مرتع و آبخیزداری، مجله منابع طبیعی ایران
دوره ۷۵، شماره ۴، زمستان ۱۴۰۱
ص ۶۶۰-۶۳۹

تاریخ دریافت: ۱۴۰۱/۰۸/۱۶
تاریخ تصویب: ۱۴۰۱/۱۰/۲۹

مقایسه روش‌های شناسایی داده‌های پرت و تاثیر آنها در مطالعات اندازه‌گیری و ارزیابی مراتع

❖ **مسلم رستم‌پور***؛ استادیار گروه مرتع و آبخیزداری و عضو گروه پژوهشی خشکسالی و تغییر اقلیم، دانشکده منابع طبیعی و محیط زیست، دانشگاه بیرجند، بیرجند، ایران

چکیده

این تحقیق به مقایسه روش‌های شناسایی داده پرت تک متغیره در بین داده‌های درصد پوشش گیاهی در یک مطالعه ارزیابی تاثیر شدت چرا در مراتع مناطق خشک می‌پردازد. بدین منظور، پس از اندازه‌گیری درصد پوشش گیاهی در مرتع و قبل از تحلیل آماری، وجود یا عدم وجود داده پرت به عنوان پیش فرض آزمون‌های پارامتریک فرضیه مقایسه‌ای بررسی شد. در این تحقیق از هشت روش شامل نمودار جعبه‌ای (Boxplot) و دامنه میان چارکی (روش Tukey)، انحراف معیار از میانگین (قانون Three-sigma)، انحراف مطلق از میانه (روش Hampel)، میانگین پیراسته، مقادیر صدک ۱ و ۹۹، آزمون کای اسکوئر (χ^2)، آزمون گرابز (ESD) و آزمون روزنر (generalised ESD) استفاده شد. نتایج نشان داد که داده‌های درصد پوشش گیاهی مراتع با شدت چرای سبک و متوسط توزیع نرمال ندارند (آزمون شاپیرو-ویلک: $P\leq$ ۰/۰۵). حتی حذف داده پرت نیز منجر به نرمال شدن داده‌ها نشد، اما منجر به همگن شدن واریانس خطا شد (آزمون لیون: $P\geq$ ۰/۰۵). از هشت روش مورد استفاده، روش Z اصلاح شده و آزمون‌های گرابز و روزنر ($P\geq$ ۰/۰۵)، هیچکدام از داده‌های درصد پوشش گیاهی را به عنوان داده پرت تشخیص ندادند. از بین روش‌های مورد مطالعه، نمودار جعبه‌ای و روش انحراف مطلق از میانه که به میانگین وابسته نیستند، برای داده‌های پوشش گیاهی مناسب‌ترند. از این‌رو قبل از انجام هرگونه آزمون فرضیه مقایسه‌ای، استفاده ترکیبی از دو روش چشمی و آماری برای بررسی وجود یا عدم وجود داده‌های پرت توصیه می‌شود.

کلمات کلیدی: آمار پارامتری، پوشش گیاهی، داده‌های دورافتاده، میانگین، مرتع.

* نویسنده مسئول: شماره تماس: ۰۹۱۵۱۶۳۷۸۶۹
Email: rostampour@birjand.ac.ir

۱. مقدمه

آمار و اکولوژی کمّی دو جزء جدایی‌ناپذیر آنالیز و ارزیابی مراتع و اکولوژی گیاهی هستند [۳۳، ۲۶]. یکی از اهداف مهم تحقیقات حوزه علوم مرتع، بررسی اثر تیمارهای مختلف اعم از طبیعی، انسانی، دامی و... بر خصوصیات پوشش گیاهی مراتع است [۸]. برای تحلیل داده‌های چنین تحقیقاتی از آزمون فرضیه مقایسه‌ای (اعم از تک متغیره و چند متغیره) استفاده می‌شود. چنین آزمون‌هایی بر پایه مقیاس داده‌ها (فاصله‌ای، نسبتی، رتبه‌ای و اسمی)، وجود یا عدم وجود داده‌های پرت و وضعیت نرمال بودن داده‌ها، به دو دسته آزمون های پارامتری و ناپارامتری تقسیم می‌شود [۲۹، ۵۰]. آزمون فرضیه‌های آماری مقایسه‌ای پارامتری مستلزم تایید فرض‌های زیربنایی همچون تصادفی بودن نمونه‌برداری، استقلال خطاها، همگن بودن واریانس خطاها، نرمال بودن داده‌ها و عدم وجود داده‌های پرت می‌باشد [۱۴]. نادیده گرفتن این پیش فرض‌ها، نه تنها کاربرد آزمون مربوطه را زیر سوال می‌برد، بلکه منجر به تولید نتایج نادرست و گاهی متناقض نیز می‌شود [۴۹].

پوشش گیاهی یکی ازشاخص‌های مهم در اندازه‌گیری و ارزیابی مراتع است و در مطالعات وسیعی استفاده شده است و نقشی تعیین کننده در ارزیابی ساختار و عملکرد مراتع دارند [۱۸]. عامل درصد پوشش گیاهی یکی از فاکتورهای مهم در تعیین شایستگی مرتع [۴]، وضعیت مرتع [۱]، سلامت مرتع [۳۸] و تولید علوفه [۶] است. در مناطق خشک و بیابانی که داده‌های پوشش بسیار پراکنده و تغییرات زمانی و مکانی خصوصیات پوشش گیاهی بسیار متغیر است [۳۳] وجود داده‌های پرت در بین داده‌های پوشش گیاهی محتمل است، از این رو آگاهی از وجود داده‌های پرت، قبل از هر فرضیه آزمایی مقایسه‌ای ضروری است.

داده‌های پرت، داده‌های هستند که طور قابل توجهی از سایر اعضای آن جمعیت انحراف دارند که نشان می‌دهد این عدد یا اعداد توسط روش یا سازوکار دیگری تولید شده باشند [۳]. داده‌های پرت یا نقاط دورافتاده در منابع علمی تحت عنوان پرت یا دورافتاده[1]، غیرعادی[2]، ناهماهنگ[3]، ناموزون[4]، آلوده[5] و یا نابجا[6] بکار رفته است. بررسی وجود داده‌های پرت به دلیل اثرات زیاد آنها بر برآوردگرهای آماری همواره مورد توجه آماردانان بوده است و روش‌های گوناگونی برای تشخیص این داده‌ها ارایه شده است [۲۱]. بسیاری از داده‌ها مانند داده‌های اقتصادی، روانشناسی، علوم اجتماعی، علوم پزشکی، و مهندسی با مشکل وجود نقاط دورافتاده در مجموعه داده‌ها مواجه هستند [۴۲]. Aguinis و همکاران (۲۰۱۳) [۲] با مطالعه ۴۶ مقاله علمی، ۱۴ تعریف از داده‌های پرت ارائه کردند. این محققین، با بررسی حدود ۲۳۲ مقاله در مجلات علمی، ۳۹ روش آماری برای تشخیص داده‌های پرت گزارش کردند. Wang و همکاران (۲۰۱۹) [۴۸] با مرور مقالات مربوط به روش‌های تشخیص داده‌های پرت، مزایا و معایب روش‌های رایج تشخیص داده‌های پرت را بررسی کردند. Mowbray و همکاران (۲۰۱۹) [۳۵] روش‌های هیستوگرام، نمودار جعبه‌ای، دامنه میان چارکی و Z را به عنوان تکنیک‌های رایج شناسایی داده‌های پرت تک متغیره و روش‌های حذف، جایگزینی و تبدیل داده‌های پرت را بررسی کردند. درخصوص روش‌های تشخیص داده‌های پرت در حوزه‌های مختلف علوم مثل اکولوژی [۷]، آلودگی محیط زیست [۴۷]، پایش محیط زیست [۲۰] و کشاورزی [۳۴] تحقیقاتی انجام شده است.

داده‌های پرت ممکن است نتایج آنالیز آماری را به

[1] Outlier
[2] Extreme
[3] Rogue
[4] Discordant
[5] Contaminant
[6] Aberrant

مقدار زیادی تحت تاثیر قرار دهد، به طوری که حذف این عدد از مجموعه داده‌ها نتایج کاملاً متفاوتی به دست بیاورد [۴۱] و به همین دلیل شناسایی و تشخیص داده‌های پرت و بررسی اثر آن‌ها بر جنبه‌های مختلف یک تحلیل آماری برای یک تحلیلگر از اهمیت ویژه‌ای برخوردار است [۲۱]. تنها یک دادهٔ پرت ممکن است بتواند کل نتایج تحلیل آماری را بی‌اعتبار کند [۱۴]. در تحلیل‌های آماری تشخیص داده‌های پرت به دو دلیل از اهمیت زیادی برخوردار است: ۱) وجود داده‌های پرت در یک مجموعه داده اثر نامطلوبی بر استنباط آماری از آن مجموعه داده می‌گذارد و منجر به تولید آماره‌های آزمون و حدود اطمینان نامناسب می‌شود، ۲) در برخی از موارد داده‌های پرت اطلاعات مفیدی از مدل داده‌ها به تحلیل‌گر ارائه می‌دهند و از این رو شناسایی آن برای تحلیل‌گر حائز اهمیت است [۱۶].

روش معمول در مواجهه با داده‌های پرت، شناسایی مکان و نوع آن‌ها و سپس حذف یا جایگزینی موارد شناسایی شده است [۱۱]. اگرچه بررسی ظاهری مشاهده‌ها اولین راه برای تشخیص داده‌های پرت است، این روش به تنهایی مطمئن نیست، علاوه بر این در نمونه‌های پیچیده‌تر با حجم داده‌های بیشتر این کار تقریباً ناممکن است. در این موارد بهتر است به جای بررسی ظاهری داده‌ها، معیارهای منطقی‌تری استفاده شود. با این حال با استفاده از این معیارها نیز نمی‌توان داده‌های پرت را به طور قطعی تشخیص داد، بلکه تنها مقادیر مشکوکی را می‌توان پیدا کرد که می‌توان در رابطه با پرت بودنشان بیشتر تحقیق کرد [۳]. برخی از رویکردهای تحلیل‌های آماری تک متغیره و چندمتغیره مثل تجزیه واریانس در طرح آزمایش‌های منابع طبیعی، رگرسیون چندگانه، آنالیز گرادیان مثل PCA[1] و CCA[2] یا قادر به تشخیص داده‌های پرت نیستند و یا در حضور داده‌های پرت نتایج غیر قابل اعتماد و دور از واقعیت تولید می‌کنند [۱۷]. از این رو هدف این تحقیق، بررسی چند روش آماری و گرافیکی برای شناسایی داده‌های پرت پوشش گیاهی در سه شدت چرای دام در مراتع خشک و بیابانی شهرستان خوسف می‌باشد.

