نشریه علمی-پژوهشی مهندسی و مدیریت آبخیز
جلد ۹، شماره ۴، ۱۳۹۶، صفحات ۳۹۸-۴۱۲

Journal of Watershed Engineering and Management
Volume 9, Issue 4, 2018, Pages 398-412

استفاده از روش‌های تک‌متغیره و چندمتغیره به‌منظور شناسایی داده‌های پرت در منشأیابی رسوبات، مطالعه موردی: حوزه آبخیز تنگ بستانک

احمد نوحه‌گر[۱]، محمد کاظمی[*۲]، سیدجواد احمدی[۳]، حمیدغلامی[۴] و رسول مهدوی[۵]

[۱] استاد، دانشکده محیط زیست، دانشگاه تهران، [۲] دانشجوی دکتری آبخیزداری، دانشکده منابع طبیعی، دانشگاه هرمزگان، [۳] دانشیار، پژوهشکده چرخه سوخت سازمان انرژی اتمی و [۴و۵] استادیار، دانشکده منابع طبیعی، دانشگاه هرمزگان

تاریخ دریافت: ۱۳۹۵/۰۶/۲۷ تاریخ پذیرش: ۱۳۹۵/۱۲/۰۹

چکیده

کارایی روش منشأیابی با ردیاب‌ها یا انگشت‌نگاری به‌عنوان روشی موفق و مؤثر برای تعیین منابع رسوبات به اثبات رسیده است. اولین و مهم‌ترین مرحله روش منشأیابی رسوب، انتخاب ترکیب مناسبی از ردیاب است که قادر به جداسازی منابع رسوب باشند. وجود داده‌های پرت بر انتخاب ترکیب مناسبی از ردیاب‌ها اثر گذاشته، ممکن است مانع انتخاب متغیرهای مهم (ردیاب‌های مؤثر) شده و توان جداسازی یا درصد طبقه‌بندی صحیح را کاهش دهد. بنابراین داده‌های یادشده باید شناسایی و در صورت وجود شواهد کافی دال بر پرت بودن، نسبت به تصحیح یا حذف آن‌ها اقدام شود. در مطالعه حاضر هدف شناسایی داده‌های پرت در بین مجموعه ردیاب‌های اندازه‌گیری شده در حوزه آبخیز تنگ بستانک، برای تشخیص بهترین ترکیب ردیاب‌ها بود. بر این اساس از روش‌های تک‌متغیره شناسایی داده‌های پرت همچون آزمون گراب، آزمون گوس، آزمون دیکسون، نمودار جعبه‌ای، میانه به اضافه یا منهای میانه انحراف‌های تمام داده‌ها از میانه و میانگین به اضافه و منهای سه برابر انحراف از معیار داده‌ها و نیز از روش‌های چندمتغیره شناسایی داده‌های پرت همچون تحلیل مؤلفه‌های اصلی، فاصله ماهالانوبیس، مربع فاصله ماهالانوبیس، نمودار چندک مربع فاصله ماهالانوبیس به روی درجه آزادی در برابر توزیع مربع کای، نمودارهای جعبه‌ای مربع فاصله ماهالانوبیس استفاده شد. در مجموع داده‌ای پرت شناخته شد که دست‌کم نیمی از روش‌های مذکور به پرت بودن آن اذعان داشته باشند. نتایج نشان داد روش میانه به اضافه و منهای میانه انحراف‌های تمام داده‌ها از میانه تعداد بیشتری از داده‌ها را به‌عنوان داده پرت معرفی می‌کند و همچنین روش‌های چندمتغیره اشتراک کمتری در تشخیص داده‌های پرت با یکدیگر دارند. آزمون‌های تک‌متغیره اجماع بهتری نسبت به شناسایی و معرفی داده‌های پرت دارند. برای استفاده از روش‌های تک-متغیره برای شناسایی داده‌های پرت روش‌های میانه به اضافه یا منهای میانه انحراف‌های تمام داده‌ها از میانه، نمودار جعبه‌ای و آزمون دیکسون به‌ترتیب حساسیت آن‌ها پیشنهاد می‌شود. همچنین نتایج نشان داد، بیشینه اجماع روش-های به‌کار رفته برای روش‌های تک‌متغیره چهار و برای روش‌های چندمتغیره دو مورد هست و در کل اجماع نیمی از روش‌ها برای پرت بودن داده‌ها مشاهده نشد.

واژه‌های کلیدی: آنالیز مؤلفه‌های اصلی، انگشت‌نگاری، ردیاب، فاصله ماهالانوبیس، گراب

---

[*] مسئول مکاتبات: mohamad.kazemi86@gmal.com

مقدمه

منشأیابی رسوب از اصول اولیه کنترل و مبارزه با فرسایش خاک محسوب می‌شود، زیرا با شناسایی مناطق برداشت می‌توان به جای پرداختن به معلول‌ها، علت‌ها را شناسایی نمود و فعالیت‌های اجرایی مبارزه با فرسایش را در مناطق برداشت متمرکز کرد (Feng و همکاران، ۲۰۱۱). لازمه اجرای برنامه‌های حفاظت خاک و کنترل رسوب، کسب اطلاعات از اهمیت نسبی منابع رسوب و سهم آن‌ها در تولید رسوب و در نتیجه شناسایی مناطق بحرانی در داخل آبخیز است (Walling و همکاران، ۲۰۰۸). به دلیل مشکلات مرتبط با روش‌های سنتی برای تعیین منابع اولیه رسوب داخل یک آبخیز، روش‌های منشأیابی به‌عنوان یک روش غیرمستقیم جمع‌آوری اطلاعات مورد توجه قرار گرفته است.

تلفیق مدل‌های ترکیبی در ارتباط با ردیاب‌های مرکب در اواخر دهه ۱۹۸۰ و اوایل دهه ۱۹۹۰ به مطالعات ردیابی منبع ورود کرد و امکان به‌دست آوردن تخمین‌های سهم نسبی از منابع مختلف را فراهم نمودند (Walling، ۲۰۰۵). در مطالعات امروزی از روش منشأیابی مرکب 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 یاد شده برای تعیین سهم نسبی هر یک از منابع رسوب با استفاده از مدل‌های ترکیبی چندمتغیره به‌کار می‌رود

. منشأیابی مرکب عبارت است از بررسی چندین منبع بالقوه رسوب از طریق چندین گروه متفاوت از خصوصیات که برای افزایش تشخیص بین منابع و اجتناب از منبع رسوب غیرواقعی است و شامل طیف وسیعی از خصوصیات مختلف شناختی می‌باشد (Collins و همکاران،۲۰۱۰). در این مدل‌های ترکیبی شناسایی مناطق مولد رسوب اولین و مهم‌ترین قدم، تشخیص ترکیب بهینه از عناصر ژئوشیمیایی و آلی می‌باشد. در این راستا وجود داده‌های پرت مانع از شناسایی ترکیب مؤثر و بهینه از عناصر خاکی به‌عنوان ردیاب خواهد شد (Walling و همکاران، ۲۰۰۸؛ Collins و Walling، ۲۰۰۷) و یا به‌عبارتی وجود داده‌های پرت بر انتخاب ترکیب مناسبی از متغیرها اثر گذاشته و ممکن است مانع انتخاب متغیرهای مهم شود (Wiegand و همکاران، ۲۰۰۹). همچنین قابل ذکر است که یکی از اصول مدل‌های مرکب در شناسایی مناطق مولد رسوب نرمال بودن متغیرها (ردیاب‌ها) می‌باشد.

در تحلیل‌های چندمتغیره علاوه بر این که 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 (Zhang و همکاران، ۲۰۰۸) و اگر این چولگی ناشی از عدم وجود داده‌های پرت باشد، مشکلی وجود ندارد یا به‌عبارتی 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 (Tabachnick و Fidell، ۱۹۹۶). بنابراین، ضروری به‌نظر می‌رسد که برای صحت و دقت طبقه‌بندی ابتدا داده‌های پرت را شناسایی و سپس وارد مراحل بعد شد (Hair و همکاران، ۱۹۹۸).

داده‌های پرت در عناصر ژئوشیمی و آلی خاک علاوه بر اشتباه یا خطا در اندازه‌گیری ممکن است بر اثر عواملی نظیر کانی‌سازی، آلتراسیون و یا فعالیت‌های انسانی باشد. تاکنون روش‌های زیادی برای تشخیص داده‌های پرت به‌کار گرفته شده است اما هیچ‌کدام مقبولیت جهانی نیافته‌اند (Reimann و همکاران، ۲۰۰۵). بنابراین برای اطمینان و بررسی بیشتر می‌توان نتایج اجماع چندین روش مختلف را به‌عنوان ملاک تصمیم‌گیری مد نظر قرار داد. هدف از تحقیق حاضر استفاده از روش‌های تک‌متغیره و چند‌متغیره برای شناسایی داده‌های پرت در منشأیابی رسوب در حوزه آبخیز تنگ بستانک می‌باشد.

