوده و شامل واحدهای رسوبات کواترنری، کنگلومرا و مارن می‌باشند. شکل ۱ موقعیت جغرافیایی حوضه را نشان می‌دهد.

در این مطالعه نقشه‌های کاربری اراضی، انواع فرسایش، واحدهای سنگ‌شناسی و واحدهای کاری حوضه تهیه شدند. نقشه انواع فرسایش حاوی فرسایش‌های سطحی (ورقه‌ای و شیاری) و فرسایش‌های زیرسطحی (فرسایش‌های خندقی، آبراهه‌ای و رودخانه‌ای) است. نقشه واحدهای سنگ‌شناسی و شامل واحدهای رسوبات کواترنری، کنگلومرا و مارن می‌باشد. نقشه کاربری اراضی دارای دو واحد یعنی اراضی کشاورزی و مراتع است. در تهیه نقشه‌های کاربری اراضی و انواع فرسایش از تصاویر ماهواره‌ای ETM^+، عکس‌های هوایی، نقشه‌های توپوگرافی و بازدیدهای صحرایی

استفاده شد. نقشه واحدهای کاری با روی‌هم‌گذاری و ترکیب سه نقشه کاربری اراضی، انواع فرسایش و واحدهای سنگ‌شناسی تهیه گردید که حدود ۹ واحد کاری به‌دست آمد.



شکل ۱- مرز و موقعیت جغرافیایی حوضه قره‌آغاج.

نمونه‌برداری و اندازه‌گیری آزمایشگاهی: از هر یک از واحدهای کاری با فرسایش‌های سطحی (ورقه‌ای و شیاری)، نمونه‌های خاک از عمق ۵-۰ سانتی‌متری و از واحدهای کاری با فرسایش‌های زیرسطحی (خندقی، رودخانه‌ای و آبراهه‌ای)، از دیواره‌ها به مقدار کافی (تقریباً ۲ کیلوگرم) و به‌وسیله یک بیلچه استیل برداشت شد (والینگ و همکاران، ۱۹۹۹). نمونه‌ها طوری برداشت شده‌اند که معرف تغییرات واحد کاری مربوطه از جمله شیب باشند. تعداد نمونه‌ها از هر واحد کاری ۵ و جمع نمونه‌ها ٤٥ عدد می‌باشد و تعداد ۵ نمونه نیز از رسوبات نهشته شده در بستر رودخانه در خروجی حوضه برداشت شد. نمونه‌ها بعد از خشک شدن در هوای آزاد و دمای اتاق، توسط یک هاون کوبیده شده و

برای جدا کردن بخش زیر ٦٣ میکرون از الک گذرانده شد. از ذرات رد شده از الک برای اندازه‌گیری ردیاب‌های انتخابی (کالینز و همکاران، ۱۹۹۸؛ والینگ و همکاران، ۱۹۹۹) در مرحله بعد استفاده گردید.

در این مطالعه ٣٦ ردیاب شامل عناصر رادیواکتیو سزیم ۱۳۷ و توریم ۲۳۲، عناصر ژئوشیمیایی بریلیم، بیسموت، سریم، سزیم، ایندیم، لانتانیم، نیوبیم، کادمیم، کبالت، کرم، مس، نیکل، سرب، قلع، روی، آهن، آلومینیوم، منگنز، یتریم، گالیم، ژرمانیم، وانادیم، تنگستن، تانتالیم، تلوریم، تیلوریم، تالیم، توریم، هافنیوم، زیرکنیم و سلنیم و کربن آلی و نیتروژن کل و فسفر قابل جذب با توجه به مطالعات گذشته (کالینز و همکاران، ۱۹۹۸؛ والینگ و همکاران، ۱۹۹۹؛ کالینز و والینگ، ۲۰۰۷) انتخاب شدند. تجزیه آزمایشگاهی و تعیین غلظت عناصر ژئوشیمیایی با استفاده از ترکیبی از دو روش طیف‌سنجی جرمی پلاسمای جفت شده القایی[1] و طیف‌سنجی نشری اتمی پلاسمای جفت شده القایی[2] و هضم توسط ٤ اسید HF، HNO_3، $HCLO_3$ و HCL، فسفر قابل جذب به روش اولسن و به طریق طیف‌سنجی، کربن آلی به روش والکر و بلاک و نیتروژن کل به روش کجلدال انجام شده است (علی احیایی و بهبهانی‌زاده، ۱۹۹۳). در ضمن عناصر رادیواکتیو به روش گاما اسپکترومتری (والینگ و کولینز، ۲۰۰۰) اندازه‌گیری شده‌اند.

روش‌های تشخیص داده‌های پرت: در مجموع روش‌های تشخیص داده‌های پرت را می‌توان به ۳ دسته تقسیم کرد (هایر و همکاران، ۱۹۹۸). این روش‌ها شامل تشخیص یک‌متغیره، دومتغیره و چندمتغیره می‌باشند. چون در این مطالعه روابط دومتغیره مورد بررسی قرار نمی‌گیرد، بنابراین از ذکر روش‌های تشخیص مربوطه خودداری می‌شود. روش‌های مختلفی برای بررسی یک‌متغیره داده‌های پرت (مشاهده پرت از نظر یک متغیر) وجود دارند که می‌توان آنها را به ۲ گروه دامنه و آزمون‌های آماری تقسیم کرد. در روش‌های دامنه توزیع مشاهدات بررسی شده و داده‌های خارج از یک دامنه معین به‌عنوان داده پرت تلقی می‌شوند. مهم‌ترین موضوع در این ارتباط تعیین دامنه یادشده برای مشخص کردن داده‌های پرت است. روش سنتی در این مورد، میانگین ($\overline{X}$) به اضافه یا منهای ۳ برابر انحراف معیار (S)[3] می‌باشد (ژانگ و همکاران، ۱۹۹۸؛ هایر و همکاران، ۱۹۹۸؛ چیانگ و همکاران، ۲۰۰۳) و داده‌های بزرگ‌تر از میانگین به اضافه ۳ برابر انحراف معیار و کوچک‌تر از میانگین منهای ۳

1- Inductively Coupled Plasma Mass Spectrometry
2- Inductively Coupled Plasma Atomic Emission Spectrometry
3- $\overline{X} \pm 3S$

برابر انحراف معیار پرت محسوب می‌شوند. چون این روش تحت‌تأثیر داده‌های پرت است (در محاسبه میانگین و انحراف معیار از تمام داده‌ها از جمله داده‌های پرت استفاده می‌شود)، از این‌رو روش‌های دیگری از جمله میانه به اضافه یا منفی میانه انحراف‌های تمام داده‌ها از میانه[1] و نمودار جعبه‌ای[2] (توکی، ۱۹۷۷؛ ریمن و همکاران، ۲۰۰۵) ارایه شده که تحت‌تأثیر داده‌های پرت قرار نمی‌گیرند. میانه انحراف‌های تمام داده‌ها از میانه (MAD) از رابطه زیر محاسبه می‌شود:

$$MAD = ۱/٤۸۲\text{Median}(|x_i - x_{median}| \qquad (۱)$$

مقدار ثابت ۱/٤۸۲ برای تبدیل MAD به برآورد نااریبی از انحراف معیار (امید ریاضی انحراف معیار نمونه برابر با انحراف معیار جامعه) انحراف معیار داده‌های گوسی (نرمال) است (چیانگ و همکاران، ۲۰۰۳).