۲. مواد و روش‌ها

تحقیق حاضر در بخشی از مراتع حاشیه کویر، ۴۵ کیلومتری جنوب غرب شهرستان خوسف، استان خراسان جنوبی انجام شد. مراتع مورد مطالعه در یک منطقه دشتی با شیب حدود ۰ تا ۵ درصد قرار دارند. اقلیم منطقه نیمه‌بیابانی و مقدار متوسط بارندگی سالیانه ۹۴/۱۹ میلی‌متر، میانگین درجه حرارت سالیانه ۲۲/۲۱ درجه سانتی‌گراد و تیپ اراضی دشت است. پوشش غالب منطقه مورد مطالعه، تاغ، رمس و اشنان است که مورد چرای شتر قرار می‌گیرد [۴۳]. پس از ارزیابی اولیه میدانی با مشاهده تعداد شترها و آثار چرای شتر بر اشنان به عنوان گونه کلید، سه منطقه با شدت چرای سبک، متوسط و سنگین انتخاب شد. مبنای انتخاب شدت‌های چرا براساس میزان بهره‌برداری گونه اشنان در منطقه معرف هر مرتع با مقایسه قرق تاسیس شده در مجاورت همان مرتع در قالب طرح پایش کیفی مراتع کشور بود [۴۳]. در هر منطقه، تعداد ۳۶ پلات ۱۶ متر مربعی (مجموعاً ۱۰۸ پلات) مستقر شد و درصد پوشش تاجی با استفاده از اندازه‌گیری قطر بزرگ و کوچک و تعیین سطح گیاه اندازه‌گیری شد. تعداد و اندازه پلات و روش اندازه‌گیری براساس دستورالعمل طرح پایش کیفی مراتع کشور تعیین شد [۴۳]. از آنجایی که برخی از روش‌های آماری مورد استفاده در این تحقیق (مثل آزمون‌های Z و Z اصلاح شده) ، عدد ۳۰ را به عنوان مرز بین داده‌های کم و زیاد در نظر می‌گیرند، و برخی از آزمون‌های آماری (مثل آزمون‌های گرابزر و روزنر)، مناسب داده‌های بین ۳۰ تا

[1] Principal Components Analysis
[2] Canonical Correspondence Analysis

۴۰ عدد است [۲۴ و ۲۸]، تعداد ۳۶ پلات برای هر منطقه، در نظر گرفته شد.

۲.۱. روش‌های شناسایی داده پرت قبل از تجزیه و تحلیل آماری

ابتدا در هر کدام از مراتع مورد مطالعه، آمار توصیفی شامل میانگین حسابی، میانگین وینزوری، انحراف معیار و ضریب تغییرات داده‌های پوشش گیاهی محاسبه شد، سپس جهت تعیین وضعیت نرمال بودن داده‌ها و همگنی واریانس‌ها، به ترتیب از آزمون شاپیرو-ویلک و آزمون لیون استفاده شد. برای شناسایی داده‌های پرت قبل از آنالیزهای آماری، از هشت روش آماری زیر استفاده شد:

۱. شناسایی داده های پرت براساس نمودار جعبه‌ای و دامنه میان چارکی (IQR)[1]

در این روش که به روش توکی نیز موسوم است، ابتدا مقادیر چارک‌های اول و سوم محاسبه شد، سپس با تفاضل بین چارک اول و سوم، مقدار دامنه میان چارکی (IQR) تعیین شد. برای نمایش بهتر دامنه میان چارکی، نمودار جعبه‌ای ترسیم شد. نمودار جعبه‌ای روشی استاندارد برای نمایش توزیع داده‌ها است که براساس شاخص‌های آماری حداقل، چارک اول، میانه، چارک سوم و حداکثر ترسیم می‌شود. از این نمودار می‌توان داده(های) پرت را تشخیص داد (شکل ۱). در این نمودار کوچکترین مقدار (حداقل)، کمترین مقداری است که حداکثر ۱/۵ برابر دامنه میان چارکی از چارک اول فاصله دارد و بزرگترین مقدار (حداکثر)، بیشترین مقداری است که حداکثر ۱/۵ برابر دامنه میان چارکی از چارک سوم فاصله دارد. برای شناسایی داده پرت و بسیار پرت[2] براساس مقدار دامنه میان چارکی از روابط ۱ تا ۴ استفاده شد.

داده پرت

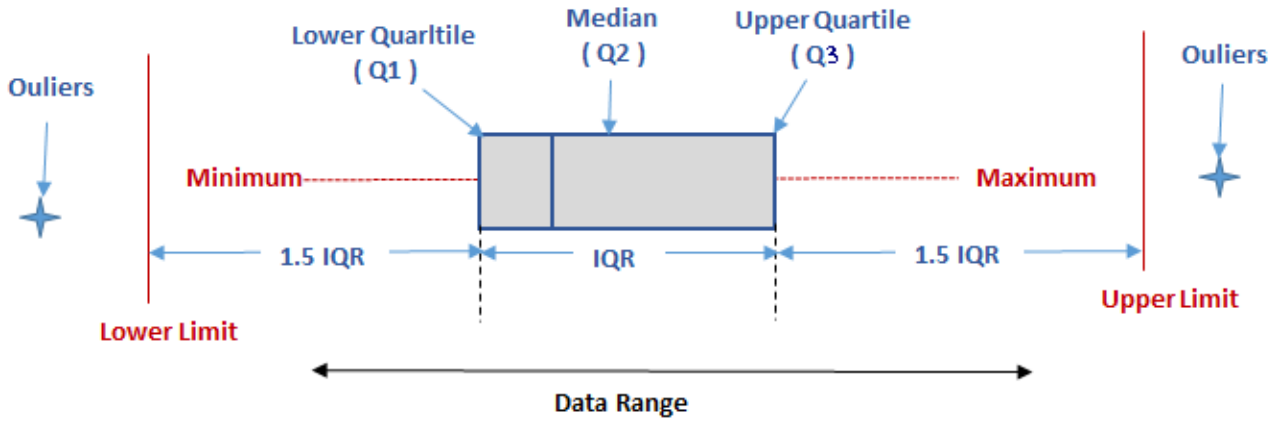
رابطه ۱ (دامنه میان چارکی×۱/۵)+چارک سوم=حد بالا

رابطه ۲ (دامنه میان چارکی×۱/۵)-چارک اول=حد پایین

داده بسیار پرت

رابطه ۳ (دامنه میان چارکی×۳)+چارک سوم=حد بالا

رابطه ۴ (دامنه میان چارکی×۳)-چارک اول=حد پایین



شکل ۱. نمایش داده‌های پرت توسط نمودار جعبه‌ای

اجزای نمودار عبارتند از: حداقل (Minimum)، چارک اول (Q1)، میانه (Median)، چارک سوم (Q3)، حداکثر (Maximum)، دامنه میان چارکی (IQR)، حد پایین (Lower Limit)،حد بالا (Upper Limit) و داده‌های پرت (Outliers)

[1] Interquartile Range
[2] Extreme Outlier

براساس روابط فوق، داده‌ای پرت یا بسیار پرت است که از حد بالا، بزرگتر و از حد پایین، کوچکتر باشد. در این تحقیق، مقادیر خارج از نمودار جعبه‌ای، نشان‌دهنده داده پرت است. همچنین به منظور شناسایی داده پرت، بررسی انحراف از توزیع نرمال و تعیین وضعیت تقارن داده‌ها یا عدم تقارن داده‌ها (وجود چولگی مثبت و منفی) از نمودار هیستوگرام استفاده شد.

۲. شناسایی داده پرت براساس انحراف معیار از میانگین (قانون سه سیگما)

پس از محاسبه میانگین و انحراف معیار داده‌های پوشش گیاهی در هر کدام از مراتع مورد مطالعه، مقدار Z از طریق رابطه ۵ محاسبه شد.

رابطه ۵ $$Z_{Score} = \frac{x - mean}{\text{Standard Deviation}}$$

براساس آماره Z، داده‌ای پرت است که مقدار Z آن بیشتر از ۲/۵+ یا کوچکتر از ۲/۵- باشد و داده‌ای بسیار پرت است که مقدار Z آن بیشتر از ۳+ یا کوچکتر از ۳- باشد (شکل ۲). به عبارت دیگر، مقادیر کوچکتر و بزرگتر از ۲/۵± و ۳± انحراف معیار از میانگین داده‌ها، جزو داده پرت و بسیار پرت محسوب می‌شوند (روابط ۶ تا ۹).

داده پرت

رابطه ۶ (انحراف معیار ×۲/۵)+میانگین=حد بالا

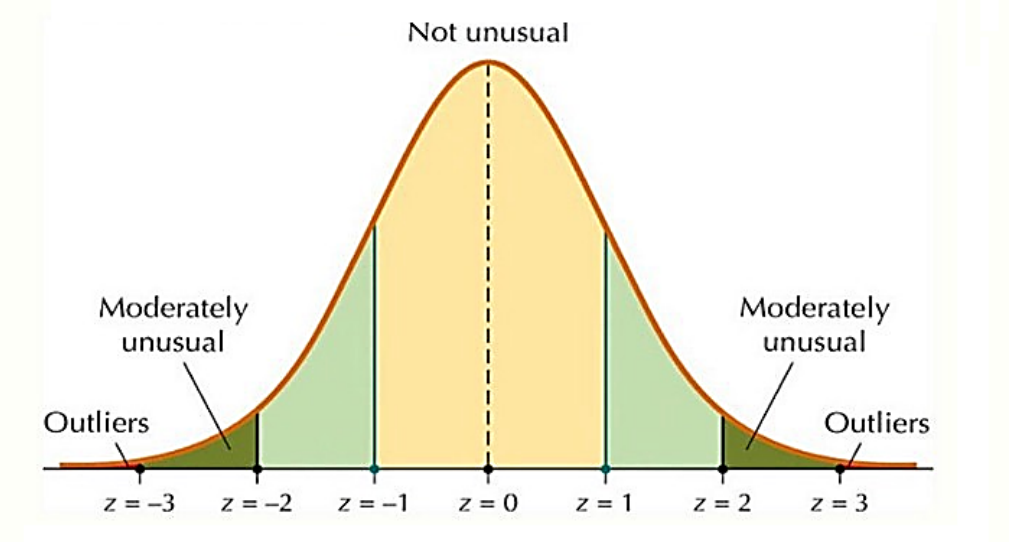
رابطه ۷ (انحراف معیار ×۲/۵)-میانگین=حد پایین

داده بسیار پرت

رابطه ۸ (انحراف معیار ×۳)+میانگین=حد بالا

رابطه ۹ (انحراف معیار ×۳)-میانگین=حد پایین

براساس روابط فوق، داده‌ای پرت یا بسیار پرت است که از حد بالا، بزرگتر و از حد پایین، کوچکتر باشد.