مواد و روش‌ها

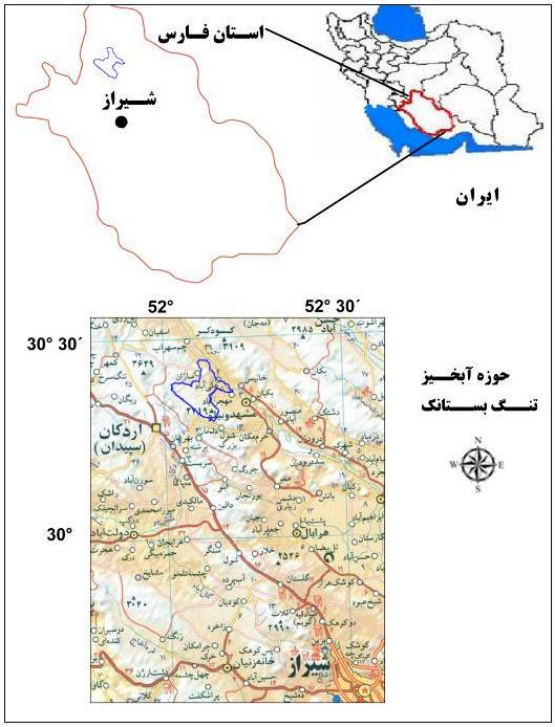
مشخصات منطقه مورد پژوهش: حوضه مورد مطالعه در این بررسی، تحت عنوان حوزه آبخیز تنگ بستانک در حدود ۸۰ کیلومتری شمال غرب شهرستان شیراز و در موقعیت جغرافیایی "۴۳ ´۰۳ ˚۵۲ تا "۳۶ ´۱۳ ˚۵۲ شرقی و "۳۳ ´۱۶ ˚۳۰ تا "۱۸ ´۲۵ ˚۳۰ شمالی واقع

شده است. این حوضه از نظر تقسیمات حوزه‌های آبخیز کشوری، جزء زیرحوزه آبخیز نیریز و شیراز بوده که آب‌های آن پس از وارد شدن به رودخانه کر، وارد دریاچه بختگان می‌شود. میانگین بارش سالانه این حوضه ۶۰۹ میلی‌متر می‌باشد، این حوضه طبق روش اقلیم‌نمای دومارتن اصلاح شده دارای اقلیم مدیترانه‌ای سرد می‌باشد. شکل ۱، موقعیت منطقه و راه‌های دسترسی به آن را نشان می‌دهد.

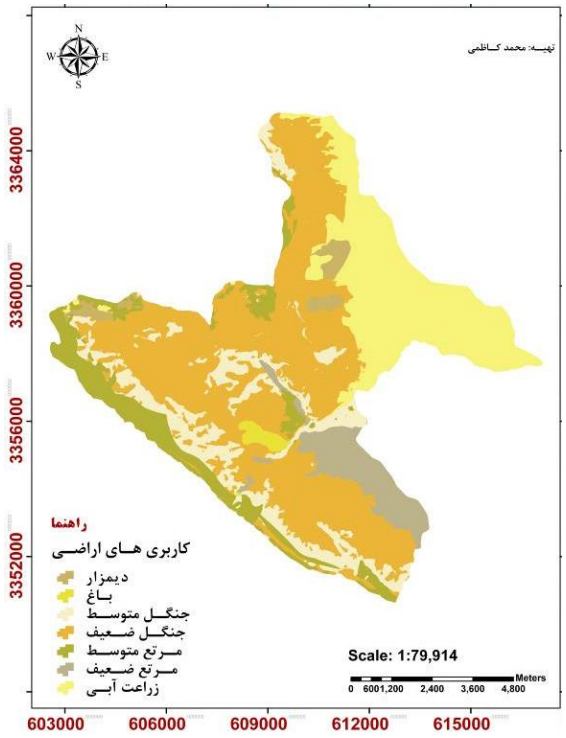
در این مطالعه نقشه کاربری اراضی تحت کاربری‌های اراضی زراعی، جنگل با پوشش ضعیف، جنگل با پوشش متوسط، باغات، مرتع با پوشش گیاهی ضعیف، مرتع با پوشش گیاهی متوسط و دیم‌زارها به‌عنوان واحدهای مولد رسوب و مطابق با روش الگوریتم بیشینه تشابه (ML) با تصویر ماهواره لندست ۸ سنجنده OLI تهیه شد.

اندازه‌گیری عناصر ژئوشیمیایی: آماده‌سازی نمونه‌ها به‌منظور اندازه‌گیری عناصر ژئوشیمیایی در چهار مرحله به‌ترتیب زیر انجام شد. از هر یک از واحدهای کاربری اراضی نمونه خاک از عمق صفر تا پنج سانتی‌متری و حدوداً به اندازه دو کیلوگرم با یک بیلچه استیل برداشت شد (Walling و همکاران، ۱۹۹۹). از هر واحد کاری با توجه به وسعت دست‌کم پنج نمونه برداشت و نمونه‌ها طوری انتخاب شدند که معرف تغییرات در نوع کاربری‌ها باشند. کلاً تعداد ۴۳ نمونه برداشت شد و بعد از خشک کردن نمونه‌ها در هوای آزاد و دمای اتاق از الک‌های ۶۲/۵، ۷۵، ۱۵۰، ۳۰۰، ۶۰۰، ۱۱۸۰ و ۱۷۰۰ میکرون عبور داده شدند. سپس به‌روی مقدار خاکی که کمتر از ۶۳ میکرون بود عملیات آزمایشگاهی زیر صورت گرفت.

مرحله یک: ابتدا سه گرم از هر نمونه وزن شده و در محلول Aqua Regia (ترکیب اسید کلریدریک (HCl) و اسید نیتریک (HNO_3) با نسبت ۱:۳ یعنی سه قسمت اسید کلریدریک و یک قسمت اسید نیتریک) هضم شد. در این مطالعه از ۲۱ میلی‌لیتر اسید کلریدریک و هفت میلی‌لیتر اسید نیتریک استفاده شد (Collins و همکاران، ۲۰۱۰ و ۲۰۱۲؛ Nosrati و همکاران، ۲۰۱۱). مرحله دو: سپس نمونه‌ها به مدت دو ساعت در دمای ۹۵ درجه سانتی‌گراد به‌روی حمام آبی قرار داده شد. مرحله سه: در این مرحله ابتدا نمونه‌ها از کاغذ صافی واتمن عبور داده و سپس به‌منظور اطمینان از شفافیت و عاری بودن نمونه‌ها از ذرات معلق دوباره از کاغذ صافی استات سلولز ۰/۲ میکرومتر عبور داده شد. مرحله چهار: اندازه‌گیری به‌وسیله ICP-Mass.



شکل ۱- موقعیت حوزه آبخیز تنگ بستانک و راه‌های دسترسی به آن



شکل ۲- نقشه واحدهای کاربری اراضی به‌عنوان منابع مختلف تولید کننده رسوب

در این مطالعه تعداد ۱۷ عنصر شامل ۱۵ عنصر ژئوشیمیایی و دو عنصر آلی استفاده شد. عناصر ژئوشیمیایی شامل باریم، کادمیوم، کروم، مس، لیتیم، منگنز، نئودیوم، نیکل، فسفر، سیلیکون، استرانسیوم، تیتانیوم، وانادیوم، زینک و دو ماده آلی نیتروژن و کربن بودند که در آزمایشگاه سازمان انرژی اتمی کشور اندازه‌گیری شدند. قابل ذکر است که اندازه‌گیری میزان کربن آلی به روش والکر و بلاک و نیتروژن کل به روش کجلدال انجام شده است (Aliehyai و Behbahanizade، ۱۹۹۳).

روش‌های تشخیص داده پرت: روش‌های تشخیص داده پرت را می‌توان در مجموع به سه دسته تقسیم کرد. روش‌های تک‌متغیره، دومتغیره و چندمتغیره (Hair و همکاران، ۱۹۹۸). در تحقیق حاضر از روش‌های تک‌متغیره و چندمتغیره استفاده شد.

روش‌های تک‌متغیره شناسایی داده‌های پرت: این روش‌ها را می‌توان به دو گروه روش‌های دامنه‌ای و آزمون‌های آماری تقسیم نمود. 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 یاد شده برای مشخص کردن داده‌های پرت هست. روش سنتی در این مورد میانگین به‌اضافه و منهای سه برابر انحراف از معیار[1] می‌باشد (Chiang و همکاران، ۲۰۰۳). داده‌های بزرگ‌تر از میانگین به اضافه سه برابر انحراف معیار و کوچک‌تر از میانگین منهای سه برابر انحراف معیار پرت محسوب می‌شوند. در این روش ۹۹/۸ درصد از داده‌ها پرت محسوب نمی‌شوند. روش مذکور دارای دقت پایینی می‌باشد و چون این روش تحت تأثیر 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[2] و نمودار جعبه‌ای[3] (Tukey، ۱۹۷۷؛ Reimann و همکاران، ۲۰۰۵) ارائه شده که تحت تأثیر داده‌های پرت قرار نمی‌گیرند.