نمودار جعبه‌ای نیز از روش‌های دامنه محسوب می‌گردد. این روش نموداری برای نشان دادن موقعیت، پراکنش و چولگی داده‌ها می‌باشد (توکی، ۱۹۷۷) و به فراوانی برای تشخیص داده‌های پرت استفاده می‌شود (ریمن و همکاران، ۲۰۰۵). این نمودار با استفاده از یک مستطیل (باکس) و دو خط یا میله[3] در دو طرف مستطیل و به‌وسیله میانه، چارک‌های اول ($Q_۱$) و سوم ($Q_۳$) و کم‌ترین و بیش‌ترین مقادیر رسم می‌شود. طول مستطیل برابر با فاصله چارکی[4] (تفاوت بین چارک سوم و چارک اول یا $IQR = Q_۳ - Q_۱$) است. در یک نوع نمودار جعبه‌ای که از آن برای تشخیص داده‌های پرت استفاده می‌شود، داده‌هایی که کوچک‌تر از $Q_۱ - ۱/۵IQR$ یا بزرگ‌تر از $Q_۳ + ۱/۵IQR$ باشند جزء داده‌های پرت خفیف و داده‌هایی که کوچک‌تر از $Q_۱ - ۳IQR$ یا بزرگ‌تر از $Q_۳ + ۳IQR$ باشند جزء داده‌های پرت قوی محسوب می‌شوند. داده‌های پرت خفیف را با علامت (○) و داده‌های پرت قوی را با علامت (✻) نشان می‌دهند.

از آزمون‌های آماری نظیر آزمون گراب[5] نیز برای تشخیص یک متغیره داده‌های پرت استفاده می‌شود (لالور و ژانگ، ۲۰۰۱). در آزمون گراب فرض بر این است که داده‌ها از توزیع نرمال پیروی می‌کنند. در آزمون یادشده، در هر مرحله یک داده پرت تشخیص داده می‌شود. در صورتی‌که داده پرتی

1- Median ± 3MAD
2- Box Plot
3- Whisker
4- Interquartile Range (IQR)
5- Grubbs' Test

شناسایی شود، داده یادشده حذف می‌گردد و آزمون برای بقیه داده‌ها دوباره انجام می‌شود. این کار آن‌قدر ادامه می‌یابد تا هیچ داده پرتی وجود نداشته باشد. فرض صفر این است که هیچ نوع داده پرتی وجود ندارد و فرض مخالف این است که حداقل یک داده پرت وجود دارد. آماره آزمون گراب (G) از رابطه زیر محاسبه می‌شود:

$$G = \frac{\max|x_i - \overline{x}|}{S} \qquad (۲)$$

که در آن، x_i کوچک‌ترین یا بزرگ‌ترین داده، $\overline{x}$ میانگین داده‌ها و S انحراف معیار داده‌ها می‌باشند. فرض صفر موقعی رد خواهد شد (فرض مخالف یا وجود حداقل یک داده پرت) که:

$$G > \frac{(n-1)}{\sqrt{n}} \sqrt{\frac{t^2_{(\alpha/2n, n-2)}}{n-2+t^2_{(\alpha/2n, n-2)}}} \qquad (۳)$$

که در آن، n اندازه نمونه و $t^2_{(\alpha/2n, n-2)}$ مقدار بحرانی آماره توزیع t اسیودنت با درجه آزادی $n-2$ و سطح معنی‌داری $\alpha/2n$، می‌باشند. در پژوهش حاضر به‌منظور تشخیص تعداد بیشتری از داده‌ها، α برابر با ۰/۰۵ انتخاب شد.

در این پژوهش به‌منظور تشخیص یک‌متغیره داده‌های پرت از روش‌های آزمون گراب، میانگین به اضافه یا منهای ۳ برابر انحراف معیار، میانه به اضافه یا منفی میانه انحراف‌های تمام داده‌ها از میانه و نمودار جعبه‌ای استفاده شد.

تشخیص چندمتغیره داده‌های پرت شامل بررسی چندمتغیره هر یک از مشاهدات براساس ترکیبی از متغیرها است. چون بیشتر تحلیل‌های چندمتغیره دارای بیش از دو متغیر است، تعیین داده‌های پرت از نظر ترکیبی از متغیرها نیز ضروری است. به‌منظور تشخیص چندمتغیره داده‌های پرت باید از روش‌ها یا معیارهایی استفاده شود که موقعیت چندبعدی (فضایی) هر یک از مشاهدات را نسبت به یک نقطه مشترک نشان دهند (هایر و همکاران، ۱۹۹۸). روش‌های مختلفی نظیر تجزیه به مؤلفه‌های اصلی (ژانگ و همکاران، ۱۹۹۹؛ کاوسینوس و همکاران، ۲۰۰۳؛ چیانگ و همکاران، ۲۰۰۳؛ لالور و ژانگ، ۲۰۰۱)، رگرسیون چندمتغیره (لالور و ژانگ، ۲۰۰۱)، شبکه‌های عصبی (لالور و ژانگ، ۲۰۰۱)، الگوریتم ژنتیک (ویگاند و همکاران، ۲۰۰۹) و روش‌های مبتنی بر فاصله ماهالانوبیس[1] (هایر و همکاران، ۱۹۹۸؛ روسیوو

1- Mahalanobis Distance

و ون دریسن، ۱۹۹۹؛ فیلزموزر و همکاران، ۲۰۰۵) به این منظور ارایه شده‌اند. از میان روش‌های یادشده، روش‌های مبتنی بر فاصله ماهالانوبیس معروف‌ترین می‌باشد (هایر و همکاران، ۱۹۹۸). در روش‌های یاد شده، فاصله ماهالانوبیس به‌عنوان معیاری از موقعیت چندبعدی هر یک از مشاهدات نسبت به مرکز ثقل کل مشاهدات عمل می‌کند. به‌عبارت دیگر فاصله یادشده، معیاری از فاصله هریک از مشاهدات در فضای چندبعدی از مرکز میانگین تمام مشاهدات است. برتری عمده فاصله ماهالانوبیس نسبت به سایر فاصله‌ها، در نظر گرفته شدن ماتریس کواریانس در آن است (فیلزموزر و همکاران، ۲۰۰۵)، چون شکل و اندازه داده‌های چندمتغیره به‌وسیله ماتریس کواریانس تعیین می‌شود. برای یک نمونه چندمتغیره P (تعداد متغیرها) بعدی، فاصله ماهالانوبیس برای مشاهده i ام از رابطه زیر به‌دست می‌آید:

$$MD_i = \left((X_i - \overline{X})^T C^{-1} (X_i - \overline{X})\right)^{1/2} \qquad (٤)$$

که در آن، X_i؛ بردار متغیرها برای مشاهده i ام، $\overline{X}$؛ بردار میانگین متغیرها (مرکز ثقل مشاهدات) و C؛ ماتریس کواریانس نمونه است.