شکل ۲. نمایش داده‌های پرت براساس قانون سه سیگما

از آنجایی که نمره Z اصلاح شده به مقدار میانه وابسته است از این رو نسبت به مقدار Z، حساسیت کمتری به داده‌های پرت دارد. پس از محاسبه مقدار Z، در مرحله قبل، با اعمال ضریب d (۰/۶۷۴۵)و مقدار انحراف مطلق از میانه (MAD)، مقدار Z اصلاح شده از طریق رابطه ۱۰ محاسبه شد [۴۵].

رابطه ۱۰ $$Modified\ Z_{score} = \frac{0.6745 \times (x - median)}{\text{MAD}}$$

که در آن: x، median و MAD، به ترتیب، سری داده، میانه و انحراف مطلق از میانه است. براساس این روش، مقدار ۳/۵ خط برش محسوب می‌شود و مقادیر کوچکتر و بزرگتر از ۳/۵، داده پرت محسوب می‌شوند (روابط ۱۱ و ۱۲).

رابطه ۱۱ ۳/۵<مقدار Z اصلاح شده=حد بالا

رابطه ۱۲ ۳/۵>مقدار Z اصلاح شده=حد بالا

براساس روابط فوق، داده‌ای پرت است که از حد بالا، بزرگتر و از حد پایین، کوچکتر باشد. در این تحقیق، برای مقادیر Z و Z اصلاح شده، آمار توصیفی شامل حداقل، حداکثر، چارک اول و سوم و میانه داده‌ها محاسبه شد.

۳. شناسایی داده پرت براساس انحراف مطلق از میانه[1] (MAD)

روش دیگر معروف به روش همپل است، در این روش، مقدار انحراف مطلق از میانه از رابطه ۱۳ محاسبه می‌شود [۴۶].

$$MAD = median(|x_i - m|)$$ رابطه ۱۳

که در آن: x و m به ترتیب، سری داده و میانه است. در این روش مقادیر خارج از بازه میانه ±۳ انحراف مطلق از میانه داده‌ها، به عنوان داده پرت محسوب می‌شود (روابط ۱۴ و ۱۵).

رابطه ۱۴ (انحراف مطلق از میانه ×۳)+میانه=حد بالا

رابطه ۱۵ (انحراف مطلق از میانه ×۳)-میانه=حد پایین

براساس روابط فوق، داده‌ای پرت است که از حد بالا، بزرگتر و از حد پایین، کوچک تر باشد.

۴. شناسایی داده پرت براساس میانگین پیراسته[2] (Trimmed)

پس از محاسبه میانگین حسابی مقادیر درصد پوشش گیاهی در سه مرتع مورد مطالعه، میانگین پیراسته با حذف ۵ درصد داده‌ها در طرفین (جمعاً ۱۰ درصد حذف و ۹۰ درصد باقی بماند) و ۱۰ درصد از طرفین (جمعاً ۲۰ درصد حذف و ۸۰ درصد باقی بماند) تحت عنوان میانگین پیراسته ۱۰ درصد و ۲۰ درصد محاسبه شد. مقادیر حذف شده، به عنوان داده پرت محسوب می‌شود.

۵. شناسایی داده پرت براساس مقادیر صدک ۱ و ۹۹ (P1 و P99)

یکی دیگر از روش‌های شناسایی داده پرت، محاسبه صدک ۱ ام و ۹۹ ام سری داده‌هاست. مقادیر کمتر از صدک ۱ ام و بیشتر از ۹۹ ام، جزو داده‌های پرت محسوب می‌شوند. این مقادیر برای صدک‌های ۲/۵ ام و ۹۷/۵ ام نیز محاسبه شد.

۶. شناسایی داده پرت براساس آزمون کای اسکوئر (χ^2)

آماره کای اسکوئر براساس تفاضل مربع بین داده‌ها و میانگین نمونه محاسبه می‌شود. البته شرط انجام این آزمون، آگاهی از واریانس جمعیت است. در صورت مشخص نبودن واریانس جمعیت، واریانس را از نمونه می‌گیرد [۲۸]. فرضیه جایگزین در این آزمون: بزرگترین و کوچکترین عدد به عنوان داده پرت در نظر گرفته می‌شود.

۷. شناسایی داده پرت به روش گرابز[3] (ESD)

آزمون گرابز یکی از آزمون‌های شناسایی داده های پرت تک متغیره است که از روابط ۱۶ و ۱۷ محاسبه می‌شود. این آزمون زمانی استفاده می‌شود که حجم داده‌ها بیشتر از ۳۰ باشد و داده‌ها توزیع نرمال داشته باشند [۲۴].

$$T = \frac{X_{max} - \bar{x}}{S_x}$$ رابطه ۱۶

$$T = -\frac{X_{min} - \bar{x}}{S_x}$$ رابطه ۱۷

[1] Median Absolute Deviation
[2] Trimmed Mean
[3] Extreme Studentized Deviate

که در آن: X_{max}، X_{min}، $\bar{x}$ و S_x به ترتیب بزرگترین عدد، کوچکترین عدد، میانگین و انحراف معیار نمونه است.

این آزمون دو فرضیه را بررسی می‌کند:

H_0: بزرگترین و کوچکترین عدد، داده پرت محسوب نمی‌شود.

H_1: بزرگترین و کوچکترین عدد، داده پرت محسوب می‌شود.

۸. شناسایی داده پرت به روش روزنر (generalised ESD)

روزنر، با اصلاح فرمول گرابز، رابطه ۱۸ را توصیه کرد. این آزمون زمانی استفاده می شود که حجم داده‌ها بیشتر از ۲۵ عدد باشد و داده‌ها از توزیع نرمال برخوردار باشند.

$$R_i = \frac{max_i|x_i - \bar{x}|}{S_x}$$ رابطه ۱۸

که در آن $\bar{x}$ و S_x به ترتیب میانگین و انحراف معیار نمونه است. در این روش، مشاهده‌ای که مقدار $|x_i - \bar{x}|$ را حداکثر می‌کند، حذف و سپس آماره فوق را n - 1 مشاهده دوباره محاسبه می‌کند. این منجر به R آزمون شامل R_1، R_2 ... R_r می‌شود [۲۴].

این آزمون دو فرضیه را بررسی می‌کند:

H_0: بزرگترین و کوچکترین عدد، داده پرت محسوب نمی‌شود.

H_1: بزرگترین و کوچکترین عدد، داده پرت محسوب می‌شود.

۲.۲. تجزیه و تحلیل آماری پس از حذف داده پرت

پس از شناسایی داده‌های پرت و حذف آن‌ها از مجموعه داده‌های پوشش گیاهی در سه مرتع مورد مطالعه، آماره‌های توصیفی شامل حداقل، حداکثر، میانگین حسابی، میانگین وینزوری، انحراف معیار و ضریب تغییرات بازمحاسبه شدند، همچنین نرمال بودن داده‌ها و همگنی واریانس‌ها، توسط آزمون شاپیرو-ویلک و لیون بررسی شد. به دلیل نرمال نبودن داده‌ها پس از حذف داده‌های پرت، از آزمون کروسکال-والیس برای بررسی اثر شدت چرای دام بر درصد پوشش گیاهی استفاده شد. برای مقایسه بین سه مرتع مورد مطالعه از آزمون دان[1] استفاده شد [۵۰]. کلیه آزمون‌های آماری توسط بسته‌های outliers، EnvStats، ggpubr و base در محیط R انجام شد [۴۰].

۳. نتایج

نتایج آمار توصیفی داده‌های اولیه پوشش گیاهی در سه مرتع با شدت چرای سبک، متوسط و سنگین در جدول ۱ نشان داده شده است. مقدار درصد پوشش گیاهی از ۲۳/۳۰ درصد در مرتع با شدت چرای متوسط تا ۴۳/۳۹ درصد در مرتع با شدت چرای کم، متغیر است. بیشترین تغییرات پوشش گیاهی با انحراف معیار ۲۲/۳۴ مربوط به مرتع با شدت چرای سبک و کمترین تغییرات پوشش گیاهی با انحراف معیار ۱۱/۳۸ مربوط به مرتع با شدت چرای سنگین است. نتیجه آزمون شاپیرو-ویلک نشان داد که پوشش گیاهی مرتع با شدت چرای سنگین توزیع نرمال دارد، در حالی که داده‌های درصد پوشش گیاهی مراتع با شدت چرای کم و متوسط، نرمال نیستند. به منظور درک بهتر از داده‌های پوشش گیاهی، بخشی از داده‌ها در جدول ۲ مشاهده می‌شود.

براساس روش دامنه میان چارکی (IQR)، مقادیر حد پایین و بالا مشخص شد، نتیجه این روش نشان داد که بالاترین حد برای داده پرت، عدد ۶۷/۵ است، و درصد پوشش ۷۰ درصد در مرتع با شدت چرای متوسط، به عنوان داده پرت محسوب می‌شود. براساس این روش، هیچ‌کدام از داده‌ها، به عنوان داده بسیار پرت شناسایی نشدند.

[1] Dunn test

جدول ۱. آمار توصیفی داده‌های درصد پوشش گیاهی مراتع مورد مطالعه قبل از حذف داده‌های پرت

شاخص‌های آماری	شدت چرا		
	سبک	متوسط	سنگین
میانگین حسابی	۴۳/۳۹	۲۳/۳۰	۲۵/۴۲
میانگین پیراسته (۲۰ درصد)	۴۰/۲۷	۲۹/۵۰	۲۴/۷۳
میانگین وینزوری	۴۰/۹۴	۳۰/۲۸	۲۵/۰۳
انحراف معیار	۲۲/۳۴	۱۵/۲۴	۱۱/۳۸
ضریب تغییرات (درصد)	۵۱	۴۷	۴۵
ارزش P آزمون شاپیرو-ویلک	۰/۰۰۹	۰/۰۱	۰/۵۰
نتیجه آزمون نرمال بودن داده‌ها	غیرنرمال	غیرنرمال	نرمال
ارزش P آزمون لیون	۰/۰۰۱		
نتیجه آزمون همگن-ناهمگن بودن واریانس‌ها	ناهمگن		

ارزش P بزرگتر از ۰/۰۵، نشان‌دهنده معنی‌داری نتیجه آزمون نرمال بودن داده‌ها و همگن-ناهمگن بودن واریانس‌ها است.

