همبستگی کاذب و خاصیت بسته بودن داده‌های ترکیبی در علوم زمین

چکیده:

داده‌های ترکیبی که معمولا نتیجه اندازه‌گیری‌ها در علوم زمین هستند، خاصیت مهمی به نام بسته بودن دارند. پژوهشگرانی که بدون توجه به این خاصیت، از روش‌های متداول آماری با اعمال تبدیل لگاریتمی برای تعدیل چولگی و یا نرمال‌کردن داده‌ها استفاده می‌کنند در واقع وجود همبستگی کاذب در داده‌های ترکیبی را نادیده می‌گیرند که این باعث نتایج آماری نادرست می‌شود. در این مقاله بعد از معرفی داده‌های ترکیبی و خاصیت بسته بودن آنها، تبدیل‌هایی برای باز کردن سیستم بسته داده‌ها معرفی شده‌اند. این تبدیل‌ها عبارت‌اند از تبدیل نسبت لگاریتمی جمعی، تبدیل نسبت لگاریتمی مرکزی شده و تبدیل نسبت لگاریتمی ایزومتریک که همگی برحسب لگاریتم نسبت‌ها تعریف می‌شوند. بعد از معرفی و برشمردن مزایا و معایب این تبدیل‌ها نسبت به همدیگر، یکی از آنها به نام تبدیل clr روی یک مجموعه داده مربوط به آنالیز شیمیایی خاک اعمال شده است. بعلاوه نتایج اعمال تحلیل خوشه‌ای بر داده‌های تبدیل شده با استفاده از ماتریس ضرایب همبستگی اسپیرمن به عنوان ماتریس فاصله مورد بررسی قرار گرفته است. همچنین تأثیر اعمال تبدیل clr بر حذف همبستگی کاذب، تعدیل چولگی و نقاط پرت در داده‌ها با کمک برخی نمودارهای آماری و با استفاده از نرم‌افزار آماری R بررسی شده است.

واژه‌های کلیدی: داده‌های ترکیبی و خاصیت بسته بودن آنها، تبدیل‌های لگاریتم نسبتی، روش‌های آماری استوار ، همبستگی کاذب

مقدمه:

در علم آمار، داده‌های ترکیبی[1] بخش‌هایی از یک کل را توصیف می‌کنند و به‌صورت بردارهایی از مقادیر نسبت، درصد، عیار یا فراوانی ثبت می‌شوند. در واقع این نوع داده‌ها شامل مشاهدات چند متغیره با مقادیر مثبت هستند که مجموع آنها یک مقدار ثابت است، یعنی آنچه برای تحلیل پیش‌رو داریم اطلاعات نسبی است نه مطلق. به عنوان مثال، اگر تمام عناصر شیمیایی موجود در یک نمونه از خاک از نظر عیار به‌طور کامل تجزیه و تحلیل شوند، مجموع این عیارها در هر نمونه برابر ۱۰۰۰۰۰۰mg/kg می‌شود. واضح است که یک مجموعه داده، تنها زمانی می‌تواند ترکیبی باشد که حداقل متشکل از دو جزء باشد. هنگام کار با داده‌های ترکیبی حتی در موردی که فقط یک جزء را گزارش می‌کنیم، در واقع به طور ضمنی آن را یا به یک کل از پیش تعریف شده یا به یک قسمت مکمل مرتبط کرده‌ایم. به‌عنوان مثال، عیار اندازه‌گیری شده یک عنصر در یک سنگ معدنی، بدون اطلاع از عیار دیگر عناصر، به‌تنهایی حاوی اطلاعاتی نیست یا بودجه تخصیصی دولت در آموزش و پرورش به خودی خود بی‌معنی است بدون اینکه بدانیم کل بودجه چقدر بوده یا چقدر بودجه برای دیگر وزارتخانه‌ها در نظر گرفته شده است. یک بردار از مشاهدات ترکیبی، x، که نمونه‌ای شامل n جزء است، به‌صورت زیر بیان می‌شود:

$$x=(x_1,x_2,\cdots,x_n), x_i>0, \sum_{i=1}^{n} x_i=\kappa, \qquad (۱)$$

با این تعریف، از نظر آماری دو بردار ax و x به ازای هر ثابت مخالف صفر a دارای اطلاعات یکسانی هستند. به‌عبارت دیگر اطلاعات مورد نیاز در نسبت بین اجزاء قرار دارد نه در خود آنها. خاصیت مجموع ثابت که در