خصوصیات مربع فاصله ماهالانوبیس (MD^2) به گونه‌ای است که اجازه استفاده از آزمون‌های آماری از جمله آزمون t و آزمون مربع کای را برای بررسی داده‌های پرت می‌دهد. هایر و همکاران (۱۹۹۸) از مقایسه $\frac{MD^2}{df}$ و توزیع t استفاده کرده‌اند. هرگاه مقدار یاد شده با توجه به درجه آزادی[1] (df) و سطح معنی‌داری مورد نظر بیشتر از t جدول باشد داده مربوطه پرت محسوب می‌شود. df (درجه آزادی) برابر با تعداد متغیرها است. بعضی از محققان (از جمله روسیوو و ونزومرن، ۱۹۹۰) از مقایسه MD^2 و توزیع مربع کای استفاده کرده‌اند. در تعدادی از مطالعات (روسیوو و ون‌دریسن، ۱۹۹۹؛ فیلزموزر و همکاران، ۲۰۰۵) با این استدلال که پارامترهای فاصله ماهالانوبیس (میانگین متغیرها و ماتریس کواریانس) متأثر از داده‌های پرت هستند و در نتیجه فاصله یادشده نیز تحت‌تأثیر داده‌های پرت خواهد بود، از فاصله ماهالانوبیس قوی[2] استفاده کرده‌اند که به داده‌های پرت حساس نیست. برای محاسبه فاصله یادشده از برآوردکننده‌های قوی از جمله کم‌ترین دترمینان ماتریس کواریانس استفاده می‌شود. برای محاسبه کم‌ترین دترمینان ماتریس کواریانس، تعداد h مشاهده (h معمولاً برابر با ۰/۷۵ تعداد کل مشاهدات) را طوری پیدا می‌کنند که ماتریس کواریانس آن دارای کم‌ترین دترمینان

1- Degree of Freedom
2- Robust Mahalanobis Distance

باشد. در ادامه ماتریس کواریانس و بردار میانگین متغیرها از h مشاهده یادشده برآورد شده و در نتیجه فاصله ماهالانوبیس حاصل حساسیت کمتری نسبت به داده‌های پرت خواهد داشت. روش‌های گرافیکی (گارت، ۱۹۸۹؛ فیلزموزر و همکاران، ۲۰۰۵؛ ژانگ و همکاران، ۲۰۰۸) که فاصله ماهالانوبیس را در مقابل توزیع مربع کای رسم می‌کنند نیز ارایه شده‌اند.

تجزیه به مؤلفه‌های اصلی، روشی جهت کاهش تعداد ابعاد داده‌های مورد مطالعه است. در تحلیل یادشده امکان تبدیل تعداد زیادی از متغیرهای اولیه به تعداد معدودی از متغیرهای جدید (ابعاد یا مؤلفه‌های اصلی) که بیش‌ترین واریانس مشاهده شده در متغیرهای اولیه را بیان کرده باشد بررسی می‌شود. از روش یادشده به‌منظور کشف روابط و همبستگی بین متغیرها نیز استفاده می‌گردد (هایر و همکاران، ۱۹۹۸). در سال‌های اخیر از روش تجزیه به مؤلفه‌های اصلی برای تشخیص داده‌های پرت نیز استفاده کرده‌اند (ژانگ و همکاران، ۱۹۹۹؛ چیانگ و همکاران، ۲۰۰۳؛ لالور و ژانگ، ۲۰۰۱). در این پژوهش از این روش نیز به‌علت توان بالای آن در تشخیص داده‌های پرت، سادگی و قابل انجام بودن آن با نرم‌افزارهای آماری موجود استفاده شد. تمام مؤلفه‌های اصلی که بیش از ۹۹ درصد واریانس متغیرها (ردیاب‌ها) را بیان کرده باشد انتخاب شدند. از فاصله نمرات نمونه[1] به‌عنوان معیاری برای بررسی داده‌های پرت که از رابطه زیر برآورد می‌شود استفاده شد (ژانگ و همکاران، ۱۹۹۹):

$$DSC_i = \sqrt{x_1^2 + x_2^2 + ... + x_n^2} \qquad (۲)$$

در این رابطه، DSC_i: فاصله نمرات نمونه i ام، n: تعداد مؤلفه‌های استخراجی و x_1، x_2 و x_n: به‌ترتیب برابر با نمرات مؤلفه اول، دوم و n ام برای نمونه i ام می‌باشند. هرچه فاصله نمرات نمونه بزرگ‌تر باشد امکان پرت بودن نمونه بیشتر است.

در تشخیص چندمتغیره داده‌های پرت از مربع فاصله ماهالانوبیس (مقایسه $\frac{MD^2}{df}$ و توزیع t)، تجزیه به مؤلفه‌های اصلی، نمودارهای جعبه‌ای مربع فاصله ماهالانوبیس و نمودار چندک- چندک[2] مربع فاصله ماهالانوبیس در مقابل توزیع مربع کای استفاده شد. براساس این نمودار چندک- چندک، داده‌هایی پرت هستند که از روند کلی بخش اصلی داده‌ها پیروی نکنند. در کل در این پژوهش داده‌ای پرت محسوب می‌شود که براساس بیش از ۵۰ درصد روش‌های یاد شده پرت باشد.

1- Distance of Sample Scores
2- Quantile-Quantile Plot

نتایج و بحث

بررسی یک‌متغیره داده‌های پرت بر مبنای آزمون گراب و معیارهای میانگین به اضافه یا منهای ۳ برابر انحراف معیار، میانه به اضافه یا منفی میانه انحراف‌های تمام داده‌ها از میانه و نمودار جعبه‌ای در جدول ۱ مشاهده می‌شود. براساس آزمون گراب تنها دو نمونه (نمونه‌های ۲ و ۲٦) از نظر چهار عنصر پرت هستند. نمونه ۲ از نظر سه عنصر و نمونه ۲٦ تنها از نظر یک عنصر پرت می‌باشند.