جدول ۲. قسمتی از داده‌های درصد پوشش گیاهی (از کم به زیاد، ۳ پلات کوچکترین عدد و ۳ پلات بزرگترین عدد) (n=۱۰۸)

شدت چرا		
سبک	متوسط	سنگین
۱۵	۱۱	۰
۱۷	۱۵	۱۰
۱۸	۱۵	۱۰
۷۵	۶۰	۴۵
۱۰۰	۶۰	۴۵
۱۰۰	۷۰	۴۷

جدول ۳. شناسایی داده پرت مقادیر درصد پوشش گیاهی براساس دامنه میان چارکی (IQR)

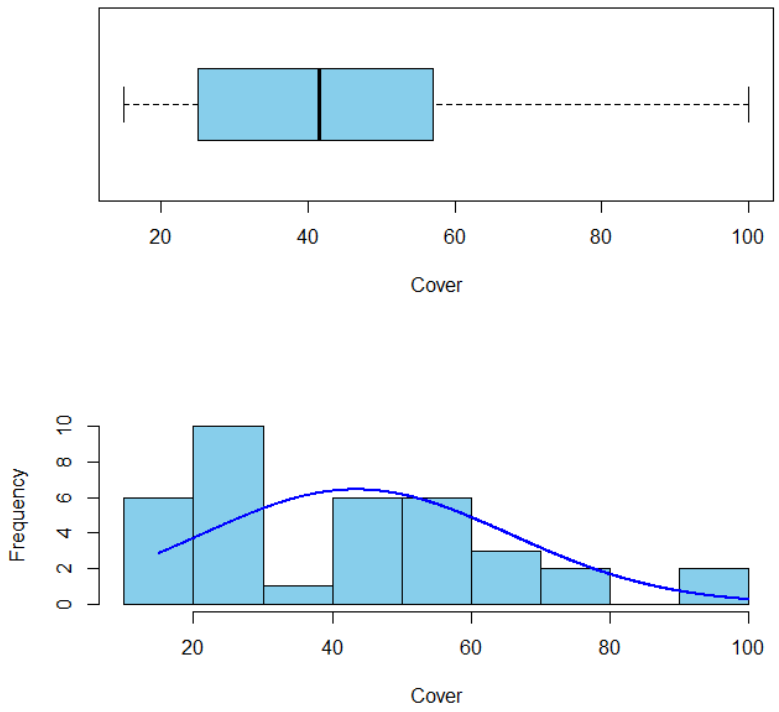
شاخص‌های آماری	شدت چرا		
	سبک	متوسط	سنگین
حداقل (Min)	۱۵	۱۱	۰
حداکثر (Max)	۱۰۰	۷۰	۴۷
چارک اول (Q1)	۲۵	۲۰	۱۸
چارک سوم (Q3)	۵۷	۳۹	۳۲/۷۵
دامنه میان چارکی (IQR)	۳۲	۱۹	۱۴/۷۵
حد بالا برای داده پرت	۱۰۵	۶۷/۵	۵۴/۸۷
حد پایین برای داده پرت	-۲۳	-۸/۵	-۴/۱۲
حد بالا برای داده بسیار پرت	۱۵۳	۹۶	۷۷
حد پایین برای داده بسیار پرت	-۷۱	-۳۷	-۲۶/۲۵
مقدار داده پرت	-	۷۰	-
مقدار داده بسیار پرت	-	-	-

همچنین نمودارهای جعبه‌ای و هیستوگرام مراتع مورد مطالعه نشـان می‌دهد که توزیع درصـد پوشـش گیاهی مراتع با شدت چرای سبک و متوسط دارای چولگی مثبت هستند و مقادیر درصد پوشش ۱۰۰ درصد و ۷۰ درصد با بقیه مقادیر فاصـله دارند (شـکل‌های ۳ و ۴). نمودار جعبه‌ای مرتع با شـدت چرای متوسـط نشـان می‌دهد که عدد ۷۰ درصـد، خارج از محدوده جعبه قرار دارد و به عنوان داده پرت در نظر گرفته می‌شود. در مرتع با شدت چرای سـنگین، نمودارهای جعبه‌ای و هیسـتوگرام نشـان می‌دهد که توزیع پوشش گیاهی متقارن بوده و این حاکی از عدم وجود چولگی اسـت (شـکل ۵). در این مرتع، داده پرت در بین داده‌های درصد پوشش گیاهی مشاهده نشد و آزمون شـاپیرو-ویلک نیز نشـان داد داده‌ها توزیع نرمال دارند.

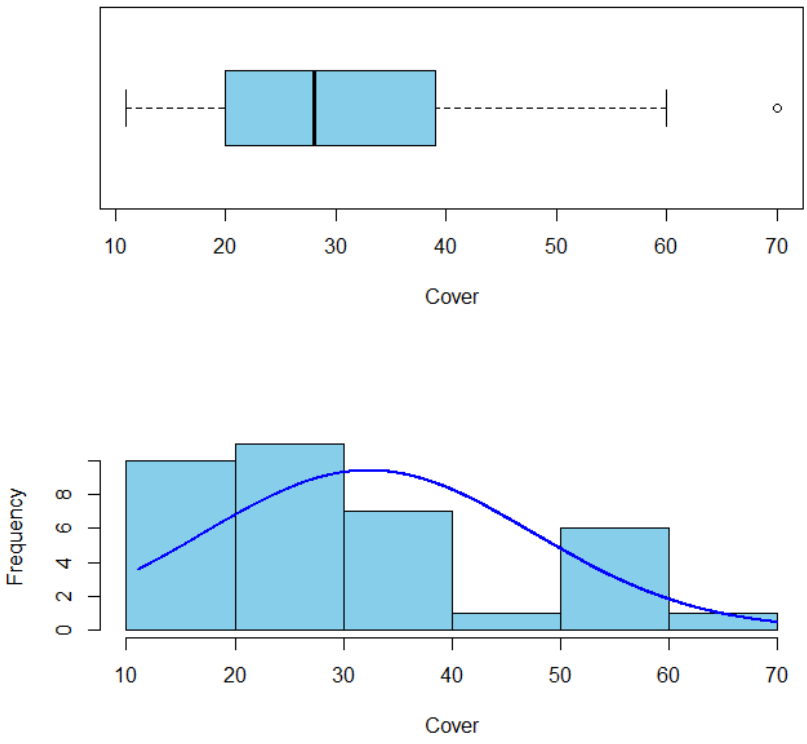
در این تحقیق، از قانون سـه سـیگما نیز در شـناسـایی داده‌های پرت اسـتفاده شـد. پس از محاسـبه میانگین و انحراف معیار داده‌های پوشش گیاهی، نتایج نشان داد که هیچکدام از داده‌ها، از حد بالا و پایین خارج نشـدند و این روش هیچ داده‌ای را به عنوان داده پرت تشخیص نداد.

مقادیر Z و Z اصلاح شده نیز محاسبه شد. آماره‌های توصیفی این مقادیر به ترتیب در جداول ۵ و ۶ ارائه شده است. نتایج روش Z نشان داد که در مرتع با شدت چرای سبک، حداکثر درصد پوشش گیاهی (۱۰۰ درصد) مقدار Z برابر با ۲/۵۳ دارد و این مقدار از عدد ۲/۵ بالاتر اسـت، از این رو ۱۰۰ درصـد پوشـش گیاهی عدد پرت محسـوب می‌شود. چون هیچکدام از مقادیر از Z ۳ بیشتر نشدند، از این‌رو داده بسیار پرتی برای مراتع مورد مطالعه شـناسـایی نشد. با اصلاح مقدار Z و محاسبه آمار توصیفی آن، نتایج نشـان می‌دهد که هیچکدام از مقادیر، داده پرت محسـوب نمی‌شوند.

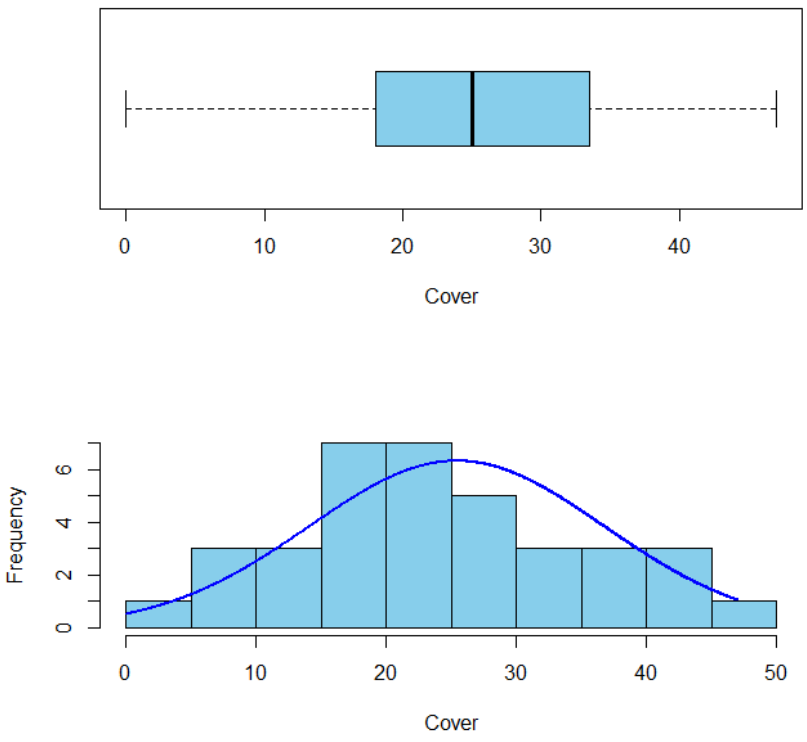
براساس روش انحراف مطلق از میانه (MAD)، حد بالا برای داده پرت، عدد ۷۰/۲۵ محاسـبه شـد، از این‌رو عدد ۷۰ درصد پوشش گیاهی در مرتع با شدت چرای متوسط به عنوان عدد پرت در نظر گرفته می‌شود (جدول ۷).



شکل ۳. نمودار جعبه‌ای و هیستوگرام درصد پوشش گیاهی در شدت چرای سبک



شکل ۴. نمودار جعبه‌ای و هیستوگرام درصد پوشش گیاهی در شدت چرای متوسط



شکل ۵. نمودار جعبه‌ای و هیستوگرام درصد پوشش گیاهی در شدت چرای سنگین

جدول ۴. شناسایی داده پرت مقادیر درصد پوشش گیاهی براساس انحراف معیار از میانگین (قانون سه سیگما)

شاخص‌های آماری	شدت چرا		
	سبک	متوسط	سنگین
میانگین	۴۳/۳۹	۲۳/۳۰	۲۵/۴۲
انحراف معیار	۲۲/۳۴	۱۵/۲۴	۱۱/۳۸
حد بالا برای داده پرت	۱۱۰/۴۱	۷۸/۰۲	۵۹/۵۶
حد پایین برای داده پرت	-۲۳/۶۴	-۱۳/۴۱	-۸/۷۳
مقدار داده پرت	-	-	-

جدول ۵. خلاصه آماری مقادیر Z مربوط به درصد پوشش گیاهی در مراتع مورد مطالعه

شاخص‌های آماری	شدت چرا		
	سبک	متوسط	سنگین
حداقل (Min)	-۱/۲۷	-۱/۴	-۲/۲۳
حداکثر (Max)	۲/۵۳	۲/۴۷	۱/۹۰
چارک اول (Q1)	-۰/۸۲	-۰/۸۱	-۰/۶۵
چارک سوم (Q3)	۰/۶۱	۰/۴۴	۰/۶۴
میانه	-۰/۰۸	-۰/۲۸	-۰/۰۴
مقدار داده پرت	۱۰۰	-	-
مقدار داده بسیار پرت	-	-	-