میانه انحراف تمام داده‌ها از میانه (MAD) از رابطه (۱) محاسبه می‌شود. در منشأیابی رسوبات با استفاده از مدل‌های ترکیبی داده‌های نرمال و یا دارای چولگی (ناشی از داده‌های حد) پیش فرض ورود به این مدل‌ها

[1] X ± 3S
[2] Median ± 3MAD
[3] Box plot

می‌باشند. اما اگر چولگی و عدم تبعیت از توزیع نرمال داده‌ها به‌خاطر وجود داده‌های پرت باشد، نتایج مدل-های ترکیبی دارای اشکال می‌باشد و لازم است برای انتخاب بهترین ترکیب از ردیاب‌ها (ترکیب بهینه ردیاب‌ها) داده‌های پرت حذف شوند (Tabachnick و Fidell، ۱۹۹۶). به‌عبارت دقیق‌تر، داده‌ها (مقادیر ردیاب‌ها) برای ورود به تابع تحلیل تشخیص (DFA) و طبقه‌بندی صحیح هر ردیاب در هر منبع باید عاری از داده‌های پرت باشند.

$$MAD = 1.482 Median(|X_i - X_{Median}|) \quad (۱)$$

مقدار ثابت ۱/۴۸۲ برای تبدیل MAD به برآورد نااریبی از انحراف معیار (امید ریاضی انحراف معیار نمونه برابر با انحراف معیار جامعه) داده‌های نرمال است (Chiang و همکاران، ۲۰۰۳). نمودار جعبه‌ای نیز از روش‌های دامنه محسوب می‌شود. این روش نموداری برای نشان دادن موقعیت، پراکنش و چولگی داده‌ها می‌باشد (Tukey، ۱۹۷۷) و به فراوانی برای تشخیص داده‌های پرت استفاده می‌شود (Reimann و همکاران، ۲۰۰۵). این نمودار با استفاده از یک مستطیل (باکس) و دو خط یا میله در دو طرف مستطیل و به‌وسیله میانه و چارک‌های اول (Q_1) و سوم (Q_3) و کمترین و بیشترین مقادیر رسم می‌شود. طول مستطیل برابر با تفاوت بین چارک سوم و چارک اول (IQR) است. در یک نوع از نمودار جعبه‌ای که از آن برای تشخیص داده‌های پرت استفاده می‌شود، داده‌هایی که کوچک‌تر از Q_1-1.5IQR و نیز بزرگ‌تر از Q3+1.5IQR باشند جزء داده‌های پرت خفیف و داده-هایی که کوچک‌تر از Q_1-3IQR و بزرگ‌تر از Q3+3IQR باشند، پرت قوی محسوب می‌شود. از آزمون‌های آماری نظیر آزمون دیکسون[1] برای تشخیص داده پرت استفاده می‌شود. در این آزمون با فرض تبعیت داده‌ها از توزیع نرمال به‌شرطی که مقدار محاسباتی Q بزرگ‌تر از Q بحرانی (Q جدول) باشد داده پرت محسوب می‌شود (Alfassi، ۲۰۰۵؛ Rorabacher، ۱۹۹۱). در روش فوق برای شناسایی داده پرت لازم است تا ابتدا مقادیر حد با توجه به نمودار شاخه و برگ شناخته شوند (EPA، ۲۰۰۶). در این روش فرض صفر این است که مقدار حدی با بقیه نمونه‌های جمعیت تفاوت ندارد و فرض یک این است که مقدار حدی با بقیه نمونه‌های جمعیت تفاوت دارد. آماره آزمون دیکسون از رابطه (۲) محاسبه می‌شود (تعداد نمونه‌ها در هر ردیاب باید بیشینه ۱۰۰ و یا کمتر از ۱۰۰ باشد).

$$Q = \frac{x_n - x_{n-1}}{x_n - x_1} \quad (۲)$$

که در آن، x_n میزان داده حد، x_{n-1} میزان داده قبل از مقادیر حدی و x_1 کمترین مقدار طبقه‌بندی شده در نمودار شاخه و برگ مقادیر حد می‌باشد.

آزمون‌های دیگری نظیر آزمون گراب[2] نیز برای تشخیص یک‌متغیره داده‌های پرت استفاده می‌شود (Lalor و Zhang، ۲۰۰۱). در آزمون گراب فرض بر این است که داده‌ها از توزیع نرمال پیروی می‌کنند و اگر توزیع داده‌ها غیر از این و خصوصاً توزیع متقارن باشد (مثلاً توزیع لوگ نرمال)، نتایج این آزمون اشتباه خواهد بود. در آزمون یاد شده، در هر مرحله یک داده پرت تشخیص داده می‌شود. در صورتی که داده پرتی شناسایی شود، داده یاد شده حذف می‌شود و آزمون برای بقیه داده‌ها دوباره انجام می‌شود. این کار آن قدر 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 دست‌کم یک داده پرت وجود دارد. آماره آزمون گراب (G) از رابطه (۳) محاسبه می‌شود (در این آزمون هم تعداد نمونه‌ها در هر ردیاب باید درنهایت ۱۰۰ باشد).

$$G = \frac{max|X_i - \bar{X}|}{S} \quad (۳)$$

که در آن، X_i کوچک‌ترین یا بزرگ‌ترین داده، $\bar{X}$ میانگین داده‌ها و S انحراف از معیار داده‌ها می‌باشند. فرض صفر موقعی رد خواهد شد که (فرض مخالف یا کمینه وجود یک داده پرت) شرط زیر برقرار باشد (رابطه ۴).

$$G > \frac{(n-1)}{\sqrt{n}} \sqrt{\frac{t^2_{\left(\frac{\alpha}{2n, n-2}\right)}}{n-2+t^2_{\left(\frac{\alpha}{2n, n-2}\right)}}} \quad (۴)$$

که در آن، n اندازه نمونه و $t^2_{\left(\frac{\alpha}{2n, n-2}\right)}$ مقدار بحرانی آماره توزیع t استیودنت با درجه آزادی n-2 و سطح معناداری α/2n می‌باشد. روش دیگری که برای

[1] Dixon' Q test

[2] Grubbs' test

شناسایی داده‌های پرت استفاده می‌شود، به روش J.Gauss مشهور است و مطابق با رابطه (۵) محاسبه می‌شود.

$$g = \frac{x_{extr} - \bar{x}}{S} \qquad (۵)$$

که در آن، x_{extre} مقادیر حدی، S انحراف از معیار و x میانگین داده‌ها می‌باشد. چنان‌چه مقدار g محاسباتی از رابطه (۵) از مقدار G جدول بزرگ‌تر باشد، آن‌گاه داده پرت محسوب می‌شود (Szalma، ۱۹۸۴).

روش‌های چندمتغیره شناسایی داده‌های پرت:

تشخیص چندمتغیره داده‌های پرت شامل بررسی چندمتغیره هر یک از مشاهدات بر اساس 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 چند بعدی (فضایی) هر یک از مشاهدات را نسبت به یک نقطه مشترک نشان دهند (Hair و همکاران، ۱۹۹۸). روش‌های مختلفی نظیر تجزیه به مؤلفه‌های اصلی (Caussinus، ۲۰۰۳)، روش‌های مبتنی بر فاصله ماهالانوبیس (Filzmoser و همکاران، ۲۰۰۵) بدین منظور ارائه شده‌اند. از میان این روش‌ها فاصله ماهالانوبیس معروف‌تر است (Hair و همکاران، ۱۹۹۸). در روش‌های یاد شده، فاصله ماهالانوبیس به‌عنوان معیاری از موقعیت چند بعدی هر یک از مشاهدات نسبت به مرکز ثقل کل مشاهدات عمل می‌کند. به عبارت دیگر فاصله یاد شده، معیاری از فاصله هر یک از مشاهدات در فضای چند بعدی از مرکز میانگین تمام مشاهدات است. برتری عمده فاصله ماهالانوبیس نسبت به سایر فاصله‌ها، در نظر گرفته شدن ماتریس کواریانس در آن است (Filzmoser 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 iام از رابطه (۶) به‌دست می‌آید.

$$MD_i = [(X_i - \bar{X})^T C^{-1} (X_i - \bar{X})]^{1/2} \qquad (۶)$$

که در آن، X_i بردار متغیرها برای مشاهده iام، $\bar{X}$ بردار میانگین متغیرها (مرکز ثقل مشاهدات) و C ماتریس کوواریانس نمونه است. خصوصیات مربع فاصله ماهالانوبیس (MD^2) به گونه‌ای است که اجازه استفاده از آزمون‌های آماری از جمله t و آزمون مربع‌کای را برای بررسی داده‌های پرت می‌دهد. Hair و همکاران (۱۹۹۸) از مقایسه MD^2/df و توزیع t استفاده کردند. هرگاه مقدار یاد شده با توجه به درجه آزادی (df) (تعداد ردیاب‌ها منهای یک) و سطح معناداری مورد نظر بیشتر از t جدول باشد، داده مربوطه پرت محسوب می‌شود. بعضی از محققان از جمله Rousseeuw و Van Driessen (۱۹۹۹) از مقایسه MD^2 و توزیع مربع کای استفاده کردند. همچنین، قابل ذکر است که روش‌های گرافیکی (Garrett، ۱۹۸۹؛ Filzmoser و همکاران، ۲۰۰۵؛ Zhang و همکاران، ۲۰۰۸) که فاصله ماهالانوبیس را در مقابل توزیع مربع‌کای رسم می‌کنند نیز ارائه شده‌اند.