[1] Compositional Data

ادبیات داده‌های ترکیبی به خاصیت بسته‌بودن داده‌ها[2] معروف است، به معنای تفاوت قابل توجه ماهیت داده‌های ترکیبی و بسیاری از دیگر داده‌های چند متغیره می‌باشد.

از آنجایی که داده‌های ترکیبی نیز به‌صورت اعداد بیان می‌شوند، ممکن است فرد بر حسب عادت آنها را به عنوان داده‌های چند متغیره عادی تفسیر یا حتی تجزیه و تحلیل کند. این عمل می‌تواند منجر به پارادوکس‌ها و یا تفسیرهای نادرستی شود، که برخی از آنها مانند همبستگی کاذب[3] حتی از بیش از یک قرن پیش به خوبی شناخته شده‌اند (Pearson, 1987). پیرسون نشان داد که دو متغیری که هیچ همبستگی بین خود ندارند، با تقسیم بر یک مخرج مشترک، همبستگی پیدا می‌کنند (به‌طور مثال همبستگی حاصل بین عیارهای اندازه‌گیری شده دو عنصر). وی همبستگی حاصل بین چنین متغیرها یا به‌عبارت بهتر نسبتها را همبستگی کاذب نامید. همچنین پیرسون، در این مقاله کلاسیک، به خطراتی اشاره کرده است که متوجه تحلیل‌گری است که سعی در تفسیر همبستگی بین نسبت‌هایی دارد که صورت‌ها یا مخرج‌های آن دارای اجزای مشترک هستند. زمان زیادی طول کشید تا زمین‌شناسان متوجه همبستگی کاذب در داده‌های ترکیبی شدند (Chayes, 1960). هر چند بحث در باره این نوع همبستگی بارها در متون آماری نظیر (Pendleton et al., 1983) مورد توجه قرار گرفته است، با وجود این در طول سال‌ها این خطرات، بیشتر فراموش و یا نادیده گرفته شده‌اند، تا اینکه اولین بار در سال ۱۹۸۶ برای تحلیل آماری داده‌های ترکیبی نظریه پردازی شد (Aitchison, 1986).

داده‌های زمین‌شناسی به‌طور معمول دارای مقادیر گمشده هستند، نقاط پرت دارند و از توزیع نرمال پیروی نمی‌کنند (حیدریان دهکردی و همکاران، ۱۳۹۶; محمدی‌اصل و همکاران، ۱۳۹۹; حسین‌پور نجاتی و همکاران، ۱۴۰۰) . در این مقاله خاصیت دیگری از این داده‌ها معرفی شده که به آن خاصیت بسته بودن می‌گوییم که باعث ایجاد همبستگی کاذب در داده‌های می‌شود. بسیار دیده می‌شود پژوهشگران، برای کاهش چولگی و نرمال کردن داده‌ها، ابتدا یک تبدیل لگاریتمی روی داده‌ها اعمال و در ادامه از روش‌های متداول آماری برای تحلیل داده‌ها استفاده می‌کنند . در این مقاله روی این نکته تاکید شده است که داده‌های زمین شناسی دارای خاصیت بسته بودن هستند که باعث مشکلی به نام همبستگی کاذب در داده‌ها می‌شود. با اعمال تبدیل لگاریتمی، خاصیت بسته بودن داده‌ها و در نتیجه همبستگی کاذب برطرف نمی‌شود که این باعث نتایج آماری غیر قابل قبول می‌گردد. در این مقاله، تبدی‌های نسبت لگاریتمی معرفی می‌شوند که اعمال انها باعث باز شدن سیستم داده‌ها و رفع مشکل همبستگی کاذب و در نتیجه تحلیل‌های آماری صحیح می‌شود.