جدول ۶. خلاصه آماری مقادیر Z اصلاح شده مربوط به درصد پوشش گیاهی در مراتع مورد مطالعه

شاخص‌های آماری	شدت چرا		
	سبک	متوسط	سنگین
حداقل (Min)	-۰/۷۳	-۰/۸۱	-۱/۶۲
حداکثر (Max)	۱/۶۱	۲/۰۱	۱/۴۳
چارک اول (Q1)	-۰/۴۵	-۰/۳۸	-۰/۴۵
چارک سوم (Q3)	۰/۴۳	۰/۵۳	۰/۵۰
میانه	۰/۰۰	۰/۰۰	۰/۰۰
مقدار داده پرت	-	-	-

جدول ۷. شناسایی داده پرت مقادیر درصد پوشش گیاهی براساس انحراف مطلق از میانه (آزمون همپل)

شاخص‌های آماری	شدت چرا		
	سبک	متوسط	سنگین
میانه	۴۱/۵	۲۸	۲۵
انحراف مطلق از میانه	۲۴/۴۶	۱۴/۰۸	۱۰/۳۸
حد بالا برای داده پرت	۱۱۴/۸۹	۷۰/۲۵	۵۶/۱۳
حد پایین برای داده پرت	-۳۱/۸۹	-۱۴/۲۵	-۶/۱۳
مقدار داده پرت	-	۷۰	-

پس از حذف ۲۰ درصد داده‌ها ، میانگین پیراسته ۲۰ درصد محاسبه شد (جدول ۸). نتایج نشان داد که مقادیر ۱۰۰ درصد و ۷۰ درصد در مراتع با سدت چرای سبک و متوسط، جزو داده‌های پرت محسوب می‌شوند. میانگین پیراسته ۱۰ درصد، هیچ کدام از مقادیر را به عنوان داده پرت تشخیص نداد.

جدول ۸. شناسایی داده پرت مقادیر درصد پوشش گیاهی براساس میانگین پیراسته

شاخص‌های آماری	شدت چرا		
	سبک	متوسط	سنگین
میانگین پیراسته (۲۰درصد)	۴۰/۲۷	۲۹/۵۰	۲۴/۷۳
مقدار داده پرت	۱۰۰	۷۰	-
میانگین پیراسته (۱۰درصد)	۴۱/۲۳	۳۱/۰۷	۲۵/۲۷
مقدار داده پرت	-	-	-

همچنین با محاسبه مقادیر صدک ۱ ام و ۹۹ ام، تمامی مقادیر درصد پوشش گیاهی کمتر از ۱۵/۷، ۱۲/۴ و ۳/۵ درصد و بیشتر از ۱۰۰، ۶۶/۵۰ و ۴۶/۳ درصد به ترتیب برای مراتع با شدت چرای سبک، متوسط و سنگین، داده پرت محسوب می‌شود (جدول ۹). برای صدک ۲/۵ ام و ۹۷/۵ ام نیز محاسبه شده است و مقادیر داده پرت برای هر دو سری یکسان است.

جدول ۹. شناسایی داده پرت مقادیر درصد پوشش گیاهی براساس مقادیر صدک‌های مختلف

شاخص‌های آماری	شدت چرا		
	سبک	متوسط	سنگین
صدک ۱ ام	۱۵/۷	۱۲/۴	۳/۵
صدک ۹۹ام	۱۰۰	۶۶/۵۰	۴۶/۳
مقدار داده پرت	۱۵	۱۱-۷۰	۰-۴۷
صدک ۲/۵ ام	۱۶/۷۵	۱۴/۵۰	۸/۷۵
صدک ۹۷/۵ ام	۱۰۰	۶۱/۲۵	۴۵/۲۵
مقدار داده پرت	۱۵	۱۱-۷۰	۰-۴۷

در این تحقیق، علاوه بر شاخص‌های آماری و نمودارهای مربوطه، از سه آزمون مربوط به داده‌های تک متغیره شامل آزمون کای اسکوئر، گرابز و روزنر نیز استفاده شد. در این روش‌ها، مقادیر حداکثر و حداقل داده به عنوان داده پرت آزمون شدند. آماره‌های χ^2، G و R_1 و R_2 روش‌های مربوطه محاسبه شد (جدول ۱۰). نتایج نشان داد که از بین سه روش مورد مطالعه، فقط آزمون کای اسکوئر، مقادیر ۱۰۰، ۷۰ و ۰ درصد پوشش گیاهی در سه مرتع مورد مطالعه را به عنوان عدد پرت تشخیص داد.

نتایج آزمون شاپیرو-ویلک نشان داد که مقادیر درصد پوشش گیاهی دو مرتع با شدت چرای سبک و متوسط توزیع نرمال نداشتند. برخی از روش‌ها و آزمون‌های انجام شده در این تحقیق، نشان داد که داده پرت در این دو

مرتع وجود دارد، داده‌های پرت در این دو مرتع حذف شدند و مجدداً آمار توصیفی مربوط به درصد پوشش گیاهی محاسبه شد (جدول ۱۱). در مرتع با شدت چرای سبک، میانگین حسابی درصد پوشش گیاهی از ۴۳/۳۹ درصد به ۴۰/۰۶ درصد و در مرتع با شدت چرای متوسط از ۲۳/۳۰ درصد به ۳۱/۲۳ درصد تغییر پیدا کرد. این تغییر لندک باعث تغییر عدم تقارن توزیع نشد و نتیجه آزمون شاپیرو-ویلک نشان داد که داده‌های درصد پوشش گیاهی حتی پس از حذف داده‌های نرمال نشدند. در خصوص همگنی واریانس‌ها که یکی دیگر از پیش نیازهای آزمون تجزیه واریانس است، نتایج نشان داد که پس از حذف داده پرت، واریانس گروه‌ها همگن شدند.

جدول ۱۰. شناسایی داده پرت مقادیر درصد پوشش گیاهی براساس آزمون‌های آماری کای اسکوئر، گرابز و روزنر

نام آزمون	شدت چرا					
	سبک		متوسط		سنگین	
آزمون کای اسکوئر	χ^2	p-value	χ^2	p-value	χ^2	p-value
بزرگترین عدد	۶/۴۲	۰/۰۱۱	۶/۱۲	۰/۰۱۳	۳/۵۹	۰/۰۶
کوچکترین عدد	۱/۶۱	۰/۲۰	۱/۹۵	۰/۱۶	۴/۹۹	۰/۰۲۵
مقدار داده پرت	۱۰۰		۷۰		۰	
آزمون گرابز	G	p-value	G	p-value	G	p-value
بزرگترین عدد	۲/۵۳	۰/۱۵	۲/۴۷	۰/۱۸	۲/۲۳	۰/۳۸
کوچکترین عدد	۱/۲۷	۱	۱/۴۰	۱	۱/۹۰	۰/۹۵
مقدار داده پرت	-		-		-	
آزمون روزنر	R_2	R_1	R_2	R_1	R_2	R_1
	۲/۸۵	۱/۵۳	۲/۰۵	۲/۴۷	۲/۲۳	۱/۹۵
مقدار داده پرت	-		-		-	

ارزش P کوچکتر از ۰/۰۵، نشان‌دهنده معنی‌داری نتیجه آزمون‌های کای اسکوئر، گرابز و روزنر است.

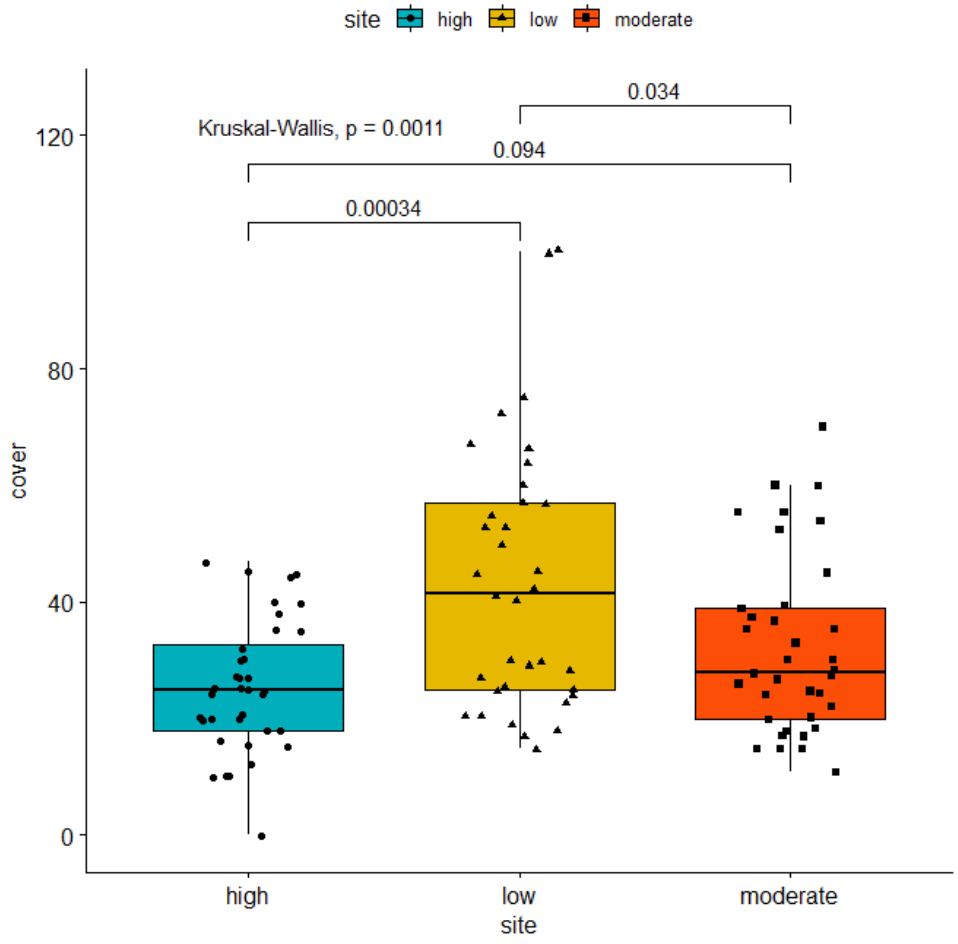
جدول ۱۱. آمار توصیفی داده‌های درصد پوشش گیاهی مراتع مورد مطالعه بعد از حذف داده‌های پرت

شاخص‌های آماری	شدت چرا	
	سبک	متوسط
حداقل	۱۵	۱۱
حداکثر	۷۵	۶۰
میانگین حسابی	۴۰/۰۶	۳۱/۲۳
میانگین وینزوری	۳۹/۲۱	۲۸/۹۰
انحراف معیار	۱۷/۹۹	۱۴
ضریب تغییرات (درصد)	۴۵	۴۵
سطح معنی‌داری آزمون شاپیرو-ویلک	۰/۰۳	۰/۰۱
وضعیت نرمال بودن داده‌ها	غیرنرمال	غیرنرمال
سطح معنی‌داری آزمون لیون	۰/۱۱	
نتیجه آزمون همگن-ناهمگن بودن واریانس‌ها	همگن	