تجزیه به مؤلفه‌های اصلی، روشی برای کاهش تعداد ابعاد داده‌های مورد مطالعه است. در تحلیل یاد شده 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 یاد شده به‌منظور کشف روابط و همبستگی بین متغیرها نیز استفاده می‌شود (Hair و همکاران، ۱۹۹۸). در سال‌های اخیر از روش تجزیه به مؤلفه‌های اصلی برای تشخیص داده‌های پرت نیز استفاده کرده‌اند (Zhang و همکاران، ۱۹۹۹؛ Chiang و همکاران، ۲۰۰۳؛ Lalor و Zhang، ۲۰۰۱) در این پژوهش از این روش نیز به‌علت توان بالای آن در تشخیص داده‌های پرت، سادگی و قابل انجام بودن آن با نرم افزارهای آماری استفاده شد. تمام مؤلفه‌های اصلی که بیش از ۹۹ درصد واریانس متغیرها (ردیاب‌ها) را بیان کرده باشد انتخاب می‌شوند. از فاصله نمرات نمونه به‌عنوان معیاری برای بررسی داده‌های پرت که از رابطه زیر (رابطه ۷) برآورد می‌شود، استفاده می‌شود (Zhang و همکاران، ۱۹۹۹).

$$DSC_i = \sqrt{X_1^2 + X_2^2 + \cdots + X_n^2} \qquad (۷)$$

که در آن، DSC فاصله نمرات نمونه iام، n تعداد مؤلفه‌های استخراجی و X_1، X_2 و X_n به‌ترتیب برابر با نمرات مؤلفه اول، دوم و nام برای نمونه iام می‌باشند. هرچه فاصله نمرات نمونه بیشتر باشد، امکان پرت بودن نمونه‌ها بیشتر است. همچنین، در تشخیص چندمتغیره داده‌های پرت از مربع فاصله ماهالانوبیس (مقایسه MD^2/df و توزیع t) نمودار جعبه‌ای مربع فاصله ماهالانوبیس و نمودار چندک‌چندک[1] مربع فاصله ماهالانوبیس در مقابل مربع توزیع کای استفاده شد. بر اساس نمودار چندک چندک نمونه‌هایی که از روند کلی داده‌ها تبعیت نکنند، پرت محسوب می‌شوند (درجه آزادی تعداد ردیاب‌ها منهای یک می‌باشد).

نتایج و بحث

نتایج آزمون گراب (G): برای عنصر Ba (باریم) طبق نمودار شاخه و برگ (برای هر ردیاب) نقطه ۲۴ یک مقدار حدی (EXTEREME) است. G جدول معادل ۲/۹ می‌باشد و بیشینه مقدار G برابر ۲/۶۹ هست. چون مقدار G کمتر از G جدول هست لذا فرض اول قبول است و داده پرت نیست. کمترین میزان G مربوط به نقطه ۲۷ می‌باشد که مقدار آن معادل ۱/۴۵ است. برای این نقطه هم فرض اول قبول و نقطه پرت محسوب نمی‌شود. برای عنصر Cd (کادمیوم) مقدار حدی وجود ندارد. برای نقطه ۳۳ مقدار G کمترین مقدار هست (۱/۵) که کوچک‌تر از مقدار جدول می‌باشد و فرض اول قبول می‌شود و داده پرتی مشاهده نمی‌شود.

برای عنصر Co (کبالت) نمودار شاخه و برگ مقدار حدی نشان نمی‌دهد و داده پرتی وجود ندارد. برای عنصر Cu (مس) نمودار شاخه و برگ مقدار حدی نشان نمی‌دهد. مقدار G دست‌کم معادل ۱/۳۹ هست که ازمقدار G جدول کمتر بوده، آزمون نشان می‌دهد داده پرتی نیست. برای عنصر Li (لیتیم) مقدار G بیشینه و کمینه به‌ترتیب معادل ۲/۱۹ و ۱/۳۹ هست که مقدار G جدول کوچک‌تر بوده و داده پرتی وجود ندارد. برای عنصر Mn (منگنز) در نمودار شاخه و برگ برای نقاط ۱۱، ۱۵ و ۲۴ مقادیر بحرانی دیده می‌شود. اما مقدار G بیشینه در این نقاط برابر ۲/۷۷ هست که از مقدار جدول کوچک‌تر بوده، بنابراین، این نقاط پرت محسوب نمی‌شوند. برای عنصر Nd (نئودیمیوم) نقطه ۱۶ بحرانی محسوب می‌شود، مقدار G بیشینه ۳/۳۴ هست که از میزان جدول هم بزرگ‌تر هست و این نقطه برای این عنصر پرت محسوب می‌شود.

برای عنصر Ni (نیکل) در نمودار شاخه و برگ مقدار بحرانی دیده نمی‌شود. میزان G بیشینه آن برابر ۲/۸۵ هست که از مقدار جدول کوچک‌تر می‌باشد و داده پرتی محسوب نمی‌شود. برای عنصر P (فسفر) نقطه ۲۰ مقداری بحرانی دارد و G بیشینه معادل سه است که از میزان جدول بزرگ‌تر است و داده پرت محسوب می‌شود. برای عنصر Si (سیلیکون) نقطه ۱۶ دارای مقدار بیشینه می‌باشد و بحرانی تلقی می‌شود که میزان G آن معادل ۲/۴۴ هست و چون کوچک‌تر از مقدار G جدول است، داده پرتی محسوب نمی‌شود. برای عنصر Sr (استرانسیوم) مقادیر بحرانی وجود ندارد و میزان G بیشینه برابر ۱/۱۴ است که نشان می‌دهد برای این عنصر در هیچ نقطه‌ای مقدار پرتی وجود ندارد. برای عنصر Ti (تیتانیوم) در نمودار شاخه و برگ مقدار بحرانی در نقطه ۲۰ مشاهده می‌شود که G بیشینه معادل ۳/۲۱ هست و داده برای این نقطه پرت محسوب می‌شود.

برای عنصر V (وانادیوم) مقادیر بحرانی وجود ندارد و مقادیر بیشینه G برابر ۱/۵ و کمینه آن معادل ۱/۱۴ است که کوچک‌تر از مقدار G جدول است و داده پرتی برای این عنصر دیده نمی‌شود. برای عنصر Zn (زینک) مقادیر بحرانی وجود ندارد. مقدار G بیشینه معادل ۲/۲۵ و مقدار G کمینه ۱/۸ است که حکایت از عدم وجود داده پرت دارد. برای عنصر N (نیتروژن) مقادیر بحرانی دیده نمی‌شود. مقدار G بیشینه معادل دو و مقدار G کمینه ۱/۵ است که حکایت از عدم وجود داده پرت دارد. برای عنصر C (کربن) مقادیر بیشینه وجود ندارد و مقدار G بیشینه معادل ۱/۶۶ و مقدار G کمینه ۱/۴۹ است که عدم وجود داده پرت را نشان می‌دهد. بنابراین مطابق با این روش برای عناصر Nd (نقطه ۱۶)، P (نقطه ۲۰) و Ti (نقطه ۲۰) داده‌های پرت مشاهده شد.

[1] Quantile-quantile plot

نتایج نمودار جعبه‌ای: مطابق با نتایج این نمودار برای عنصر Ba نقاط ۱۵ و ۲۴ پرت محسوب می‌شوند. برای Cd نمودار جعبه‌ای داده پرتی را نشان نمی‌دهد. برای عنصر Co نمودار جعبه‌ای داده پرتی را نشان نمی‌دهد. برای عناصر Cu، Li نقاط پرتی نشان داده نشد. برای عنصر Mn نقاط ۱۱، ۱۵ و ۲۴ به‌عنوان داده‌های پرت شناخته شدند. برای عنصر Nd نمودار جعبه‌ای نقطه ۱۶ را داده پرت معرفی کرد. برای عنصر Ni مطابق با نمودار جعبه‌ای داده پرتی دیده نمی‌شود. برای عنصر P نقطه ۲۰ داده پرت محسوب شده، برای عنصر Si در نمودار جعبه‌ای نقاط ۱۱ و ۱۶ داده پرت محسوب می‌شوند. برای عنصر Sr داده پرتی وجود ندارد. برای عنصر Ti مطابق با نمودار جعبه‌ای نقطه ۲۰ داده پرت تلقی شده است. برای عناصرV، Zn، N و C مطابق با نمودار جعبه‌ای نقطه پرتی دیده نمی‌شود. بنابراین مطابق با این روش برای عناصر Ba (نقاط ۲۴ و ۱۵)، Mn (نقاط ۱۱،۱۵ و ۲۴)، Nd (نقطه ۱۶)، P (نقطه ۲۰)، Si (نقاط ۱۱ و ۱۶) و Ti (نقطه ۲۰) داده‌های پرت وجود دارد.