همبستگی کاذب و علیت

ضریب همبستگی پیرسون یک شاخص است که اطلاعاتی در مورد رابطه خطی یا یکنواخت بین دو متغیر که با مقیاس فاصله‌ای یا نسبتی (متغیر از نوع پیوسته) اندازه گیری شده‌اند، ارائه می‌دهد همبستگی کاذب، طبق تعریق یک همبستگی معنی‌دار آماری بین دو متغیر است که به دلیل نگرش تحلیل‌گر به متغیرها و نحوه مدیریت آنها به‌دست آمده است نه وجود هرگونه رابطه ذاتی بین متغیرها. در واقع این همبستگی معنادار لزوماً به این معنی نیست که بین دو متغیر همبسته یک رابطه علت و معلولی وجود دارد یا اینکه این دو متغیر با یک پیوند ذاتی به‌هم مرتبط هستند. بسیاری از رویدادها وجود دارند که ارتباط آنها ضریب همبستگی

[2] Closed Data
[3] Spurious Correlation

بزرگی را ایجاد می‌کند، اما نتیجه‌گیری اینکه یکی باعث دیگری شده است، درست نیست. در اینجا چند نمونه آورده شده است:

- آمار نشان می‌دهد شانس زنده ماندن از اولین حمله قلبی در افراد سیگاری نسبت به افراد غیرسیگاری بسیار بیشتر است. آیا سیگار برای سلامتی انسان مفید است؟
- محققان یک همبستگی قوی و مثبت بین اندازه کفش و میزان مطالعه افراد پیدا کردند. آیا دلیل مطالعه بیشتر اندازه پا است؟
- بین میزان فروش بستنی و تعداد غرق‌شدگان، همبستگی بالایی وجود دارد. آیا برای کاهش تعداد مرگ و میر در اثر غرق‌شدگی باید مصرف بستنی را کاهش داد؟

چرا همبستگی های کاذب رخ می‌دهد؟ گاهی همبستگی بی‌دلیل رخ می‌دهد که به‌عنوان همبستگی‌های بی‌معنی شناخته می‌شوند. اما همبستگی‌های کاذب همیشه بی‌دلیل به وجود نمی‌آیند بلکه حاصل آماده‌سازی یا عدم جمع‌آوری مناسب داده‌ها است. اغلب، دو متغیر به‌دلیل متغیر سومی که بر هر دو تأثیر می‌گذارد، با هم مرتبط هستند. به‌عبارت دیگر عامل یا متغیر دیگری که این دو متغیر را به‌هم مرتبط می‌کند، هنگام بررسی و اندازه‌گیری این دومتغیر نادیده گرفته شده است. به این متغیر که علت مشترک تغییرات همسو در دو متغیر است، متغیر میانجی[4] گفته می‌شود. این زمانی اتفاق می‌افتد که در ظاهر، متغیر x با متغیر y ارتباط داشته باشد ولی این همبستگی به دلیل وجود یک متغیر میانجی است که همزمان تغییرات همسوی دو متغیر را توجیه می‌کند. در مثال‌های بالا در خصوص توجیه همبستگی‌های کاذب بالا لازم به ذکر است افراد سیگاری از اولین حمله قلبی خود بیشتر از افراد غیر سیگاری جان سالم به‌در می‌برند، زیرا سیگاری‌ها معمولاً اولین حمله قلبی خود را در سنین پایین‌تری تجربه می‌کنند. همچنین افراد بزرگسال بیشتر از کودکان مطالعه می‌کنند لذا سن بالاتر معادل اندازه کفش بالاتر و ساعت مطالعه بیشتر است. معمولا در هوای گرم از یک طرف میزان مصرف بستنی و از طرف دیگر تعداد غرق‌شدگی افزایش می‌یابد. پس در مثال‌های بالا نقش سن و دمای هوا به عنوان متغیر میانجی نادیده گرفته شده است. توجه کنید، وقتی پژوهشگر به وجود همبستگی کاذب بین دو متغیر ظنین است، بایستی با شناسایی متغیر میانجی، از ضریب همبستگی جزئی استفاده کند، ولی در علوم زمین (مانند ژئوشیمی) و در مهندسی (مانند مهندسی معدن و خاک) وقتی به‌طور مثال با داده‌های ترکیبی مانند درجه خلوص عناصر در یک نمونه سر و کار داریم، به‌طور معمول با استفاده از تبدیلات مناسبی که در ادامه به آنها اشاره خواهیم کرد. سیستم به‌اصطلاح بسته داده‌ها را باز، سپس تحلیل‌های آماری مورد نظر را روی داده‌های باز شده پیاده می‌کنند.