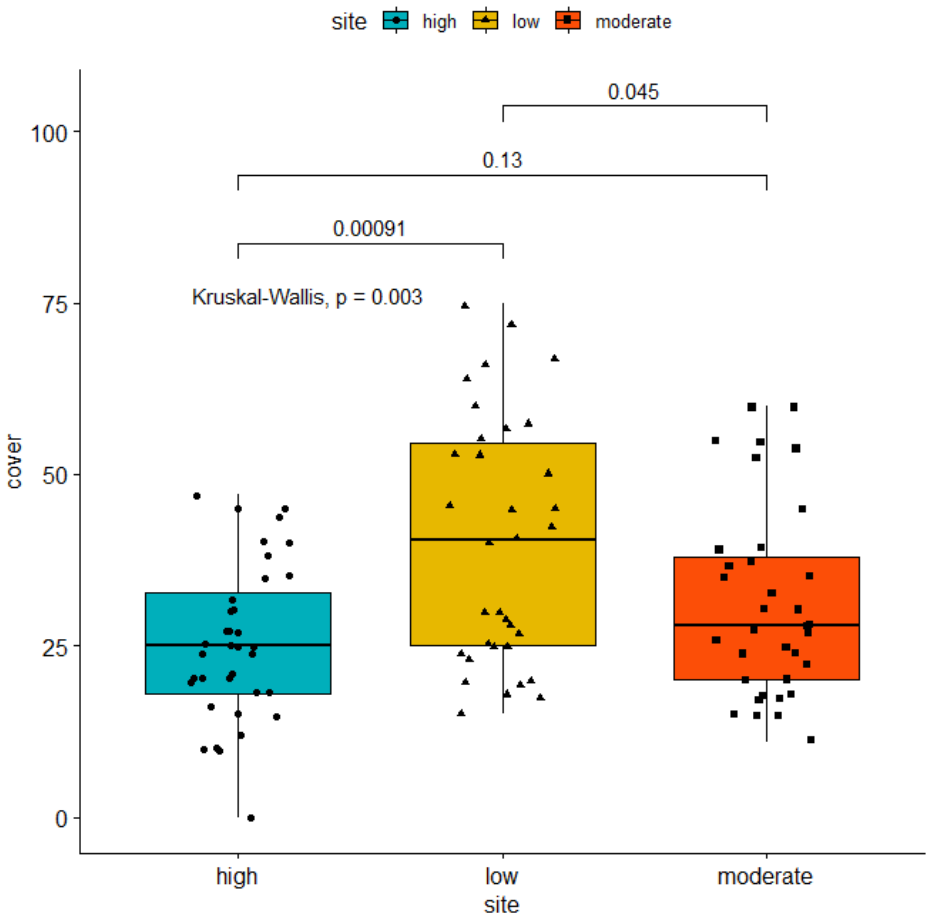
ارزش P بزرگتر از ۰/۰۵، نشان‌دهنده معنی‌داری نتیجه آزمون نرمال بودن داده‌ها و همگن-ناهمگن بودن واریانس‌ها است.

به دلیل عدم تامین کامل شـــرایط پارامتری برای بررسـی اثر شـدت چرای دام بر درصـد پوشـش گیاهی، از آزمون ناپارامتری کروسـکال-والیس اسـتفاده شـد. نتیجه آزمون قبل از حذف داده‌های پرت، نشـان داد که بین سـه مرتع مورد مطالعه از لحاظ درصـد پوشـش گیاهی تفاوت معنی‌دار وجود دارد. بیشترین درصد پوشش گیاه در مرتع با شـدت چرای سـبک مشـاهده می‌شـود. بین دو مرتع با شـدت چرای متوسـط و سـنگین، تفاوت آماری معنی‌داری مشاهده نمی‌شود (شکل ۶).

پس از حذف داده‌های پرت در دو مرتع با شدت چرای سـبک و متوسـط، نتیجه آزمون کروسـکال-والیس تغییر زیادی نکرد، تنها ارزش P آزمون از ۰/۰۰۳۴ به ۰/۰۰۳ تغییر کرد، همچنین مقادیر ارزش P مقایســه‌های دوتایی آزمون دان نیز انـدکی تغییر کرد، امـا تـاثیری در نتیجـه نهایی نداشـت (شـکل ۷). در کل، نتایج نشـان می‌دهد که اثر شـدت چرای دام بر درصـد پوشـش گیاهی مراتع مورد مطالعه معنی‌داری است ($P\leq$ ۰/۰۵).



شکل ۶. مقایسه میانگین‌های درصد پوشش گیاهی قبل از حذف داده‌های پرت



شکل ۷. مقایسه میانگین‌های درصد پوشش گیاهی بعد از حذف داده‌های پرت

۴. بحث و نتیجه‌گیری

در این تحقیق هشت روش آماری و گرافیکی جهت شناسایی داده های پرت در مطالعه بررسی اثر شدت چرای دام بر درصد پوشش گیاهی در مراتع خشک و بیابانی شهرستان خوسف ارزیابی شدند. آمار توصیفی و استنباطی ارائه شدند. با بررسی تحقیقات انجام شده در منابع علمی علوم مرتع مشاهده می‌شود محققین معمولاً هنگام تجزیه و تحلیل آماری آزمون‌های فرضیه مقایسه‌ای مستقیماً سراغ بیان نتایج آمار استنباطی می‌روند، حال آنکه آمار توصیفی، که نشان دهنده وضعیت داده‌ها و شکل توزیع داده‌هاست، نادیده گرفته می‌شود. یکی از پیش‌فرض‌های آزمون‌های فرضیه مقایسه‌ای، عدم وجود داده‌های پرت است، نتایج این تحقیق نشان داد که می‌توان از طریق نتایج آمار توصیفی (مثل میانگین، انحراف معیار، دامنه میان چارکی و...) وجود یا عدم وجود داده‌های پرت را ارزیابی کرد. نتایج تحقیق حاضر نشان داد که میانگین درصد پوشش گیاهی در دو مرتع با شدت چرای متوسط و سنگین نزدیک یکدیگر هستند، با این حال قبل از مقایسه آماری، وضعیت نرمال بودن داده‌ها بررسی شد. آزمون شاپیرو ویلک نشان داد که داده‌های مراتع با شدت چرای سبک و متوسط توزیع نرمال ندارند. حتی حذف داده پرت نیز منجر به نرمال شدن این داده‌ها نشد. از این‌رو کاربرد تجزیه واریانس قبل از بررسی اولیه داده‌های پوشش گیاهی نادرست است.

از هشت روش استفاده شده، روش قانون سه سیگما،

Z اصلاح شده و آزمون‌های گرابز و روزنر، هیچکدام از داده‌های درصد پوشش گیاهی را به عنوان داده پرت شناسایی نکردند. احتمالاً کاربرد این آزمون‌ها به حجم داده بستگی دارد، Saleem و همکاران (۲۰۲۱) [۴۴] گزارش کردند که تعداد داده‌های پرت شناسایی شده در روش انحراف معیار (قانون سه سیگما) در نمونه‌های کوچک و بزرگ، متفاوت است. دو برآوردگر مورد استفاده در روش Z، یعنی میانگین نمونه و انحراف معیار نمونه، می‌توانند تحت تأثیر داده‌های پرت قرار گیرند. برای غلبه بر این چالش از روش Z اصلاح شده که در آن، میانه و انحراف مطلق از میانه محاسبه می‌شود، استفاده می‌شود. از طرفی، روش Z برای نمونه‌های کوچک دقیق نیست [۳]. Dehghan و Faridrohani (۲۰۲۲) [۱۶] با انتقاد از روش انحراف معیار از میانگین (قانون سه سیگما) و Z، بیان کردند که در این روش‌ها، از شاخص‌های میانگین و انحراف معیار مشاهده‌ها استفاده می‌شود، حال آنکه این شاخص‌ها، برآوردگرهای استواری[1] از پارامترهای مکان و مقیاس نیستند و عملکرد روش‌های فوق که معمولاً تحت برقراری فرض نرمال از آن استفاده می‌شود، تحت تاثیر داده‌های پرت است. Ahmadi و Sarmad (۲۰۱۰) [۳] با نقد روش Z اصلاح شده، بیان کردند که در این فرمول، ثابت d (رابطه ۱۰) را زمانی می‌توان برابر ۰/۶۷۴۵ در نظر گرفت که حجم نمونه به بی‌نهایت میل کند. بنابراین در نمونه‌های با حجم کم، مقدار واقعی این ثابت متفاوت است و باید برای هر تحقیقی به صورت جداگانه بررسی شود. Jolous Jamshidi و همکاران (۲۰۲۲) [۲۵] روش Z اصلاح شده را در شناسایی داده‌های پرت در مجموعه داده‌های بزرگ مربوط به منابع آب موثر دانستند.

Cho و همکاران (۲۰۱۶) [۱۲] در خصوص داده‌های محیطی نشان دادند که با استفاده از روش روزنر حدود ۵ تا ۱۰ درصد داده‌ها، به عنوان داده‌های پرت شناسایی می‌شوند، در آن تحقیق، پس از حذف داده‌های پرت، میانگین و انحراف معیار داده‌ها نسبت به قبل از حذف داده‌ها، بین ۱۳ تا ۳۳ درصد تغییر کرد. در یک مطالعه جنگلداری، Hordo و همکاران (۲۰۰۶) [۲۲] نشان دادند که برای نمونه‌های کوچک، آزمون دیکسون وبرای نمونه‌های بزرگ، قانون دو سیگما موثرتر از آزمون گرابز است.

در تحقیق حاضر، نتیجه روش دامنه میان چارکی و نمودار جعبه‌ای (روش توکی) و روش انحراف مطلق از میانه (آزمون همپل) نتیجه یکسانی داشتند. براساس این روش‌ها، مرتع با شدت چرای متوسط دارای عدد پرت بودند. روش نمودار جعبه‌ای، یکی از ساده‌ترین و مفیدترین روش‌ها در شناسایی داده پرت محسوب می‌شود [۱۰] و نیازی به هیچ فرض توزیع آماری ندارد و توسط چارک‌ها ترسیم می‌شود که تحت تاثیر داده پرت قرار نمی‌گیرند [۲۷]، اما در مواردی که تعداد مشاهده‌ها کم باشد، قادر به شناسایی داده‌های پرت نیست [۴۵]. Leys و همکاران (۲۰۱۳) [۳۰] نتیجه گرفتند که روش انحراف مطلق از میانه نسبت به روش انحراف معیار، به داده‌ها پرت حساس نیست و از میانگین و انحراف معیار دقیق‌تر است.