نتایج آزمون دیکسون[1]: برای عنصر Ba در نقطه ۲۴ نقطه بحرانی شناخته شده است، میزان محاسبه شده معادل ۰/۴۴ است و مقدار قرائت شده از جدول (با توجه به تعداد نقاط نمونه‌برداری شده برای هر ردیاب و سطح اطمینان یک درصد) برابر ۰/۲۳۱ است. بنابراین این نقطه داده پرت محسوب می‌شود. همچنین برای نقطه ۱۵، مقدار محاسبه شده ۰/۳۹ است که از مقدار جدول بزرگ‌تر است و داده پرت محسوب می‌شود. برای عنصر Cd مقادیر بحرانی وجود ندراد و مقدار محاسباتی معادل ۰/۱۸۳ است که کوچک‌تر از مقدار جدول است و داده پرت دیده نمی‌شود. برای عنصر Co نقطه ۲۴ با مقدار ۰/۳۱ داده پرت محسوب می‌شود. برای عناصر Cr، Cu و Li داده پرت وجود ندارد. برای عنصر Mn نقاط ۲۴ و ۱۵ با مقادیر ۰/۳۷ و ۰/۹۹ داده‌های پرت محسوب می‌شوند. برای عنصر Nd نقطه ۱۶ با مقدار ۰/۹۴ پرت محسوب می‌شود. برای عناصر Ni و P داده پرت دیده نمی‌شود. برای عنصر Si نقطه ۱۱ با مقدار ۰/۵۷ داده پرت محسوب می‌شود. برای عنصر Ti نقطه ۲۰ با مقدار ۰/۵۷ که بزرگ‌تر از مقدار جدول یعنی ۰/۲۳۱ است، داده پرتی محسوب می‌شود. همچنین، برای عناصر V، Zn، N و C داده پرتی مشاهده نشد. بنابراین مطابق با این روش عناصر Ba (نقاط ۲۴ و ۱۵)، Co (نقطه ۲۴)، Mn (نقاط ۱۵ و ۲۴)، Nd (نقطه ۱۶)، Si (نقطه ۱۱) و Ti (نقطه۲۰) دارای داده‌های پرت می‌باشند.

نتایج آزمون گوس: مقادیر محاسبه شده برای مقادیر بیشینه در عناصر Ba (۲/۶۹)، Cd (۲/۰۳)، Co (۲/۱)، Cr (۲/۴۴)، Cu (۲/۱۲)، Li (۲/۲۹)، Mn (۲/۷۷)، Nd (۳/۳۴)، Ni (۲/۸۵)، P (۳)، Si (۲/۴۴)، Ti (۳/۲۱)، V (۲/۰۳)، Zn (۲/۲۴)، N (۱/۶) و C (۱/۶۶) محاسبه شد که کلیه مقادیر محاسبه شده کمتر از مقدار جدول برای آزمون گوس یعنی مقدار ۳/۵۳ می‌باشد. بنابراین مطابق با این روش هیچ‌کدام از ردیاب‌ها دارای داده پرت نیستند.

نتایج آزمون میانگین به اضافه و منهای سه برابرانحراف از معیار: مطابق با این روش برای عناصر Nd (نقطه ۳۵)، P (نقطه ۱) و Ti (نقطه ۱) داده پرت وجود دارد و برای بقیه عناصر در هیچ نقطه‌ای داده پرت مشاهده نشد.

نتایج آزمون میانه ۳± برابر MAD: این آزمون نشان داد، برای همه عناصر اندازه‌گیری شده دارای مقادیر پرت هستند و تعداد بیشتری از عناصر را در نقاط مختلف پرت تلقی کرد. نتایج حاصل از این روش در جدول زیر (جدول ۱) آمده است. در این روش کمترین داده‌های پرت مربوط به عناصر نئودیمیوم و نیتروژن هست.

همان‌گونه که در جدول ۱ مشاهده می‌شود، برای عنصر Ti نقطه ۲۰ به‌وسیله چهار روش پرت معرفی شده است. برای عنصر Si نقطه ۱۱ به‌وسیله سه روش نقطه پرت معرفی شده است. برای عنصر P نقطه ۲۰ به‌وسیله سه روش نقطه پرت معرفی شده است. برای عنصر Nd نقطه ۱۶ به‌وسیله چهار روش نقطه پرت معرفی شده است و برای عنصر Ba نقاط ۱۵ و ۲۴ به‌وسیله سه روش به‌عنوان نقطه پرت و داده پرت معرفی شده است.

مقایسه مقادیر مربع فاصله ماهالانوبیس بروی درجه آزادی (MD^2/df) با مقدار t مشاهداتی از جدول

[1] Dixon

t استیودنت برای درجه آزادی ۱۶ در سطح اطمینان ۰/۰۱ (مقدار قرائت شده برابر ۲/۵۸ بود)، نشان داد همه داده‌ها کمتر از مقدار جدول هستند و هیچ داده‌ای مطابق با این روش پرت محسوب نمی‌شود (درجه آزادی تعداد ردیاب‌ها منهای یک و سطح اعتماد داده‌ها یک درصد می‌باشد). برای بررسی بیشتر نمودار جعبه‌ای مربع فاصله ماهالانوبیس برای واحد کاربری اراضی نقطه ۲۲ پرت تشخیص داده شده است (شکل ۳).

نمودار چندک‌چندک مربع فاصله ماهالانوبیس که در مقابل توزیع مربع کای مرتب شده در شکل ۴ آورده شده است. طبق شکل نشان داده شده هیچ نمونه یا داده‌ای پرت تشخیص داده نشده است. پراکنش تمام نقاط (نمونه‌ها) از یک روند خاص پیروی می‌کند و انحراف قابل توجهی مشاهده نمی‌شود.

نتایج تجزیه تحلیل مؤلفه‌های اصلی نشان داد که نمونه‌های شماره ۴۳، ۱ و ۳۵ پرت هستند و فاصله DSC آن‌ها از بقیه نمونه‌ها بیشتر است. در تحلیل مؤلفه‌های اصلی تعداد چهار مؤلفه اصلی بیش از ۹۹ درصد از واریانس متغیرها (ردیاب‌ها) را بیان کردند. هر چه فاصله نمرات نمونه بزرگ‌تر باشد، امکان پرت‌بودن داده بیشتر است. همچنین، نمودار جعبه‌ای فاصله نمرات نمونه‌ها نیز برای تشخیص داده پرت ترسیم شد. در این نمودار برای فواصل نمونه‌ها هیچ داده‌ای پرت تشخیص داده نشد (شکل ۶). بنابراین، مقادیر مشکوک به پرت بودن مطابق با پراکنش فواصل نمرات نمونه پرت نیستند (شکل ۵).

برای تشخیص داده پرت با استفاده از فاصله ماهالانوبیس از جدول خی دو (مربع کای) در سطح احتمال یک درصد و با درجه آزادی ۱۶ (مقدار قرائت شده ۵/۸۱) استفاده شد. نتایج مقایسه فاصله ماهالانوبیس و مقدار مربع کای در جدول ۳ ارائه شده است. همان‌طور که نتایج بالا نشان داد، روش چندمتغیره تحلیل مؤلفه‌های اصلی هیچ داده‌ای را پرت معرفی نکرد. همچنین روش مربع فاصله ماهالانوبیس نیز داده‌ای را پرت نشان نداد. نمودار چندک‌چندک مربع فاصله ماهالانوبیس در مقابل توزیع مربع کای نیز داده‌ای را پرت نشان نداد. نمودار جعبه‌ای مربع فاصله ماهالانوبیس برای هر منبع رسوب، تنها نقطه ۲۲ را به‌عنوان داده پرت نشان داد. همچنین روش فاصله ماهالانوبیس نیز این نقطه را به‌عنوان داده پرت معرفی نمود. بنابراین همان‌طور که مشاهده می‌شود، تنها دو روش یاد شده بر پرت بودن نقطه ۲۲ اجماع دارند و بقیه روش‌ها چنین استدلالی ندارند.

همان‌گونه که مشاهده شد، نقطه ۲۲ تنها به‌وسیله روش میانه به‌اضافه یا منهای میانه انحراف‌های تمام داده‌ها از میانه پرت شناخته شد و بقیه روش‌های تک-متغیره آن‌را پرت تشخیص ندادند. با این حساب تنها سه روش از مجموع ۱۱ روش به‌کار رفته بر پرت بودن این نقطه تأکید دارند و برای بقیه نقاط هم تنها چهار روش از مجموع ۱۱ روش به‌کار رفته بر پرت بودن داده‌ها دلالت دارند که این حکایت از عدم توافق ۵۰ درصد بر پرت بودن داده‌هاست.