براساس معیار میانگین به اضافه یا منهای ۳ برابر انحراف معیار، ۵ نمونه از ٤٥ نمونه حداقل یک بار از نظر عناصر مورد استفاده پرت محسوب می‌شوند. از نمونه‌های پرت، ٤ مورد یک بار (نمونه‌های ۱، ۱۰، ۲۰ و ۲٦) و ۱ مورد (نمونه ۲) ٤ بار تکرار شده است.

بر مبنای معیار میانه به اضافه یا منفی میانه انحراف‌های تمام داده ها از میانه، ۱۲ مورد از ٤٥ نمونه، حداقل یک بار از نظر عناصر مورد استفاده پرت هستند. از نمونه‌های پرت، ۵ مورد یک بار، ۲ مورد دو بار، ۲ مورد سه بار، ۱ مورد (نمونه ۲٦) چهار بار، ۱ مورد (نمونه ۲۷) پنج بار و ۱ مورد (نمونه ۲) هفت بار تکرار شده‌اند. از نظر عناصر سزیم، بریلیم و بیسموت به‌ترتیب ٤، ۳ و ۳ نمونه پرت می‌باشند.

براساس نمودار جعبه‌ای ۱۹ نمونه حداقل یک بار از نظر عناصر مورد استفاده پرت محسوب می‌شوند. از نمونه‌های پرت، ۷ مورد یک بار، یک مورد دو بار (نمونه ۲۸)، ٦ مورد سه بار، ۲ مورد (نمونه‌های ۲۳ و ۲۷) چهار بار، ۲ مورد (نمونه‌های ۱٤ و ۲٦) پنج بار و ۱ مورد (نمونه ۲) هفت بار تکرار شده‌اند. در اینجا به‌علت زیاد بودن تعداد عناصر، ارایه نمودارهای جعبه‌ای برای همه عناصر امکان‌پذیر نبود، بنابراین تنها به نمودار جعبه‌ای عنصر تانتالیم اکتفا شده است (شکل ۲). نمودار یادشده علاوه‌بر نشان دادن دو نمونه پرت (نمونه‌های ۲ و ۱۳)، توزیع و پراکنش عنصر یادشده را نیز نمایش می‌دهد. طبق جدول ۱ و براساس روش نمودار جعبه‌ای، تعداد نمونه‌های پرت از نظر عناصر قلع، سزیم، منگنز، زیرکنیم و تنگستن به‌ترتیب ۱۰، ٤، ۳، ۳ و ۳ نمونه می‌باشند.

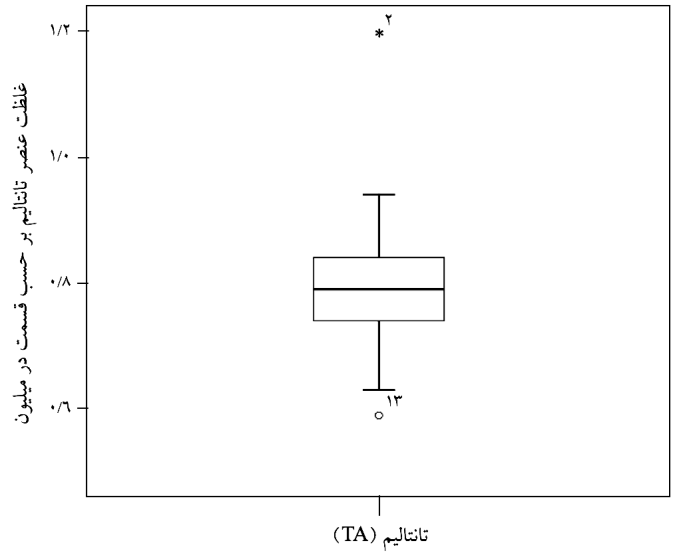
در کل با توجه به جدول ۱ و براساس هر چهار روش مورد استفاده برای تشخیص یک‌متغیره داده‌های پرت، هیچ نمونه‌ای از نظر تعداد قابل‌توجهی از عناصر پرت نیستند. نمونه ۲ براساس روش‌های میانه به اضافه یا منفی میانه انحراف‌های تمام داده‌ها از میانه و نمودار جعبه‌ای از نظر ۷ عنصر و طبق آزمون گراب و روش میانگین به اضافه یا منهای ۳ برابر انحراف معیار به‌ترتیب از نظر سه و چهار عنصر پرت محسوب می‌شود. با توجه به تعداد زیاد عناصر (۳٦ عنصر)، تعداد ۷ مورد برای نمونه ۲ نیز زیاد نیست. بنابراین شواهد و دلایل کافی مبنی بر پرت بودن هیچ نمونه‌ای وجود ندارد.

جدول ۱- نمونه‌های پرت براساس روش‌های یک‌متغیره.

عنصر	میانگین	انحراف معیار	شماره نمونه‌های پرت براساس روش‌های			
			آزمون گراب	میانگین به اضافه یا منهای ۳ برابر انحراف معیار	میانه به اضافه یا منفی میانه انحراف‌های تمام داده‌ها از میانه	نمودار جعبه‌ای
آلومینیوم	٦/٣٨	٠/٤٩				١٤
بریلیم	١/١٤	٠/١٩			٢، ١٤ و ٢٧	
بیسموت	٠/٠٧	٠/٠٢			٢٠، ٢٦ و ٢٧	
کادمیوم	٠/١٥	٠/٠٣				٢٢
سریم	٤٤/٤٠	٤/١٧	٢	٢	٢	٢
کبالت	٢٨/٦٢	٣/٠٧		١	١	١
کرم	٢٩١/٠٤	٨١/١٤				
سزیم	٢/٠٨	٠/٦٦			٢٢، ٢٤، ٢٧ و ٢٨	٢٢، ٢٤، ٢٧ و ٢٨
مس	٤٣/٩٦	٧/٢٣				
آهن	٤/٤٥	٠/٤٠				
گالیم	١٣/٩١	١/٣٥			٢	٢ و ١٤
زرمانیوم	٠/١٤	٠/٠١				١٣، ٢١ و ٢٧
هافنیوم	١/٩٨	٠/٣١			٢٣ و ٢٦	٢٠، ٢٣ و ٢٦
ایندیم	٠/٠٤	٠/٠١			١	١٧
لانتانیم	٢٢/٥٣	٢/١٢				٢
منگنز	٩٣١/١٦	٩٢/٨٢			٢٢ و ٢٣	١٧، ٢٢ و ٢٣
نیوبیم	١١/٨٠	١/٥٠	٢	٢	٢	٢ و ١٣
نیکل	٢١٠/٨٠	٣٧/٦٢				
سرب	٩/٩٦	١/٦٢		٢٠		٢٠
سلنیوم	١/٧١	٠/٤٦				
قلع	١/٤٤	٠/٢٠			٢ و ١٤	١، ٢، ١١، ١٣، ١٤، ١٦، ١٨، ٢٠، ٢٤ و ٢٦
تانتالیم	٠/٧٩	٠/١٠	٢	٢	٢	٢ و ١٣
تلوریم	٠/٠٥	٠/٠١				
توریم	٥/٠٣	٠/٨٢			٢٧	
تیلوریم	٠/٤٦	٠/٠٥				

ادامه جدول ۱- نمونه‌های پرت براساس روش‌های یک‌متغیره.