نتایج آزمون کای اسکوئر و میانگین پیراسته (۲۰ درصد) مشابه یکدگیر بود و مقادیر پوشش ۱۰۰ درصد و ۷۰ درصد در دو مرتع با شدت چرای سبک و متوسط را داده پرت تشخیص داد. Komsta (۲۰۲۲) [۲۸] استفاده از آزمون کای اسکوئر را در تشخیص داده پرت تک متغیره توصیه نکردند. نتایج این تحقیق نشان داد که از بین روش‌های مورد استفاده، روش مقادیر صدک‌های ۹۹ و ۹۷/۵ بیشترین تعداد داده پرت را تشخیص داد و احتمالاً با افزایش مقادیر صدک (مثلا ۹۵ درصد) تعداد داده بیشتری به عنوان داده پرت تشخیص داده می‌شود. میانگین پیراسته و وینزوری در مرتع با شدت چرای متوسط، بزرگتر از میانگین حسابی بود، اما در دو مرتع دیگر چندان تفاوتی بین میانگین‌ها مشاهده نشد. Buckley و Georgianna (۲۰۰۱) [۹] بیان کردند که میانگین پیراسته و وینزوری به عنوان روش‌های حذف داده پرت محسوب می‌شوند، اما این روش‌ها، میانگین را اندکی افزایش و واریانس مجموعه

[1] Robust estimation

داده‌ها را کاهش دادند. اگر چه میانگین متداول‌ترین شاخص آمار توصیفی است که در ارزیابی‌های زیست‌محیطی استفاده می‌شود، اما در مواردی که توزیع داده‌ها دارای چولگی است، توصیف کننده خوبی نیست و استفاده از آن بدون بیان انحراف معیار یا خطای معیار از میانگین می‌تواند گمراه کننده باشد [۳۱].

علت وجود داده پرت در مجموعه داده‌های پوشش گیاهی در مطالعات آنالیز مرتع را به می‌توان به دلایل مختلفی نسبت داد، به عنوان مثال خطاهای انسانی (خطا در وارد کردن داده‌ها)، خطاهای اندازه‌گیری (خطای ابزار نمونه‌برداری)، خطای دستکاری داده‌ها (خطای پیش پردازش داده‌های معیوب)، خطای نمونه‌برداری (نمونه‌برداری از پوشش گیاهی ناهمگن) می‌تولند ایجادکننده داده پرت باشد [۲۹]. حتی روش نمونه‌برداری تصادفی نیز می‌تواند در این مسئله نقش داشته باشد. در مراتع مناطق خشک و بیابانی که پوشش گیاهی پراکنده است، در اثر نمونه‌برداری تصادفی ممکن است، قطعه نمونه به صورت تصادفی در محلی قرار گیرد که سطح زمین یا کاملاً لخت و عاری از پوشش گیاهی است و یا روی توده‌ای از پوشش گیاهی قرار گیرد که درصد پوشش بالایی داشته باشد [۴۳]. از آنجایی که رویشگاه مراتع مورد مطالعه درختچه‌ای و از نوع تاغ و رمس بود، تعدادی از پلات‌های مورد مطالعه دقیقاً روی تاج پوشش متراکم تاغ افتاده و ۱۰۰ درصد پلات را پوشش گیاهی شامل شده است. شایان ذکر است که اندازه پلات براساس دستورالعمل طرح پایش کیفی مراتع کشور [۴۳] تعیین شد. به نظر می‌رسد اگر اندازه پلات به روش تجربی (دو برابر قطر تاج پوشش بزرگترین گونه) تعیین شود، حداکثر پوشش گیاهی، احتمالاً کمتر از ۱۰۰ درصد می‌شد. وجود داده پرت (۰ یا ۱۰۰ درصد) در نمونه‌برداری‌های تصادفی، امری طبیعی است [۳]. Christy و همکاران (۲۰۱۵) [۱۳] نشان دادند که در تحقیقات زیستی حداقل ۱۰ درصد از مجموعه داده‌های زیستی، داده اشتباه یا گمشده است. احتمالاً افزایش شدت نمونه‌برداری (تعداد نمونه یا اندازه پلات) به سرشکن کردن خطاهای نمونه‌برداری کمک خواهد کرد [۲۹].

وجود داده پرت همیشه ناشی از خطای انسانی و نمونه‌برداری نیست و تغییرپذیری ذاتی بین داده‌های جمع آوری شده از توزیع با چولگی زیاد نیز می‌تواند علت پرت بودن داده‌ها باشد. یک توزیع اریب ممکن است مانند نقاط پرت به نظر برسد، اما این‌ها دنباله‌های واقعی هستند. دنباله یک توزیع، بخشی از توزیع است که از قسمت مرکزی توزیع دور است، اما پرت نیستند. در چنین مواقعی استفاده از روش گرافیکی مثل هیستوگرام یا نمودار جعبه‌ای به تنهایی شاید نمی‌تولند تفاوت بین داده پرت و دنباله توزیع را نشان دهد. از این‌رو توصیه می‌شود، معنی‌داری ضرایب چولگی توسط آزمون‌های آماری مربوطه بررسی شود، درصورتی که معنی‌داری چولگی تایید شد، دنباله توزیع می‌تولند به عنوان داده پرت در نظر گرفته شود. زمانی که تعداد داده پرت بیش از یک مورد باشد، آزمون‌های چولگی و کشیدگی نسبت به روش تک متغیره مثل گرابز از دقت و توان آماری بالاتری در تشخیص داده‌های پرت برخوردارند [۳۲].

دلیل استفاده از روش‌های متعدد در تحقیق حاضر بدین خاطر بود که نشان داده شود استفاده از روش‌های مختلف، به نتایج متفاوتی می‌انجامد. عدم وجود داده پرت توسط برخی از روش‌های تحقیق حاضر، نشان دهنده نبود داده پرت نیست، بلکه آن روش قادر به تشخیص داده پرت نبوده است. از این رو قبل از انجام هرگونه آزمون فرضیه مقایسه‌ای، استفاده ترکیبی از دو روش بصری و آماری جهت بررسی وجود یا عدم وجود داده پرت توصیه می‌شود.

در آزمون‌های فرضیه مقایسه‌ای، پس از انجام آزمون مربوطه مثل تجزیه واریانس در صورت معنی‌داری F، نیاز به مقایسه میانگین‌ها است، از آنجایی‌که میانگین جزو یکی از شاخص‌های آماری مرکزی است که به شدت تحت تاثیر وجود داده‌های پرت قرار می‌گیرد [۵۰]، تشخیص داده پرت و تصمیم‌گیری در خصوص حذف، تبدیل یا ابقای آن قبل از هر تحلیل آماری ضرورت دارد. ساختار آزمون تجزیه واریانس به‌گونه‌ای است که در صورت حضور

داده‌های پرت، نتایج آزمون می‌تواند به اشتباه منجر به رد یا پذیرش فرض صفر شود [۱۴]. Dadkhah و Samadi Tudar (۲۰۱۸) [۱۴] نشان دادند که استنباط در تجزیه واریانس کلاسیک در حضور داده‌های پرت می‌تواند گمراه کننده باشد.

در سایر موارد نیز وجود داده پرت ممکن است اثرات غیرقابل پیش بینی بر نتایج آزمون‌های آماری داشته باشد. این اثرات زمانی بدتر می‌شوند که نقاط پرت در نتیجه اندازه‌گیری‌های غیرتصادفی (مثل سیستماتیک) با تکرار کم باشد. اگر تعداد داده‌های پرت بسیار کم باشد، معمولاً حذف می‌شوند یا با میانگین داده‌ها جایگزین می‌شوند [۱۹]. با این حال، اگر آنها نتیجه توزیع با دم بلند باشند (مثل چولگی مثبت یا منفی)، ممکن است محقق مجبور باشد از آزمون‌های آماری استفاده کند که به داده های پرت خیلی حساس نیستند. Benhadi-Marín (۲۰۱۸) [۷] گزارش کرد که حذف سیستماتیک یا تبدیل داده‌ها می‌تواند منجر به نادیده گرفتن فرآیندهای مهم اکولوژیکی و نتیجه‌گیری اشتباه شود. به عنوان مثال در زمینه ارزیابی تنوع زیستی، تفسیر نادرست نتایج به دلیل پردازش نادرست داده‌های پرت ممکن است تصمیم‌گیری را دشوار کند یا حتی منجر به شکست در اتخاذ بهترین برنامه مدیریتی شود [۷]. André (۲۰۲۲) [۵] پیشنهاد کرد که حذف داده پرت برخلاف منطق آزمون فرضیه است و این عمل منجر به افزایش مثبت‌های کاذب می‌شود.

در تحقیق حاضر، وجود و حذف داده پرت بر روی همگنی واریانس‌ها نیز بررسی شد. آزمون همگنی واریانس‌ها یکی از پیش شرط‌های اصلی آنالیز واریانس است [۵۰]. آزمون بارتلت، یکی از آزمون‌هایی است که حساس به انحراف از توزیع نرمال است [۳۹]، در این تحقیق، از آزمون ناپارامتری لیون استفاده شد. Nordstokke و Zumbo (۲۰۱۰) [۳۷] توصیه کرد هنگامی که توزیع یکی از جمعیت‌های نمونه‌برداری شده دارای چولگی باشد (مثل مرتع با شدت چرای متوسط در تحقیق حاضر) ، آزمون ناپارمتری مثل لیون قدرت آماری بیشتری دارد. نتایج نشان داد که اگرچه حذف داده پرت، منجر به نرمال شدن داده‌ها نشد، اما واریانس گروه‌ها را همگن کرد.

در مطالعات جامعه شناسی گیاهی که با داده‌های چندمتغیره کار می‌کند، شرایط بسیار سخت‌تر از مطالعات تک متغیره است. مطالعات بوم‌شناختی اغلب شامل تعداد زیادی متغیر و مشاهده است و اغلب در معرض خطاهای مختلفی هستند. به دلیل ماهیت چند متغیره داده‌های بوم شناختی، شناسایی داده‌های پرت با استفاده از رویکردهایی مانند نمودارهای تک متغیره یا دو متغیره دشوار است. این مشکل مستلزم استفاده از روش‌های آماری قوی در شناسایی داده‌های پرت است [۲۳]. در تحلیل های داده‌های تک متغیره و چند متغیره پوشش گیاهی در مطالعات آنالیز و ارزیابی مرتع که شامل داده های پرت هستند، همواره حذف داده پرت، درست‌ترین راه حل نبوده و توصیه می‌شود در آزمون فرضیه‌های مقایسه‌ای و رابطه‌ای، از برآوردگرهای استوار که حساسیت کمتری در مواجهه با داده‌های پرت دارند استفاده شود.

References

[1] Abdolalizadeh, Z., Ghorbani, A., Mostafazadeh, R., and Moameri, M. (2019). Evaluation of the relationship between the quantitative characteristics of vegetation and rangeland condition in the northern rangelands of in Ardebil province. Journal of Range and Watershed Management, 72(1), 167-182.