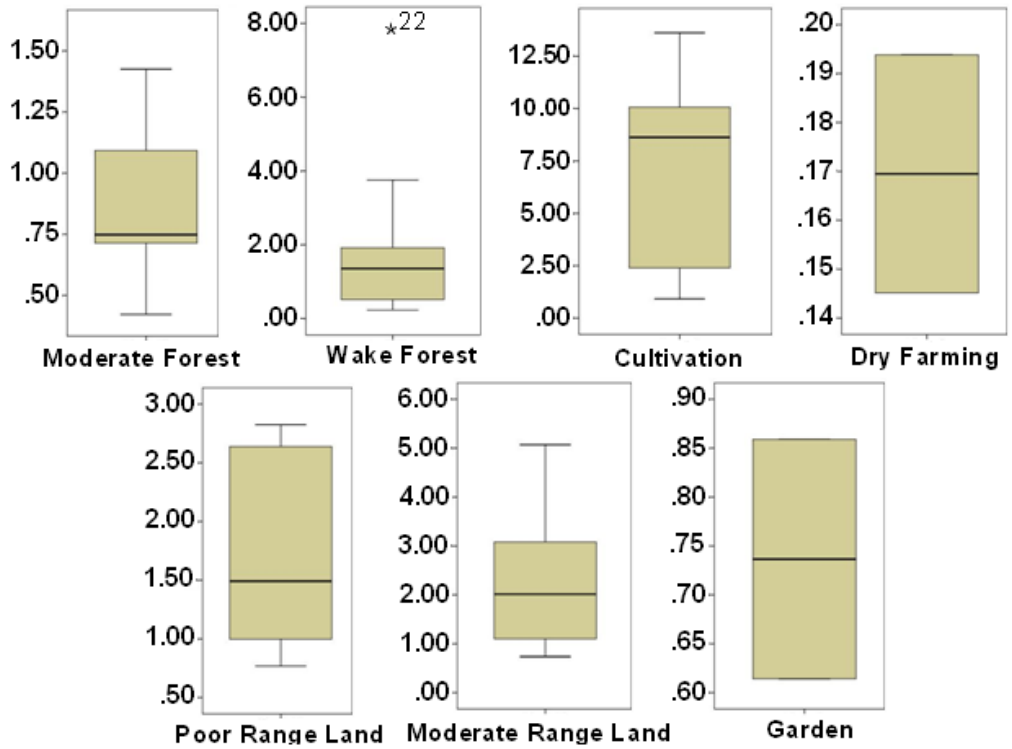
همان‌طور که مشاهده می‌شود، در کل هیچ داده پرتی وجود ندارد و همه اندازه‌گیری‌ها و مقادیر صحیح و قابل استناد برای ورود به مدل‌های ترکیبی و مراحل بعد هستند. با توجه به این که روش‌های آزمون گراب، گوس و میانگین به‌اضافه و منهای سه برابر انحراف معیار تحت تأثیر داده‌های پرت و غیرعادی (داده‌های بسیار کوچک و بسیار بزرگ) می‌باشد (چون میانگین و انحراف معیار با استفاده از تمام داده‌ها برآورد می-شوند) (Reimann و همکاران، ۲۰۰۵) کارایی زیادی در تشخیص داده‌های پرت و غیرعادی ندارند (Hakimkhani و Alijanpour، ۲۰۱۰). جدول ۱ نیز این نتیجه را تأیید می‌کند، به‌طوری که آزمون گوس و میانگین به‌اضافه و منهای سه برابر انحراف از معیار، هیچ نمونه‌ای را در مقایسه با آزمون گراب پرت معرفی نکرده و حساسیت آزمون گراب در مقایسه با این دو روش به مراتب بیشتر است. روش‌های میانه به‌اضافه یا منفی میانه انحراف‌های تمام داده‌ها از میانه و نمودار جعبه‌ای تحت تأثیر داده‌های پرت نبوده، می‌توانند تعداد بیشتری نمونه پرت را شناسایی کنند (Chiang و همکاران، ۲۰۰۳). آزمون دیکسون تعداد داده‌های پرت بیشتری را در مقایسه با آزمون‌های گوس، گراب و میانگین به‌اضافه و منهای سه برابر انحراف از معیار، نشان می‌دهد.

جدول ۱- نتایج حاصل از بررسی داده‌های پرت با روش‌های تک‌متغیره

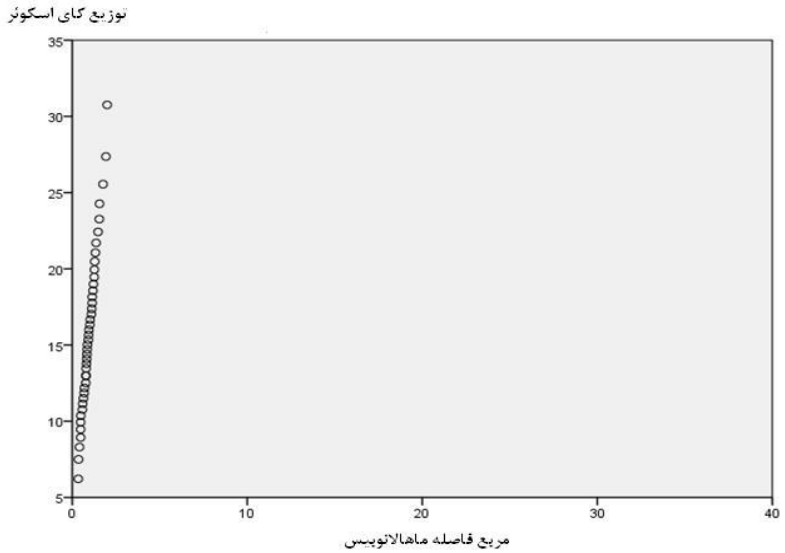
عنصر	میانگین	انحراف از معیار	میانه	میانه انحراف‌های تمام داده‌ها از میانه	آزمون گراب	نمودار جعبه‌ای مقادیر استاندارد	آزمون گوس	آزمون دیکسون	آزمون میانگین به اضافه یا منهای سه برابر انحراف از معیار	میانه به اضافه یا منهای میانه انحراف‌های تمام داده‌ها از میانه
باریم	۱۶/۵۷	۸/۳	۱۴/۰۳	۶/۹۸	-	۱۵ و ۲۴		۱۵ و ۲۴		۱، ۷، ۱۰، ۱۱، ۲۲،۲۰،۱۵، ۲۴، ۲۶، ۲۸، ۳۴، ۲۷
کادمیوم	۰/۴۴۴	۰/۲۷۹	۰/۳۸	۰/۳۴	-					۴، ۵، ۱۳، ۲۱، ۲۳، ۳۱
کبالت	۳/۳۲۹	۱/۷۶۹	۳/۳۳	۲/۳۴	-			۲۴		۱، ۱۵، ۲۴، ۲۵، ۳۴، ۳۲
کروم	۱۵/۷۶۷	۷/۱۸۴	۱۵/۵۸	۷/۴۲	-					۱، ۵، ۶، ۱۵، ۱۷، ۳۴، ۱۹، ۲۳، ۲۵، ۲۷، ۳۲، ۳۳
مس	۶/۳۵۳	۳/۰۹۶	۶/۱	۳/۴	-					۱۵، ۲۸، ۳۳، ۳۴، ۳۵، ۴۳، ۱۹، ۲۳، ۲۷، ۴۰
لیتیم	۳/۴۶۴	۱/۸۱۷	۳/۴۴	۲/۲۶	-					۱۵، ۲۶، ۳۵، ۴۳، ۱۹، ۲۳، ۲۵، ۳۲، ۴۰
منگنز	۱۵۶/۴۷	۹۳/۷۳۷	۱۲۰/۳	۷۶/۶۱	-	۱۵، ۱۱ و ۲۴		۱۵ و ۲۴		۱، ۶، ۱۱، ۱۵، ۲۰، ۲۲، ۲۴، ۲۶، ۲۸، ۲۷، ۳۴
نئودیوم	۳/۰۸۴	۱/۵۶	۳/۱۶	۱/۹۷	۱۶	۱۶		۱۶	۳۵	۱۶، ۱۰، ۲۰
نیکل	۲۶/۱۲۷	۱۳/۹۳۵	۲۵/۶۸	۱۶/۱۹	-					۱، ۳۴، ۳۵، ۳۷، ۴۳، ۲۵، ۲۷، ۳۲، ۳۳
فسفر	۱۲۹/۲۱۴	۸۳/۷۳۳	۱۰۳/۶	۷۹/۵۲	۲۰	۲۰			۱	۱۱، ۱۵، ۱۸، ۲۰، ۲۴، ۲۶، ۲۸، ۳۴، ۳۵، ۳۶، ۴۳، ۲۵
سیلیکون	۳/۴۹۳	۱/۹۷۳	۳/۱۳	۱/۷	-	۱۱ و ۱۶		۱۱		۶، ۱۱، ۱۶، ۱۸، ۲۰، ۲۱، ۲۵، ۳۲، ۳۳، ۳۵، ۳۸، ۴۳
استرانسیوم	۳۷/۷۲۶	۲۴/۲۱۷	۲۵/۶۱	۱۸/۲۴	-					۱۴، ۱۷، ۱۹، ۲۱، ۲۳، ۲۵، ۳۲، ۳۳، ۳۴، ۳۷، ۳۸، ۳۹، ۴۰، ۴۲
تیتانیوم	۸/۷۰۵	۵/۰۹۲	۷/۵۸	۵/۱۴	۲۰	۲۰		۲۰	۱	۲، ۶، ۱۷، ۱۸، ۲۰، ۲۵، ۲۶، ۲۷، ۲۹، ۳۰، ۳۴، ۳۶، ۴۱
وانادیوم	۱۰/۷۳۷	۴/۹۳	۹/۹۹	۵/۵۲	-					۱، ۶، ۱۱، ۱۵، ۲۴، ۳۴، ۳۵، ۴۳، ۲۳، ۲۷، ۳۲، ۳۳
زینک	۱۴/۳۵۶	۷/۱۷۱	۱۳/۰۵	۶/۶۳	-					۱۱،۱۵، ۱۸، ۲۰، ۲۲، ۲۴، ۲۶، ۲۸، ۳۴، ۳۵، ۴۳، ۲۵
نیتروژن	۰/۰۸	۰/۰۲۴	۰/۰۸۱۹	۰/۰۲۹	-					۴، ۱۱، ۱۳، ۲۲، ۲۴، ۲۶، ۲۷، ۲۹، ۳۷، ۴۱
کربن	۲۷/۰۷۶	۸/۰۸	۲۷	۱۱/۱۱	-					۴، ۱۱، ۱۳، ۱۷

جدول ۲- مربع فاصله ماهالانوبیس و مربع فاصله ماهالانوبیس به‌روی درجه آزادی برای تشخیص داده‌های پرت