عنصر	میانگین	انحراف معیار	شماره نمونه‌های پرت براساس روش‌های: آزمون گراب	میانگین به اضافه یا منهای ۳ برابر انحراف معیار	میانه به اضافه یا منفی میانه انحراف‌های تمام داده‌ها از میانه	نمودار جعبه‌ای
تالیم	۰/۲۲	۰/۰٦				
وانادیم	۱٤۰/٥٦	۱٤/۰۷				
تنگستن	۰/۸٥	۰/۱۱				۱۷، ۲٤ و ۲۷
یتریم	۲۰/۱٥	۱/۸۸			۱٤	۱٤
روی	۷۲/٦۹	۷/۱٦		۲	۲ و ۲۷	۲ و ۲۷
زیرکنیم	٦۱/۱۳	۱۰/۳۳			۲۳ و ۲٦	۱٤، ۲۳ و ۲٦
سزیم ۱۳۷	۷/۱۷	۷/٤۸			۳۰	۲۳ و ۳۰
توریم ۲۳۲	٥۰/۲٤	۲/۸۸				۱ و ۲٦
نیتروژن	۰/۰۱	۰/۰۰			۳۸	۲۸ و ۳۲
کربن آلی	۰/٦۹	۰/٤۲		۱۰		۱۰
فسفر	۱۹/۲۱	۹/۷٦	۲٦	۲٦	۲٦	۲٦



شکل ۲- نمودار جعبه‌ای غلظت عنصر تانتالیم به‌منظور تشخیص داده‌های پرت.

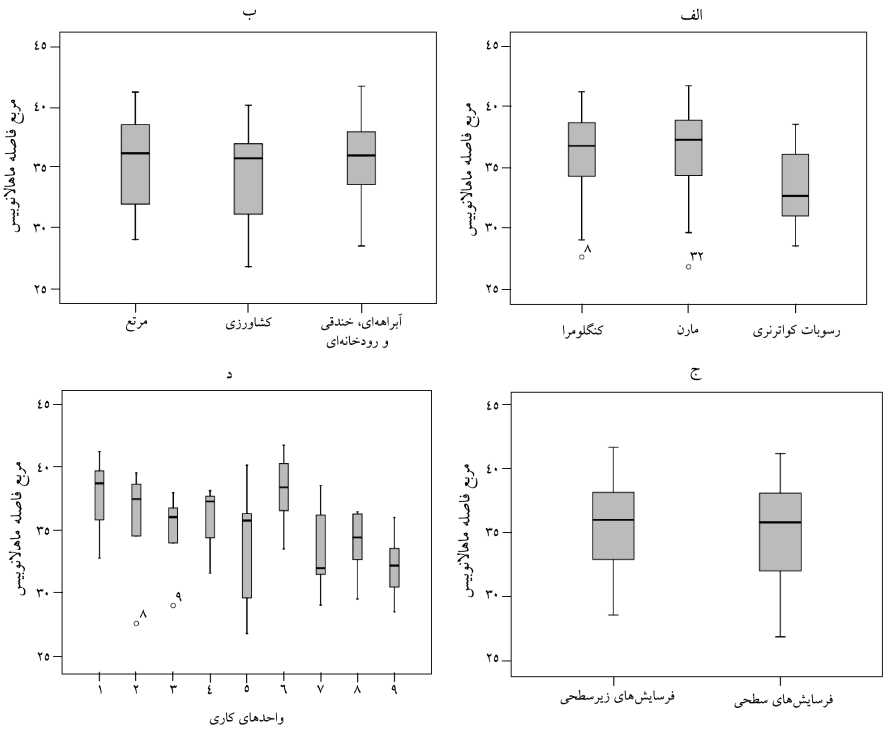
جدول ۲- مربع فاصله ماهالانوبیس (MD^2) و $\frac{MD^2}{df}$ به‌منظور تشخیص داده‌های پرت.

شماره نمونه	MD^2	$\frac{MD^2}{df}$	شماره نمونه	MD^2	$\frac{MD^2}{df}$	شماره نمونه	MD^2	$\frac{MD^2}{df}$	شماره نمونه	MD^2	$\frac{MD^2}{df}$
۱	۳۹/۴۸	۱/۱۰	۱۳	۳۲/۸۵	۰/۹۱	۲۴	۳۸/۴۸	۱/۰۷	۳۵	۳۲/۲۶	۰/۹۰
۲	۴۱/۷۷	۱/۱۶	۱۴	۳۸/۰۴	۱/۰۶	۲۵	۳۳/۷۰	۰/۹۴	۳۶	۳۰/۵۷	۰/۸۵
۳	۳۱/۶۶	۰/۸۸	۱۵	۳۵/۹۲	۱/۰۰	۲۶	۳۶/۵۶	۱/۰۲	۳۷	۳۲/۰۶	۰/۸۹
۴	۳۳/۶۴	۰/۹۳	۱۶	۳۹/۷۸	۱/۱۰	۲۷	۳۷/۳۶	۱/۰۴	۳۸	۳۸/۶۲	۱/۰۷
۵	۳۶/۴۲	۱/۰۱	۱۷	۳۹/۶۰	۱/۱۰	۲۸	۴۱/۱۹	۱/۱۴	۳۹	۳۶/۱۰	۱/۰۰
۶	۲۹/۷۲	۰/۸۳	۱۸	۳۶/۱۳	۱/۰۰	۲۹	۳۵/۸۶	۱/۰۰	۴۰	۲۸/۶۴	۰/۸۰
۷	۳۷/۵۶	۱/۰۴	۱۹	۳۴/۵۵	۰/۹۶	۳۰	۳۸/۲۱	۱/۰۶	۴۱	۳۲/۷۲	۰/۹۱
۸	۲۷/۷۳	۰/۷۷	۲۰	۳۴/۶۵	۰/۹۶	۳۱	۳۸/۰۳	۱/۰۶	۴۲	۳۱/۵۷	۰/۸۸
۹	۲۹/۱۳	۰/۸۱	۲۱	۳۴/۱۱	۰/۹۵	۳۲	۲۶/۹۴	۰/۷۵	۴۳	۲۹/۶۴	۰/۸۲
۱۰	۴۱/۲۷	۱/۱۵	۲۲	۳۵/۲۶	۰/۹۸	۳۳	۴۰/۲۱	۱/۱۲	۴۴	۲۹/۱۷	۰/۸۱
۱۱	۳۸/۷۴	۱/۰۸	۲۳	۳۸/۷۹	۱/۰۸	۳۴	۳۶/۳۹	۱/۰۱	۴۵	۳۶/۲۹	۱/۰۱
۱۲	۳۶/۸۷	۱/۰۲									