[2] Aguinis, H., Gottfredson, R. K., and Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. Organizational Research Methods, 16(2), 270–301.

[3] Ahmadi, M. and Sarmad, M. (2010). Detecting Outliers in Normal Data Using Modified Z-Scores. Journal of Statistical Sciences, 3 (2), 119-139.

[4] Amiri, F. and Arzani, H. (2019). Suitability Model of Medical and Industrial Plants of Semirom Rangelands in Isfahan. Journal of Range and Watershed Management, 72(1), 15-28.

[5] André, Q. (2022). Outlier exclusion procedures must be blind to the researcher's hypothesis. Journal of experimental psychology. General, 151(1), 213–223.

[6] Arzani, H. and Abedi, M. (2015). Rangeland Assessment: Vegetation Measurement, University of Tehran Press, 305 p.

[7] Benhadi-Marín, J. (2018). A conceptual framework to deal with outliers in ecology. Biodivers Conserv, 27, 3295–3300.

[8] Bihamta, M. R. and Zare Chahkoei, M. A. (2011). Principles of statistics for the natural resources sciences. University of Tehran Press. 300p.

[9] Buckley, J. A. and Georgianna, T. D. (2001). Analysis of statistical outliers with application to whole effluent toxicity testing. Water Environment Research, 73(5), 575–583.

[10] Carter, N. J. , Schwertman, N. C. and Kiser, T. L. (2009). A comparison of two boxplot methods for detecting univariate outliers which adjust for sample size and asymmetry. Statistical Methodology, 6(6), 604-621.

[11] Chinipardaz, R. and Kamranfar, H. (2009). Effect of Different Types of Outliers on GARCH Models. Journal of Statistical Sciences,3 (1) ,31-46.

[12] Cho, H., Lee, K. and Ahn, S. (2016). Impact of Outliers on the Statistical Measures of the Environmental Monitoring Data in Busan Coastal Sea. Ocean and Polar Research, 38, 149-159.

[13] Christy, A., Gandhi, M.G. and Vaithyasubramanian, S. (2015). Cluster based outlier detection algorithm for healthcare data. Procedia Comput. Sci., 50, 209–215.

[14] Dadkhah, K. and Samadi Tudar, E. (2018). Robust Analysis of Variance based on Permutation Distribution of Trimmed Mean. Journal of Statistical Sciences, 12 (1),119-141.

[15] Dallmeier, F., Szaro, R. C., Alonso, A., Comiskey, J. and Henderson, A. (2013) Framework for Assessment and Monitoring of Biodiversity. In: Levin S.A. (ed.) Encyclopedia of Biodiversity, second edition, Volume 3, pp. 545-559. Waltham, MA: Academic Press.

[16] Dehghan, S. and Faridrohani, M. (2022). Multivariate Outlier Detection Based on Depth-Based Outlyingness Function. Journal of Statistical Sciences, 15 (2), 443-462.

[17] Dervilis, N., Worden, K. and Cross, E. J. (2015). On Robust Regression Analysis as a Mean of Exploring Environmental and Operational Conditions for SHM Data. Journal of Sound and Vibration, 347, 279-296.

[18] Ebrahimi, A. (2017). Effect of sampling group and life-form on estimation of relationship between forage production and canopy cover. Journal of Range and Watershed Management, 70(1), 19-30.

[19] Emami, H. and Mansoori, P. (2018). Influence Diagnostics in Semiparametric Linear Mixed Measurement Error Models. Journal of Statistical Sciences,11 (2), 219-240.

[20] Garces, H. and Sbarbaro, D. (2011). Outliers detection in environmental monitoring databases, Engineering Applications of Artificial Intelligence, 24(2), 341–349.

[21] Hajibagheri, F., Rasekh, A. and Akhoond, M. R. (2014). Detecting Outliers in Liu Regression Model. Journal of Statistical Sciences, 8 (1),19-36.

[22] Hordo, M., Kiviste, A., Sims A. and Lang, M. (2006). M Outliers and /or measurement errors on the permanent sample plot data. USDA Forest Service - General Technical Report PNW

[23] Jackson D. A. and Chen, Y. (2004). Robust Principal Component Analysis and Outlier Detection with Ecological Data. Environmetrics, 15(2),129-139628.

[24] Jain, R. B. (2010). A recursive version of Grubbs' test for detecting multiple outliers in environmental and chemical data. Clin. Biochem. 43, 1030–1033.

[25] Jolous Jamshidia, E. Yusup, Y., Stephen Kayod, J. and Kamaruddina, M. A. (2022). Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on surface water temperature. Ecological Informatics 69, 101672.

[26] Kitzes, J. (2022). Handbook of Quantitative Ecology. University of Chicago Press.

[27] Kolbaşi, A., and Ünsal A. (2019). A Comparison of the Outlier Detecting Methods: An Application on Turkish Foreign Trade Data. Journal of Mathematics and Statistical Science, 5, 213-234.

[28] Komsta, L., (2022). outliers: Tests for Outliers. R package version 0.15. https://CRAN.R-project.org/package=outliers

[29] Krebs, C. J. (2014). Ecological Methodology, 3rd ed .Addison-Wesley Educational Publishers, Inc.

[30] Leys, C., Ley, C., Klein, O., Bernard, P. and Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49, 764–766.

[31] Lintott, P. R., and Mathews, F. (2018). Basic mathematical errors may make ecological assessments unreliable. Biodiversity and conservation, 27(1), 265–267.

[32] Livesey, J. H. (2007). Kurtosis provides a good omnibus test for outliers in small samples. Clinical biochemistry, 40(13-14), 1032–1036.

[33] Moghaddam, M R. (2001). Quantitative plant ecology. University of Tehran press. 285p.

[34] Mouret, F., Albughdadi, M., Duthoit, S., Kouamé, D., Rieu, G. and Tourneret, J-Y. (2021). Outlier Detection at the Parcel-Level in Wheat and Rapeseed Crops Using Multispectral and SAR Time Series. Remote Sensing, 13(5), 956.

[35] Mowbray, F. I., Fox-Wasylyshyn, S. M. and El-Masri, M. M. (2019). Univariate Outliers: A Conceptual Overview for the Nurse Researcher. The Canadian journal of nursing, 51(1), 31–37.

[36] Nicolae-Marius, J. (2014). Software solutions for identifying outliers, Computational Methods in Social Sciences (CMSS), 2(2), 5-14.

[37] Nordstokke, D. W. and Zumbo, B. D. (2010). A new nonparametric Levene test for equal variances. Psicológica, 31(2), 401–430.

[38] Norouzi, A., haghiyan, I. and Sheidai Karkaj, E. (2020). Rangeland management plans and rangeland health (Case Study: Rangelands of Torbat-e-Heydarieh). Journal of Range and Watershed Management, 72(4), 1131-1145.

[39] Odoi, B., Twumasi-Ankrah, S., Samita, S. and Al-hassan, S. (2022). The Efficiency of Bartlett's Test using Different forms of Residuals for Testing Homogeneity of Variance in Single and Factorial Experiments-A Simulation Study. Scientific African, 17, e01323.

[40] R Core Team, (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[41] Rasekh, A., Mansouri, B. and Hedayatpoor, N. (2019). Outlier Detection in Ridge Regression Model Under Stochastic Linear Restrictions. Journal of Statistical Sciences, 13 (1) ,117-137.

[42] Roozbeh, M. and Amini, M. (2020). Feasible Generalized Rdge Robust Estimator in Semiparametric Regression Models. Journal of Statistical Sciences, 13 (2), 441-460.

[43] Rostampour, M. (2022). Rangeland Ecosystems Monitoring in different climatic regions of Iran, South Khorasan Province, Khosf Site. Research Institute of Forests and Rangelands.

[44] Saleem, S., Aslam, M. and Shaukat, M. (2021). A review and empirical comparison of univariate outlier detection methods. Pakistan Journal of Statistics, 37 (4), 447-462.

[45] Seo, S. (2006). A review and comparison of methods for detecting outliers in univariate data sets (master's thesis). Pittsburgh: University of Pittsburgh.

[46] Shimizu, Y. (2022). Multiple Desirable Methods in Outlier Detection of Univariate Data with R Source Codes. Frontiers in psychology, 12, 819854.

[47] Torres, J. M., Pastor Pérez, J., Sancho Val, J., McNabola, A., Martínez Comesaña, M. and Gallagher, J. A. (2020). functional data analysis approach for the detection of air pollution episodes and outliers: A case study in Dublin, Ireland. Mathematics, 8, 225.

[48] Wang, H., Bah, M. J. and Hammad, M. (2019). Progress in Outlier Detection Techniques: A Survey. IEEE Access, 7, 107964–108000.

[49] Wheater C. P. , Bell J. R. and Cook P. A. (2011). Practical Field Ecology: A Project Guide. Wiley-Blackwell. 389 p.

[50] Zar, J.H. (2010). Biostatistical Analysis, 5th ed. Pearson Prentice Hall: Upper Saddle River, NJ.

Journal of Range and Watershed Management
Volume 75, No. 4, Winter 2023

pp. 639-660

Accepted: 2022-11-07

Received: 2023-01-19

Comparison of outlier data detection methods and their impact on rangeland measurement and assessment studies

❖ **Moslem Rostampour*;** Assistant Professor, Department of Rangeland and Watershed Management and Research Group of Drought and Climate Change, Faculty of Natural Resources and Environment, University of Birjand, Birjand, Iran

Abstract

This study compared of univariate outlier data detection methods among vegetation data in a study of the effect of grazing intensity in the rangelands of arid regions. For this purpose, after measuring the vegetation cover in the rangeland and before the statistical analysis, the presence of outlier data was examined as the assumption of parametric comparison tests. In this study, eight methods including the boxplot and IQR (Tukey method), standard deviation of the mean (three-sigma rule), median absolute deviation (Hampel method), trimmed mean, 1st percentile and 99th percentile, The Chi Square test (χ^2), the Grubbs Test (ESD) and the Rosner test (generalized ESD) were used. The results showed that the vegetation cover of rangelands with light and moderate grazing intensity was not normally distributed (Shapiro-Wilk test: $p \leq 0.05$). Even deletion of outliers did not lead to a normal distribution, but it resulted in the homogeneity of variances (Levene's test: $p \geq 0.05$). The modified Z-score and the Grubbs and Rosner tests ($p \geq 0.05$) did not identify outliers from the vegetation cover data. Among the methods evaluated, the boxplot and MAD method, which are not dependent on the mean, are more suitable for the vegetation cover. Therefore, before performing any comparison test, a combination of visual and statistical methods is recommended to evaluate the presence of outliers.

Keywords: mean, outliers, parametric statistics, rangeland, vegetation.

*Corresponding Author: Tel: +98 9151637869 Email: rostampour@birjand.ac.ir