پیشینه تحقیق

از حدود شش دهه گذشته، پژوهشگران مشکلات مربوط به تحلیل آماری سیستم اعداد بسته (مانند داده‌های ترکیبی) را مورد بحث قرار داده‌اند (Miesch and Chapman, 1977; Aitchison, 1986; Filzmoser, et al., 2018). هنگام کار با داده‌های ترکیبی یکی از مشکلات چولگی شدید داده‌ها و همچنین غیر نرمال بودن آنها است. در اولین گام مربوط به اماده‌سازی داده‌ها قبل از مدل‌بندی مبتنی بر ماتریس همبستگی که اساس بررسی ارتباط بین متغیرها است، نیاز به یک تبدیل احساس می‌شود. نویسندگان در

[4] Mediator Variable

(Filzmoser and Hron, 2009) توضیح داده‌اند که چرا در مرحله آماده سازی برای محاسبه ماتریس همبستگی، نیاز داریم با یک تبدیل، داده‌های ترکیبی را به یک فضای نمونه مناسب منتقل کنیم. برای فهم ضرورت این کار، توجه کنید که در رابطه (۱) مقدار ثابت κ که باعث بسته بودن سیستم داده‌های ترکیبی است به خودی‌خود اهمیت ندارد، چون مقدار این ثابت با تغییر واحد اندازه‌گیری مشاهدات تغییر می‌یابد، آنچه تعیین کننده است مقیاس اندازه‌گیری می‌باشد، موضوعی که هنگام استفاده از ضریب همبستگی مشکل‌ساز خواهد شد. توجه کنید

چون $\sum_{i=1}^{n} x_i = \kappa$، به ازای هر i داریم:

$$cov\left(\sum_{i=1}^{n} x_i, x_i\right) = cov(\kappa, x_i) = 0.$$

به‌عبارت دیگر جمع هر سطر یا ستون ماتریس واریانس-کواریانس داده های ترکیبی برابر صفر می‌شود. به‌عنوان مثال برای سطر اول این ماتریس داریم:

$$cov(x_1, x_2) + cov(x_1, x_3) + \cdots + cov(x_1, x_n) = -var(x_1).$$

با توجه به اینکه مقدار واریانس همیشه مثبت است، رابطه بالا باعث می‌شود مقدار برخی کواریانس‌ها نه به خاطر وجود تغییرات در جهت معکوس بلکه به‌دلیل خاصیت بسته بودن داده‌ها، منفی شود. این موضوع در تحلیل داده‌های ترکیبی به مسئله اریبی منفی[5] معروف است که نخستین بار توسط (Pearson, 1897) و بار دیگر توسط (Chayes, 1960) بیان شد.