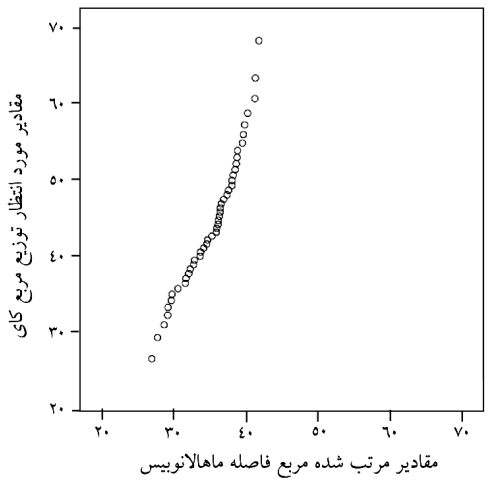
مربع فاصله ماهالانوبیس (MD^2) برای هر یک از نمونه‌ها با استفاده از معادله ٤ برآورد شد که نتایج در جدول ۲ ارایه شده است. در بررسی چندمتغیره داده‌های پرت، مقایسه $\frac{MD^2}{df}$ و توزیع t با درجه آزادی (df) (جدول ۲) برابر با ۳٦ و سطح معنی‌داری ۱ درصد نشان می‌دهد که مقدار $\frac{MD^2}{df}$ هیچ نمونه‌ای بیشتر از t جدول (یعنی ۲/٤٤) نبوده و بنابراین از نظر آزمون یادشده پرت محسوب نمی‌شوند. برای بررسی بیشتر نمونه‌های پرت، نمودارهای جعبه‌ای مربع فاصله ماهالانوبیس در شکل ۳ برای واحدهای سنگ‌شناسی (شکل ۳- الف)، کاربری اراضی (شکل ۳- ب)، انواع فرسایش (شکل ۳- ج) و واحدهای کاری (شکل ۳- د) ترسیم شده است. در شکل (۳- الف) نمونه‌های ۸ و ۳۲ و در شکل (۳- د) نمونه‌های ۸ و ۹ پرت به‌نظر می‌رسند، ولی نمونه‌های یاد شده دارای مربع فاصله ماهالانوبیس کوچک‌تری هستند و نمونه‌هایی می‌توانند پرت باشند که مربع فاصله ماهالانوبیس بزرگ‌تری نسبت به بقیه نمونه‌ها داشته باشند. در شکل‌های (۳- ب) و (۳- ج) هیچ نمونه‌ای پرت محسوب نمی‌شود.

شکل ٤ نمودار چندک- چندک مربع فاصله ماهالانوبیس در مقابل توزیع مربع کای مرتب شده را نشان می‌دهد. طبق شکل یادشده نیز هیچ نمونه‌ای یا داده‌ای پرت به‌نظر نمی‌رسد. چون پراکنش تمام نقاط (نمونه‌ها) از یک روند خاص پیروی می‌کند و انحراف قابل‌توجهی مشاهده نمی‌شود.

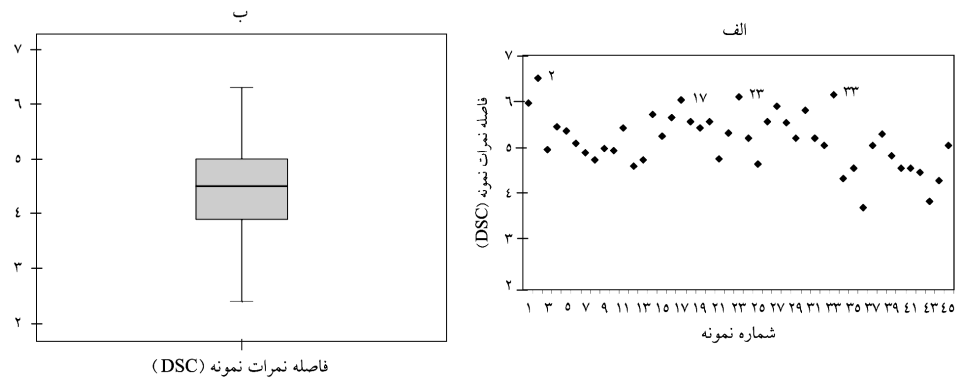
همان‌طورکه گفته شد در این پژوهش از روش تجزیه به مؤلفه‌های اصلی (به‌علت توان بالای آن در تشخیص چندمتغیره داده‌های پرت، سادگی و قابل انجام بودن با نرم‌افزارهای آماری موجود) نیز استفاده شد. ۲۱ مؤلفه اول بیش از ۹۹ درصد از واریانس متغیرها (ردیاب‌ها) را بیان کردند که از آنها برای محاسبه فاصله نمرات نمونه (معادله ٤) استفاده شد. هرچه فاصله نمرات نمونه بزرگ‌تر باشد امکان پرت بودن نمونه بیشتر است. شکل (۵- الف) پراکنش فاصله‌های نمرات نمونه‌ها را نشان می‌دهد. طبق شکل یادشده نمونه‌های ۲، ۱۷، ۲۳ و ۳۳ فاصله بیشتری از بقیه نمونه‌ها دارند. ولی فواصل یادشده قابل‌توجه نبوده و نمی‌توانند دلیل بر پرت بودن نمونه‌های یاد شده باشند. به‌منظور داشتن مقدار آستانه‌ای برای جداسازی فاصله‌های پرت از بقیه فاصله‌ها، از نمودار جعبه‌ای استفاده شد. نمودار جعبه‌ای ترسیم شده (شکل ۵- ب) نشان می‌دهد که هیچ نمونه‌ای براساس روش تجزیه به مؤلفه‌های اصلی پرت نیست. لالور و ژانگ (۲۰۰۱) نیز در مقایسه روش‌های رگرسیون چندمتغیره، شبکه عصبی و تجزیه به مؤلفه‌های اصلی نتیجه‌گیری کردند که روش اخیر کارایی بهتری در تشخیص داده‌های پرت نسبت به دو روش دیگر دارد.



شکل ۳- نمودارهای جعبه‌ای مربع فاصله ماهالانوبیس به تفکیک: الف- واحدهای سنگ‌شناسی، ب- کاربری‌های اراضی، ج- انواع فرسایش و د- واحدهای کاری.