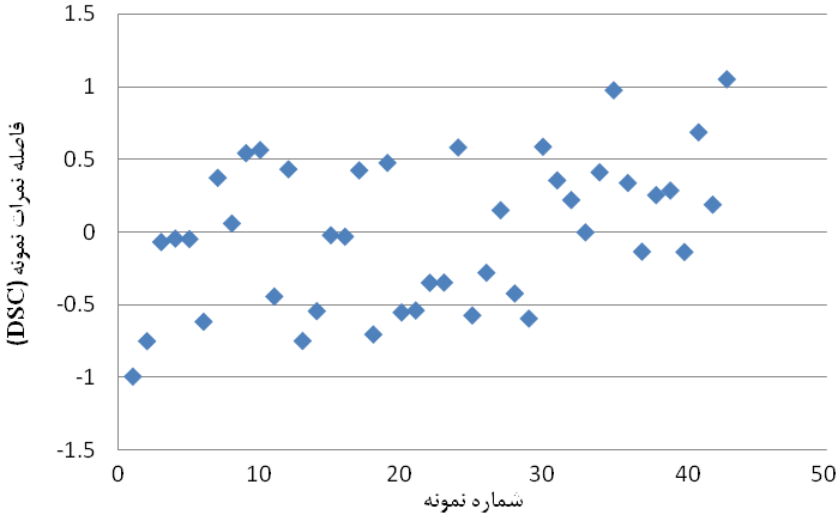
شماره نمونه	MD^2	MD^2/df	شماره نمونه	MD^2	MD^2/df
۱	۱/۵۸۴	۰/۰۹۳۲	۲۳	۱/۰۰۶	۰/۰۵۹۱
۲	۰/۲۳۲	۰/۰۱۳۶	۲۴	۱/۳۵۶	۰/۰۷۹۸
۳	۰/۵۴	۰/۰۳۱۸	۲۵	۲/۰۱	۰/۱۱۸۴
۴	۱/۹۱۶	۰/۱۱۲۷	۲۶	۱/۲۲۸	۰/۰۷۲۲
۵	۰/۶۲۴	۰/۰۳۶۷	۲۷	۱/۱۹۱	۰/۰۷
۶	۰/۸۵۸	۰/۰۵۰۵	۲۸	۱/۷۵۳	۰/۱۰۳۱
۷	۰/۴۲۱	۰/۰۲۴۷	۲۹	۰/۵۱۴	۰/۰۳۰۲
۸	۰/۷۶۷	۰/۰۴۵۱	۳۰	۰/۷۳۶	۰/۰۴۳۳
۹	۰/۹۹۷	۰/۰۵۸۶	۳۱	۱/۴۲۴	۰/۰۸۳۸
۱۰	۰/۶۱۴	۰/۰۳۶۱	۳۲	۰/۹۱۸	۰/۰۵۴
۱۱	۱/۶۹۹	۰/۰۹۹۹	۳۳	۹/۶۹۱	۰/۵۷۰۱
۱۲	۰/۹۹۲	۰/۰۵۸۳	۳۴	۲۱/۳۹۵	۰/۱۴۰۹
۱۳	۰/۷۴۷	۰/۰۴۳۹	۳۵	۹/۰۰۸	۰/۵۲۹۹
۱۴	۲/۶۳۷	۰/۱۵۵۱	۳۶	۰/۶۹۱	۰/۰۴۰۶
۱۵	۲/۶۳۷	۰/۱۵۵۱	۳۷	۳/۷۵۵	۰/۲۲
۱۶	۰/۲۹۴	۰/۰۱۷۳	۳۸	۱۳/۶۱	۰/۸۰۰۸
۱۷	۰/۱۹۳	۰/۰۱۱۴	۳۹	۱۰/۰۶	۰/۵۹۱۹
۱۸	۰/۱۴۵	۰/۰۰۸۵	۴۰	۳/۳۰۷	۰/۱۹۴۵
۱۹	۰/۷۷۸	۰/۰۴۵۷	۴۱	۵/۰۷۵	۰/۲۹۸۳
۲۰	۲/۸۵۳	۰/۱۶۷۸	۴۲	۲/۸۲۳	۰/۱۶۶۱
۲۱	۰/۴۲۶	۰/۰۲۵۱	۴۳	۷/۵۶۱	۰/۴۴۴۸
۲۲	۷/۹	۰/۴۶۴۷			



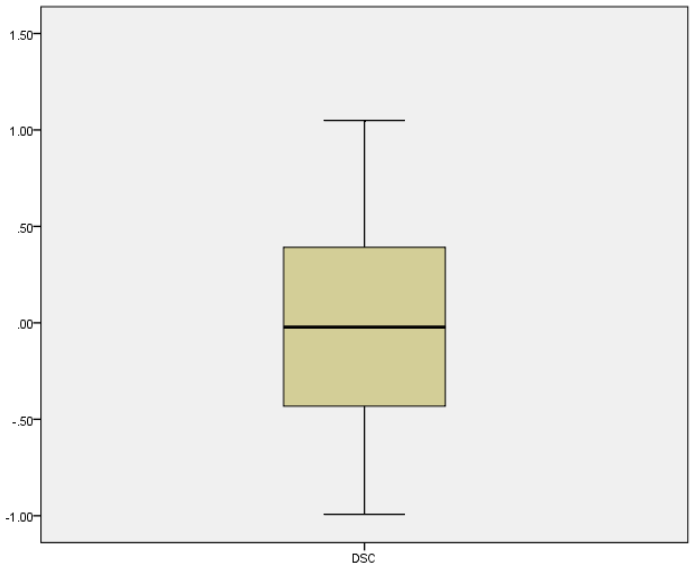
شکل ۳- نمودار جعبه‌ای مربع فاصله ماهالانوبیس برای منبع‌های زراعت آبی، جنگل ضعیف، جنگل متوسط، باغات، مرتع ضعیف، مرتع متوسط و دیم‌زار



شکل ۴- پراکنش مقادیر مربع فاصله ماهالانوبیس در مقابل مقادیر مورد انتظار توزیع مربع کای



شکل ۵- پراکنش فواصل نمرات نمونه (روش تجزیه مؤلفه‌های اصلی)



شکل ۶- نمودار جعبه‌ای فواصل نمرات نمونه‌ها

جدول ۳- داده‌های پرت مطابق با روش فاصله ماهالانوبیس

نام عنصر	نقاط پرت	میزان فاصله ماهالانوبیس داده پرت
Ba	۳۵	۷/۲۴۱
Cd	-	-
Co	-	-
Cr	۴۰	۵/۹۵
Cu	-	-
Li	۲۲	۵/۱۹۸
Mn	۴۳	۷/۶۷۷
Nd	۳۳	۱۱/۱۶۶
Ni	-	-
P	۳۸ و ۳۹	۸/۹۹۸ و ۷/۰۴
Si	۳۳ و ۴۳	۵/۹۳۱ و ۵/۸۵۷
Sr	-	-
Ti	۳۸	۱۰/۳
V	-	-
Zn	-	-
N	-	-
C	-	-

قابل ذکر است که این روش بر اساس مقادیر حدی و مقدار فاصله آن‌ها از عمده داده‌ها کار می‌کند. همان‌طور که مشاهده شد دو روش باقی‌مانده از روش-های تک‌متغیره یعنی نمودارهای جعبه‌ای و روش میانه به‌اضافه یا منهای میانه انحراف‌های تمام داده‌ها از میانه تعداد قابل توجهی از نمونه‌ها را پرت نشان دادند. خصوصاً روش میانه به‌اضافه یا منهای میانه انحراف‌های تمام داده‌ها از میانه. بنابراین این دو روش کارایی بیشتری نسبت به چهار روش قبل دارند.

برای استفاده از روش‌های تک‌متغیره به‌منظور شناسایی داده‌های پرت روش‌های میانه به‌اضافه یا منهای میانه انحراف‌های تمام داده‌ها از میانه، نمودار جعبه‌ای و آزمون دیکسون به‌ترتیب پیشنهاد می‌شود. Reimann و همکاران (۲۰۰۵) نیز نتایج مشابهی در این زمینه دست یافتند. بررسی چندمتغیره داده‌های پرت به‌روش چندک‌چندک مربع فاصله ماهالانوبیس در برابر مربع کای، مربع فاصله ماهالانوبیس در مقایسه با توزیع t استیودنت نشان داد، هیچ داده‌ای پرت نیست. همچنین روش تحلیل مؤلفه‌های اصلی تنها سه نقطه را به‌عنوان داده پرت نشان داد و نمودار جعبه‌ای فاصله ماهالانوبیس در منابع تولید رسوب تنها یک نقطه را به‌عنوان داده پرت نشان داد.

Hair و همکاران (۱۹۹۸) در مورد مربع فاصله ماهالانوبیس، Garrett (۱۹۸۹) و Filzmoser و همکاران (۲۰۰۵) در مورد نمودار چندک‌چندک به نتایج مشابهی دست یافتند. قابل ذکر است که آزمون‌های تک‌متغیره همچون گراب و گوس در تعداد نقاط برداشت شده برای هر ردیاب محدودیت دارند، در حالی‌که این محدودیت تعداد داده برای آزمون‌های چندمتغیره وجود نداشت. روشی همچون فاصله ماهالانوبیس اگرچه روشی مناسب برای تشخیص داده‌های پرت است، اما از آنجا که در به‌دست آوردن فاصله داده‌های پرت نیز مؤثرند این داده‌ها اثر منفی بر الگوریتم می‌گذارند. همچنین، قابل ذکر است که در کل با توجه به این که هیچ یک از نمونه‌ها از نظر اجماع روش‌های یک‌متغیره 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 یاد شده 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.