به‌دلیل نقش اساسی ماتریس واریانس-کواریانس (یا همبستگی) که در روش‌های آماری چند متغیره، می‌توان انتظار داشت نادیده گرفتن خاصیت بسته بودن داده های ترکیبی، ممکن است چه تلاش بیهوده‌ای برای تفسیر آماری نتایج نادرست به دلیل همبستگی‌های بی‌دلیل را شکل دهد.
اینجاست که ضرورت اعمال یک تبدیل مناسب روی داده‌های ترکیبی، پیش از هرگونه تحلیل آماری مبتنی بر همبستگی مشخص می‌شود، تا داده‌های ترکیبی به یک فضای نمونه غیر محدود تبدیل شوند.

مواد و روش‌ها

داده‌ها

در این بخش با استفاده از یک نمونه داده ژئوشیمیایی مواردی نظیر وجود همبستگی کاذب و اعمال تبدیل برای باز کردن داده‌ها و تاثیر آن مورد بررسی قرار می‌گیرد. داده‌ها مربوط به عیار عناصر اصلی Zn، Pb، Mo، Ba، Co، Sb، As، Cu، Ni و Au حاصل از ۴۳۳ نمونه از رسوبات آبراهه‌ای بوده است. منطقه مورد مطالعه در مختصات طول جغرافیایی ′۳۰ °۵۱ تا °۵۲ شرقی، عرض جغرافیایی ′۳۰ °۳۳ تا °۳۴ شمالی و ۱۲۰ کیلومتری شمال شهر اصفهان در منطقه نطنز واقع شده و دارای مساحت ۲۵۰۰ کیلومتر مربع بوده است. هدف ما در ادامه بررسی آماری این داده‌ها از منظر آمار توصیفی در راستای مطالب ذکر شده در بخش‌های قبلی بوده است و لذا در این چارچوب موقعیت مکانی داده‌ها وارد تحلیل نشده است. با وجود این اطلاعات

[5] Negative Bias Problem

بیشتر در مورد زمین‌شناسی این داده‌ها در (اعلمی‌نیا و همکاران، ۱۳۹۷) آمده است. داده‌های ژئوشیمیایی متعارف به چهار دسته اصلی تقسیم می شوند که شامل عناصر اصلی، عناصر کمیاب، ایزوتوپ های پرتوزا و ایزوتوپ های پایدار هستند. به‌دلیل ماهیت پیچیده داده‌های ژئوشیمیایی یا به‌طور کلی داده‌های زمین‌شناسی در کاربست روشهای معمول آماری برای تحلیل این داده‌ها محدودیت داریم. دلیل این محدودیت علاوه بر ماهیت ترکیبی بودن این نوع داده‌ها که موضوع اصلی این مقاله است، ناشی از همبستگی مکانی، غیر نرمال بودن، چولگی شدید، وجود نقاط پرت، داده‌های گمشده، داده‌های سانسور شده (به علت دقت دستگاههای اندازه‌گیری که به عیار کمتر از یک حد آستانه مشخص، حساس نیستند) که همگی باعث محدودیت در کاربست روش‌های معمول آماری برای تحلیل داده‌های ژئوشیمیایی می‌شود.