شکل ٤- پراکنش مقادیر مربع فاصله ماهالانوبیس در مقابل مقادیر مورد انتظار توزیع مربع کای.



شکل ۵- فواصل نمرات نمونه‌ها (روش تجزیه به مؤلفه‌های اصلی)، الف- پراکنش فواصل نمرات نمونه‌ها و ب- نمودار جعبه‌ای فواصل نمرات نمونه‌ها.

با توجه به این‌که روش‌های آزمون گراب و میانگین به اضافه یا منهای ۳ برابر انحراف معیار تحت‌تأثیر داده‌های پرت و غیرعادی (داده‌های بسیار کوچک و بسیار بزرگ) می‌باشد (چون میانگین و انحراف معیار با استفاده از تمام داده‌ها برآورد می‌شوند) (ریمن و همکاران، ۲۰۰۵)، کارایی زیادی در تشخیص داده‌های غیرعادی و پرت ندارند و جدول ۱ نیز این نتیجه‌گیری را تأیید می‌کند. به‌طوری‌که روش‌های یادشده به‌خصوص آزمون گراب کم‌ترین تعداد نمونه‌های پرت (۵ نمونه) را در مقایسه با دو روش دیگر مشخص کرده‌اند. روش‌های میانه به اضافه یا منفی میانه انحراف‌های تمام داده‌ها از میانه و نمودار جعبه‌ای تحت‌تأثیر داده‌های پرت نبوده و می‌توانند تعداد بیشتری نمونه پرت را شناسایی کنند (چیانگ و همکاران، ۲۰۰۳؛ ریمن و همکاران، ۲۰۰۵). طبق جدول ۱، تعداد نمونه‌های پرت به روش نمودار جعبه‌ای ۱۹ و به روش میانه به اضافه یا منفی میانه انحراف‌های تمام داده‌ها از میانه، ۱۲ مورد است و این نشان می‌دهد که روش نمودار جعبه‌ای حساسیت زیادی به داده‌های غیرعادی دارد. ریمن و همکاران (۲۰۰۵) نتایج مشابهی در مطالعه خود به‌دست آوردند. بنابراین روش نمودار جعبه‌ای کارایی بهتری نسبت به دو روش دیگر در تشخیص داده‌های پرت دارد. روش نمودار جعبه‌ای، مشاهدات مشکوک به پرت را مشخص می‌کند و تصمیم‌گیری نهایی در مورد پرت بودن یک مشاهده براساس اطلاعات یادشده و سایر شواهد به عهده کارشناس خواهد بود.

بررسی چندمتغیره داده‌های پرت به روش‌های مربع فاصله ماهالانوبیس (جدول ۲)، نمودارهای جعبه‌ای مربع فاصله ماهالانوبیس (شکل ۳) و نمودار چندک- چندک مربع فاصله ماهالانوبیس در

مقابل توزیع مربع کای (شکل ٤) و تجزیه به مؤلفه‌های اصلی (شکل ٥) نیز بیانگر این است که هیچ نمونه‌ای پرت و دور افتاده نمی‌باشد. هایر و همکاران (۱۹۹۸) در مورد مربع فاصله ماهالانوبیس، گارت (۱۹۸۹) و فیلزموزر و همکاران (۲۰۰۵) در مورد نمودار چندک- چندک و ژانگ و همکاران (۱۹۹۹)، کاوسینوس و همکاران (۲۰۰۳) و چیانگ و همکاران (۲۰۰۳) در مورد تجزیه به مؤلفه‌های اصلی به نتایج مشابهی رسیدند.

در کل با توجه به این‌که هیچ‌یک از نمونه‌ها از نظر روش‌های یک‌متغیره، در تعداد زیادی از متغیرها (جدول ۱) و از نظر تمام روش‌های بررسی چندمتغیره، پرت نیستند، بنابراین شواهد کافی مبنی بر پرت بودن و عضو جامعه نبودن هیچ نمونه‌ای وجود نداشته و نمی‌توان نسبت به حذف نمونه‌ای اقدام کرد. داده‌های مورد استفاده در این پژوهش از نظر وجود خطا در ثبت، آماده‌سازی نمونه‌ها، روش‌های اندازه‌گیری عناصر و غیره کاملاً کنترل شده و تقریباً فاقد اشتباهات ناشی از عوامل یادشده می‌باشند و از سوی دیگر نمونه‌های خاک از محل‌های مناسب (معرف جامعه) و غیرآلوده برداشت شده‌اند، از این‌رو امکان وجود داده‌های پرت بسیار کم است. با توجه به مطلب یاد شده، چون از نظر روش‌های مورد استفاده پرت بودن هیچ داده‌ای تأیید نشد، بنابراین این روش‌ها کارایی بالایی در تشخیص داده‌های پرت نشان دادند. نمونه‌هایی که در بعضی از روش‌های مورد استفاده به‌عنوان پرت شناسایی شده‌اند در واقع جزو داده‌های حد محسوب می‌شوند. فرق داده‌های حد با داده‌های پرت این است که داده‌های حد عضو توزیع اصلی (جامعه) داده‌های مورد مطالعه بوده و در فاصله دورتری از مرکز توزیع قرار دارند (خیلی بزرگ یا خیلی کوچک هستند)، ولی داده‌های پرت عضو توزیع اصلی نبوده و جزو یک یا چند توزیع متفاوت هستند (ریمن و همکاران، ۲۰۰۵؛ فیلزموزر و همکاران، ۲۰۰۵). در واقع فرآیند یا فرآیندهای ایجادکننده داده‌های پرت متفاوت از داده‌های اصلی است. همان‌طورکه گفته شد داده‌های پرت در ژئوشیمی علاوه‌بر اشتباه یا خطا، اغلب ناشی از فرآیندهای کانی‌سازی، آلتراسیون و فعالیت‌های انسانی است.

مزایای اصلی روش‌های استفاده شده در این پژوهش سادگی و قابل انجام بودن آنها با نرم‌افزارهای آماری موجود می‌باشند. در ضمن از آنجایی‌که هیچ‌کدام از روش‌های ارایه شده برای تشخیص داده‌های پرت مقبولیت جهانی ندارند، برای حصول اطمینان به نتایج بررسی، از چند روش بهره‌گیری شده است. توصیه می‌شود در مطالعات آینده (چه در زمینه منشأیابی رسوب و چه در سایر مطالعات آماری) از رویکرد اتخاذ شده در این پژوهش استفاده شود. یعنی به‌منظور حصول اطمینان کافی به

نتایج بررسی داده‌های پرت از چند روش مختلف به‌خصوص روش‌های مورد استفاده در این مطالعه استفاده شود و تا زمانی که پرت بودن مشاهده‌ای از نظر چند روش تأیید نشده است داده پرت محسوب نشود. در ضمن اگر داده‌ای به روش‌های یادشده پرت تشخیص داده شد تا زمانی که دلیل کافی مبنی بر پرت بودن آن در دست نباشد نباید اقدام به حذف آن کرد. روش‌های پیچیده‌تری نظیر فاصله ماهالانوبیس قوی، شبکه عصبی، الگوریتم ژنتیک نیز وجود دارند و تشخیص داده‌های پرت با استفاده از آنها نیاز به محاسبه‌های زیاد داشته و با نرم‌افزارهای آماری موجود قابل انجام نیستند.