نمونه‌هایی که در بعضی از روش‌های مورد استفاده به‌عنوان پرت شناسایی شده‌اند، در واقع جزء 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 جز و یک یا چند توزیع متفاوت هستند (Reimann و همکاران، ۲۰۰۵؛ Filzmoser و همکاران، ۲۰۰۵). در واقع فرایند یا فرایندهای ایجاد کننده داده‌های پرت متفاوت از داده‌های اصلی است. هم آن‌طور که گفته شد داده‌های پرت در ژئوشیمی علاوه بر اشتباه یا خطا، اغلب ناشی از فرایندهای کانی‌سازی، آلتراسیون و فعالیت‌های انسانی (از جمله کودپاشی مزارع) است. با توجه به نتایج حاصل از تحقیق حاضر پیشنهاد می‌شود در زمینه منشأیابی رسوب و استفاده از مدل‌های ترکیبی در مشخص کردن سهم منابع رسوب و در جهت عدم ورود داده‌های پرت به این روند، از اجماع چندین روش شناسایی داده‌های پرت استفاده شد و تنها یک شاخص یا یک آزمون ملاک عمل قرار نگیرد. همان‌گونه که قبلاً نیز اشاره شد، تمامی روش‌های شناسایی داده‌های پرت مقبولیت جهانی ندارند و نیاز است که حصول پرت بودن مشاهده‌ای از چندین روش مختلف آشکار شود. با توجه به جمیع مطالب بالا می‌توان گفت، یک داده پرت مشاهده‌ای است که به‌طور غیرعادی و یا اتفاقی از وضعیت کلی داده‌های تحت آزمایش و نسبت به قاعده‌ای که بر اساس آن تحلیل می‌شوند، تفاوت معنی‌داری داشته باشند و چندین روش بر انحراف آن از وضعیت عمومی داده‌ها اذعان داشته باشند.

منابع مورد استفاده

1. Aliehyai, M. and A.A. Behbahanizade. 1993. Methods of chemical analysis of soil. Institute of Soil and Water Research, Bulletin No 893, 26 pages (in Persian).
2. Alfassi, Z.B., Z. Boger and Y. Statistical. 2005. Treatment of analytical data: outliers (Chapter 6). CRC Press, 512 pages.
3. Caussinus, H., M. Fekri, S. Hakam and A. Ruiz-Gazen. 2003. A monitoring display of multivariate outliers. Computational Statistics and Data Analysis, 44: 237-252.
4. Chiang, L.H., R.J. Pell and M.B. Seasholtz. 2003. Exploring process data with the use of robust outlier detection algorithms. Journal of Process Control, 13: 437-449.
5. Collins, A.L., D.E. Walling, L. Webb and P. King. 2010. Apportioning catchment scale sediment sources using a modified composite fingerprinting technique incorporating property weightings and prior information. Geoderma, 155: 249–261.
6. Collins, A.L. and D.E. Walling. 2007. Sources of fine sediment recovered from the channel bed of lowland groundwater-fed catchments in the UK. Geomorphology, 88: 120-138.
7. Collins, A.L., Y. Zhang, D. McChesney, D.E. Walling, S.M. Haley and P. Smith. 2012. Sediment source tracing in a lowland agricultural catchment in southern England using a modified procedure combining statistical analysis and numerical modelling. Science of the Total Environment, 414: 301-317.
8. EPA. 2006. Data quality assessment: statistical methods for practitioners EPA QA/G-9S, EPA/240/B-06/003, U.S. Environmental Protection Agency, Office of Environmental Information, Washington DC.
9. Feng, J., Z.G. Hu, J.T. Ju and L.P. Zhu. 2011. Variations in trace element (including rare earth element) concentrations with grain sizes in loess and their implications for tracing the provenance of eolian deposits. Quaternary International, 236: 116-126.
10. Filzmoser, P., R.G. Garrett and C. Reimann. 2005. Multivariate outlier detection in exploration geochemistry. Computers and Geosciences, 31: 579-587.
11. Garrett, R.G. 1989. The chi-square plot: a tool for multivariate outlier recognition. Journal of Geochemical Exploration, 32: 319-341.
12. Hakimkhani, Sh. and A. Alijanpour. 2010. Detection of outliers in the sediment fingerprinting method. Water and Soil Conservation, 17(1): 23-43 (in Persian).
13. Hair, J.F., R.E. Andersen, R.L. Tatham and W.C. Black. 1998. Multivariate data analysis. Rentice Hall, Upper Saddle River, New Jersey.

14. Szalma, J. 1984. Mérési eredmények kiértékelésének alapjai (Introduction to evaluation of measured data). Tankönyvkiadó, Budapest.
15. Reimann, C., P. Filzmoser and R.G. Garrett. 2005. Background and threshold: critical comparison of methods of determination. Science of the Total Environment, 346: 1-16.
16. Rousseeuw, P.J. and B.C. Van Zomeren. 1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85: 633-651.
17. Rousseeuw, P.J. and K. Van Driessen. 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41: 212-223.
18. Nosrati, K., G. Govers, H. Ahmadi, F. Sharifi, M.A. Amoozegar and R. Merckx. 2011. An exploratory study on the use of enzyme activities as sediment tracers: biochemical fingerprints? International Journal of Sediment Research, 26(2): 136-151.
19. Rorabacher, D.B. 1991. Statistical Treatment for rejection of deviant values: critical values of dixon Q parameter and related subrange ratios at the 95 percent confidence level. Anal. Chem, 83(2): 139-146.
20. Tabachnick, B.G. and L.S. Fidell. 1996. Using multivariate statistics. Harper Collins College Publishers, New York, 963 pages.
21. Tukey, J.W. 1977. Exploratory data analysis. Addison-Wesley Publication, 205-235.
22. Lalor, G.C. and C. Zhang. 2001. Multivariate outlier detection and remediation in geochemical databases. The Science of the Total Environment, 281: 99-109.
23. Walling, D.E., A.L. Collins and R.W. Stroud. 2008. Tracing suspended sediment and particulate phosphorus sources in catchments. Journal of Hydrology, 350: 274-289.
24. Walling, D.E. 2005. Tracing suspended sediment sources in catchments and river systems. Science of the Total Environment, 344: 159–184.
25. Walling, D.E., P.N. Owens and G.J.L. Leeks. 1999. Fingerprinting suspended sediment sources in the catchment of the River Ouse, Yorkshire, UK. Hydrological Processes, 13: 955-975.
26. Wiegand, P., R. Pell and E. Comas. 2009. Simultaneous variable selection and outlier detection using a robust genetic algorithm. Chemo metrics and Intelligent Laboratory Systems, 98(2): 108-114.
27. Zhang, C., D. Fay, D. McGrath, E. Grennan and O.T. Carton. 2008. Statistical analyses of geochemical variables in soils of Ireland. Geoderma, 146: 378-390.
28. Zhang, C.S., P.M. Wong and O. Selinus. 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.

Using univariate and multivariate methods to detect outliers in sediment fingerprinting method, case study: Tange Bostanak Watershed

Ahmad Nohegar[1], Mohamad Kazemi*[2], Seyed Javad Ahmadi[3], Hamid Gholami[4], Rasool Mahdavi[5]
[1] Professor, Faculty of Environment Sciences, Tehran University, Iran, [2] PhD Student, Faculty of Natural Resources, Hormozgan University, Iran, [3] Associate Professor, Fuel Cycle Research Institute of Atomic Energy Organization, Iran and [4 and 5] Assistant Professor, Faculty of Natural Resources, Hormozgan University, Iran

Received: 17 September 2016 Accepted: 27 February 2017

Abstract

Efficiency of sediment fingerprinting by using tracers as a successful method to determine the sources of sediment has been proved. 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. The presence of outliers affects the selection of the suite subset and possibly prevents picking the important tracers and reducing the accuracy of classification. Therefore, the outliers must be detected in order to be corrected or omitted, if enough evidences were present. The present study aims to detect outliers in the subset of tracers, to identify the best combination. For detecting outliers, We used univariate methods such as Grubbs test, Gauss test, Dioxin test, box plot, the Median ± 3MAD, the mean ± 3standard deviation and also multivariate methods such as 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. we consider an observation as the outlier that at least half of these methods have detected it as an outlier. The results showed that Median ± 3MAD method introduced a larger number of data as outliers Methods of multivariate outlier detection has low agreement with each other. Univariate methods to identify outliers show higher agreement with each other. To use univariate analysis techniques to detect outliers namely Median ± 3MAD, box plot, and Dioxin one can recommended to test their sensitivity. The results also showed that the maximum consensus for univariate analysis techniques is four samples (observations) and for multivariate methods is two samples (observations). In general, there is no observation that is identified as an outlier by half of the used methods.

Key words: Grubbs' Test, Mahalanobis Distance, PCA, Sediment Fingerprinting, Tracers

* Corresponding author: mohamad.kazemi86@gmail.com