روش تحلیل داده‌ها

در داده‌های حاصل از اندازه‌گیری عیار عناصر در نمونه‌ها مشاهده می‌شود که توزیع عیار اغلب بسیار چوله است و بسیار دیده می‌شود که برخی پژوهشگران پیشنهاد می‌کنند یک تبدیل لگاریتمی توزیع داده‌ها را متقارن یا حتی نرمال می‌کند (Reimann et al., 2008). لازم به‌ذکر است حتی اگر از تبدیل لگاریتمی استفاده شود، این تبدیل سیستم بسته داده‌ها را باز نمی کند و همبستگی‌های بدست آمده بعد از اعمال تبدیل‌هایی غیر آنچه در ادامه به آن اشاره خواهیم کرد، می‌توانند به‌طور جدی گمراه کننده باشند، (Reimann and Filzmoser, 2000). در متون آماری مربوط به تحلیل داده‌های ترکیبی، برای باز کردن سیستم‌های بسته اعداد، سه تبدیل از نوع نسبت لگاریتمی[۶] یعنی تبدیل‌هایی بر حسب لگاریتم نسبت‌ها پیشنهاد شده است، که عبارت‌اند از تبدیل نسبت لگاریتمی جمعی[۷]، تبدیل نسبت لگاریتمی مرکزی شده[۸] و تبدیل نسبت لگاریتمی ایزومتریک[۹] که آنها را به‌ترتیب با تبدیل‌های alr، clr و ilr نمایش می‌دهیم. تبدیل alr و clr توسط (Aitchison, 1986) و تبدیل ilr توسط (Egozcue et al., 2003) تعریف شده‌اند. تبدیل alr به‌این صورت تعریف می‌شود که یکی از متغیرها به‌دلخواه به‌عنوان متغیر مرجع در نظر گرفته شده و بردار مشاهدات جدید به صورت نسبت لگاریتم هر متغیر به متغیر مرجع تعریف می‌شود. متاسفانه این تبدیل خاصیت ایزومتری ندارد، یعنی فاصله بین نقاط در فضای اولیه و فضای تبدیل شده یکسان نیستند. توجه کنید که اغلب شاخص‌های آماری، مانند ضریب همبستگی، بر اساس فواصل اقلیدسی هستند. در تحلیل داده‌های ترکیبی می‌توان از تبدیل clr و از نوع مخصوصی از تبدیل ilr یعنی مختصات محوری[۱۰] برای دریافت اطلاعات نسبی موجود در داده‌ها استفاده نمود. در تبدیل clr، با استفاده از n متغیر موجود، n متغیر جدید ساخته می‌شوند که هر کدام اطلاعات موجود در نسبت‌های لگاریتمی بین همه زوج متغیرهای اولیه را خلاصه می‌کنند. در ادامه فرمول‌بندی یکی از سه تبدیل معرفی می‌شود و خواننده علاقه‌مند برای آشنایی با فرمول‌بندی بقیه تبدیل‌ها می‌تواند به (Filzmoser and Hron, 2008) مراجعه کند. نکته اینجاست که هنگام کار با داده، هرگاه مایل به اعمال یکی از تبدیل‌ها باشیم از امکانات موجود در نرم‌افزار R استفاده می‌کنیم و نگران سختی فرمول‌ها نیستیم (Gerald van den Boogaart and Tolosana-Delgado, 2013).

تبدیل clr روی بردار مشاهدات $x = (x_1, x_2, ..., x_n)$ به‌صورت زیر تعریف می‌شود:

---

[6] Logratio Transformation
[7] Additive Logratio Transformation
[8] Centered Logratio Transformation
[9] Isometric Logratio Transformation
[10] Pivot Coordinate

$$\begin{aligned}\mathrm{clr}(x) &= \left\{\log\left(\frac{x_1}{G(x)}\right),\ldots,\log\left(\frac{x_n}{G(x)}\right)\right\}\\ &= \{\log(x_1)-\log(G(x)),\ldots,\log(x_n)-\log(G(x))\}\end{aligned}$$

که در آن $G(x)$ میانگین هندسی بردار مشاهدات است. توجه کنید با در نظر گرفتن

$$\log(G(x)) = \log\left(\exp\left[\frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right]\right) = \frac{1}{n}\sum_{i=1}^{n}\log(x_i)),$$

داریم:

$$\sum \mathrm{clr}(x) = \sum\left(\log(x_i)-\log(G(x))\right) = 0.$$

مزیت اصلی این تبدیل این است که ارتباط متغیرهای جدید و قبلی حفظ و تفسیر نتایج نسبتاً ساده است. تبدیل clr ایزومتریک است ولی در عین حال، این تبدیل مشکل خاصیت مجموع صفر را دارد که محدودیت در اعمال برخی روش‌های آماری روی داده‌های تبدیل شده، را باعث می‌شود. تبدیل ilr علاوه بر اینکه خاصیت ایزومتری دارد، مشکل هم‌خطی تبدیل clr را نیز مرتفع می‌کند ولی همبستگی بین متغیرهای تبدیل یافته برابر همبستگی متغیرهای اصلی نیست زیرا متغیرهای تبدیل یافته و اصلی، بر اساس یک تابع غیرخطی به‌هم مربوط می‌شوند. هر چند برای این منظور در (Egozcue and Pawlowsky-Glahn, 2005; Filzmoser and Hron, 2009) راه‌حلی مبتنی بر بالانس‌ها[11] مطرح شده است. همچنین (Reimann et al., 2017) نوع خاصی از تبدیل ilr براساس مختصات متقارن را برای تحلیل همبستگی داده‌های ترکیبی ارائه داده‌اند.

در پایان این بخش برای درک اهمیت اعمال این تبدیل‌ها و اینکه ضرورت استفاده از این تبدیل‌ها چگونه است، در ادامه قسمتهای مرتبط از خلاصه مقاله (Filzmoser et al., 2009) نقل به‌مضمون می‌شود. در پژوهش‌های مرتبط، تبدیل‌های ذکر شده برای باز کردن داده‌های بسته به‌ندرت اعمال می‌شوند. این تبدیل‌ها پیچیده‌تر از تبدیل لگاریتمی هستند و به‌واسطه اعمال آنها ارتباط داده‌های تبدیل شده و داده‌های ابتدائی قطع می‌شود. ممکن است از نظر پژوهشگر، نتایج به‌دست‌آمده از اعمال روش‌های آماری متداول روی داده‌های بسته، منطقی به‌نظر برسد و در نتیجه پیامدهای احتمالی کار با داده‌های بسته به‌ندرت مورد سؤال قرار بگیرد. در این مقاله نشان داده می‌شود که مشکل بسته‌بودن داده‌ها باید حتی قبل از معیارهای آماری ساده مانند میانگین یا انحراف معیار یا رسم نمودارهای توزیع داده‌ها مانند هیستوگرام و نمودار جعبه‌ای برطرف شود. برخی از معیارها مانند انحراف معیار (یا واریانس) که با استفاده از داده‌های بسته به‌دست بیایند، از نظر آماری مفهومی ندارند. بنابراین تمام آزمون‌های آماری مبتنی بر انحراف معیار در صورت استفاده با داده‌های اصلی، نتایج اشتباهی به‌دست می‌دهند.

در بخش بعد، وجود همبستگی کاذب، چولگی شدید داده‌ها و وجود داده‌های پرت با استفاده از نمودارهای آماری بررسی شده است. همچنین نشان داده می‌شود که با اعمال تبدیل clr روی داده‌ها، علاوه بر حذف همبستگی‌های کاذب، دو مشکل دیگر یعنی چولگی شدید داده‌ها و نقاط پرت تا حدی اصلاح شده‌اند. همچنین یک تحلیل خوشه‌ای روی داده‌های اصلی و تبدیل شده با استفاده از ماتریس همبستگی اسپیرمن به عنوان ماتریس فاصله، انجام و دندروگرام نتایج به همراه نقشه گرمایی ماتریس فاصله ارائه شده است. هدف مقایسه بررسی تاثیر اعمال تبدیل و انتخاب ماتریس فاصله مناسب بر روی نتایج حاصل از خوشه‌بندی و نشان دادن این موضوع بوده است که با اعمال تبدیل نتایج تا چه میزان نسبت به وقتی تبدیل اعمال نشده، متفاوت بوده‌اند.

[11] Balances

نتایج

جدول ۱، آماره‌های توصیفی مربوط به عیار عناصر را نشان می‌دهد. با توجه به ضریب چولگی ملاحظه می‌شود، توزیع عیار همه عناصر به جز Au دارای چولگی شدید مثبت می‌باشد. شکل‌های ۱ و ۲، همچنین ۳ و ۴ به ترتیب نمودار جعبه‌ای و گراف شبکه‌ای ماتریس همبستگی مربوط عیار عناصر و تبدیل clr روی آن‌ها را نشان می‌دهد.