منابع

1.Aliehyai, M., and Behbahanizade, A.A. 1993. Methods of Chemical analysis of soil. Institute of Soil and Water Research, Bulletin No 893, 26p. (In Persian)

2.Caussinus, H., Fekri, M., Hakam, S., and Ruiz-Gazen, A. 2003. A monitoring display of multivariate outliers. Computational Statistics and Data Analysis, 44: 237-252.

3.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.

4.Collins, A.L., and Walling, D.E. 2004. Documenting catchment suspended sediment sources: problems, approaches and prospects. Progress in Physical Geography, 28: 159-196.

5.Collins, A.L., and Walling, D.E. 2007. Sources of fine sediment recovered from the channel bed of lowland groundwater-fed catchments in the UK. Geomorphology, 88: 120-138.

6.Collins, A.L., Walling, D.E., and Leeks, G.L. 1998. Use of composite fingerprints to determine the spatial provenance of the contemporary suspended sediment load transported by rivers. Earth Surface Processes and Landforms, 23: 31-52.

7.Filzmoser, P., Garrett, R.G., and Reimann, C. 2005. Multivariate outlier detection in exploration geochemistry. Computers and Geosciences, 31: 579-587.

8.Garrett, R.G. 1989. The chi-square plot: A tool for multivariate outlier recognition. J. Geochemical Exploration, 32: 319-341.

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

10.Hill, S.J., Fisher, A., and Cave, M. 2004. Inductively coupled plasma spectrometry. In: Smith, K.A., and Cresser, M.S. (eds). Soil and environmental analysis. Marcel Dekker, third edition, Pp: 53-110.

11.Lalor, G.C., and Zhang, C. 2001. Multivariate outlier detection and remediation in geochemical databases. The Science of the Total Environment, 281: 99-109.

12.Loughran, R.J., and Campbell, B.L. 1995. The identification of catchment sediment sources. In: Foster, I.D.L., Gumell, A.M., and Webb, B.W. (eds.). Sediment and Water Quality in River Catchments. Wiley, Chichester, Pp: 189-205.
13.Reimann, C., and Filzmoser, P. 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environ Geol. 39: 1001-1014.
14.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.
15.Rousseeuw, P.J., and Van Zomeren, B.C. 1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85: 633-651.
16.Rousseeuw, P.J., and Van Driessen, K. 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41: 212-223.
17.Tabachnick, B.G., and Fidell, L.S. 1996. Using Multivariate Statistics. Harper Collins College Publishers, 5th ed. New York, 963p.
18.Tukey, J.W. 1977. Exploratory data analysis. Addison-Wesley publication, Pp: 205-235.
19.Walling, D.E. 2005. Tracing suspended sediment sources in catchments and river systems. Science of the Total Environment, 344: 159-184.
20.Walling, D.E., Owens, P.N., and Leeks, G.J.L. 1999. Fingerprinting suspended sediment sources in the catchment of the River Ouse, Yorkshire, UK. Hydrological Processes, 13: 955-975.
21.Walling, D.E., and Collins, A.L. 2000. Integrated assessment of catchment sediment budgets: A technical manual. University of Exeter, 168p.
22.Walling, D.E., Collins, A.L., and Stroud, R. 2008. Tracing suspended sediment and particulate phosphorus sources in catchments. J. Hydrology, 350: 274-289.
23.Wiegand, P., Pell, R., and Comas, E. 2009. Simultaneous variable selection and outlier detection using a robust genetic algorithm. Chemometrics and Intelligent Laboratory Systems, 98: 2. 108-114.
24.Zhang, C.S., Selinus, O., and Schedin, J. 1998. Statistical analyses for heavy metal contents in till and root samples in an area of southeastern Sweden. The Science of the Total Environment, 212: 217-232.
25.Zhang, C.S., Wong, P.M., and Selinus, O. 1999. A comparison of outlier detection methods: exemplified with an environmental geochemical dataset, P 183-187, In: Proceeding of the 6th International Conference on Neural Information Processing, Perth, Australia.
26.Zhang, C., Fay, D., McGrath, D., Grennan, E., and Carton, O.T. 2008. Statistical analyses of geochemical variables in soils of Ireland. Geoderma, 146: 378-390.

Gorgan University of Agricultural Sciences and Natural Resources

J. of Water and Soil Conservation, Vol. 17(1), 2010
www.gau.ac.ir/journals

Detection of outliers in the sediment fingerprinting method

Sh. Hakim Khani[1] and *A. Alijan Pour[2]

[1]Assistant Prof., Dept. of Rangelands and Watershed Management, Urmia University,
[2]Assistant Prof., Dept. of Forestry, Urmia University

Abstract

Selection of the suite subset of tracers, capable of discriminating sediment sources, is the first and the most important step in the sediment fingerprinting method. Selection of the suite subset is carried out by Discriminant function analysis. The presence of outliers affects the suite subset selection and prevents entering the important tracers into the model, hence reducing accurate classification percent of Discriminant function analysis. Therefore, the outliers must be detected and corrected or omitted, if enough evidences were present. In this study, different univariate and multivariate outlier detection methods were used to assess the presence of outliers in geochemical and organic elements and radionuclides of soil samples collected from Ghara aghaj watershed, Makoo township. According to four univariate outlier detection methods, no observations (samples) were outlier on a sufficient number of tracers. The [Median ± 3MAD] and box plot procedures showed better performance in outlier identification than the [Mean ± 3S] and Grubbs' test methods. Also, based on multivariate outlier detection methods, namely squared Mahalanobis distance, separate box plots of squared Mahalanobis distance for each of sediment sources, principal component analysis and plot of the squared Mahalanobis distances against the quantiles of the chi-square distribution, no observations were detected as outlier. From perspectives of each of the two group methods, there was no sufficient information and demonstrable proof about true outlierness of any observation. The advantages of the approach adopted in this study are the simplicity and computability of the selected outlier detection methods with commonly used statistical softwares, and the condition that an observation is regarded as outlier if its uniqueness is confirmed with several methods.

Keywords: Outlier detection, Gharaghaj watershed, Mahalanobis distance, Box plot, Tracers

* Corresponding Author; Email: aalijanpour@yahoo.com