در گراف داده‌شده رنگ سبز و قرمز به‌ترتیب نشان دهنده همبستگی مثبت و منفی و ضخامت این خطوط شدت همبستگی را نشان می‌دهد. هر چه خطوط کم‌رنگ‌تر باشند، ضریب همبستگی به صفر نزدیک‌تر است. همانطور که انتظار داشتیم به‌دلیل وجود همبستگی کاذب، ساختار همبستگی دو متغیره بین عیارها قبل و بعد از تبدیل بسیار متفاوت است. به‌عنوان مثال همبستگی دو متغیر Au-Sb قبل تبدیل مثبت و ضعیف و بعد از آن شدت همبستگی همچنان مثبت ولی شدت آن افزایش پیدا کرده است. توجه کنید که نقاط پرت روی ساختار ضریب همبستگی بین دو متغیر تأثیر مخرب دارد. لذا در شکل ۴ علیرغم انتظار وجود همبستگی مثبت بین برخی متغیرها، ممکن است برخی همبستگی‌ها ضعیف یا حتی در مواردی همبستگی منفی باشد. دلیل این موضوع که همانطور که از شکل ۲ پیداست به وجود داده‌های پرت برمی‌گردد. داده‌پرت از دو منظر داده‌پرت تک متغیره و چندمتغیره قابل بررسی است. از ملاحظه شکل ۲ فقط می‌توان به وجود داده‌های تک متغیره پی برد. برای بررسی وجود داده پرت چند متغیره می‌توان فاصله ماهالونوبیس را محاسبه کرد، جایی که میانگین و ماتریس واریانس کواریانس موجود در فرمول این معیار با استفاده از روش‌های آماری استوار برآورد می‌شود. خوشبختانه امروزه نرم افزارهای آماری نظیر R برای پژوهشگرانی که علاقه‌مند به مباحث نظری نیستند تا حدی گره‌گشاست. لازم به ذکر است، درصد قابل توجهی از ۴۳۳ داده، دارای فاصله ماهالونوبیس معناداری هستند و از این لحاظ داده‌پرت چند متغیره در این داده‌ها وجود دارد.

در چنین مواردی محاسبه ضریب همبستگی مقاوم به نقطه پرت می‌تواند اثر نقاط پرت را تعدیل کند. توجه کنید معرفی و استفاده از روش‌های آماری استوار[۱۲] موضوع این مقاله نبوده است. با این وجود چون می‌دانیم ضریب همبستگی رتبه‌ای اسپیرمن تحت تأثیر نقاط پرت نیست، در شکل‌های ۵ و ۶، نقشه گرمایی[۱۳] ضریب همبستگی اسپیرمن داده‌ها به همراه دندروگرام[۱۴] حاصل از خوشه‌بندی سلسه‌مراتبی نمایش داده شده است. توجه کنید معیار نزدیکی مورد استفاده معکوس مقدار همبستگی رتبه‌ای بین متغیرها بوده است (یعنی همبستگی بیشتر فاصله کمتر). لازم به ذکر است که نقشه گرمایی همان ماتریس ضریب همبستگی است که مقادیر مثبت و منفی همبستگی با رنگ‌های متفاوت نشان داده شده‌اند به این ترتیب که با افزایش شدت همبستگی (قدر مطلق) رنگ‌ها اشباع‌تر می‌شوند. هر سلول در ماتریس همبستگی با یک مربع در نقشه‌گرمایی نشان داده می‌شود و نقشه گرمایی همانند ماتریس همبستگی حول قطر اصلی‌اش متقارن است. توجه کنید به‌دلیل اینکه نقشه گرمایی خواناتر شود، ترتیب سطر و ستون‌های این نقشه متفاوت از ترتیب سطر و ستون‌های ماتریس داده‌هاست. به‌خصوص متغیرها با ضریب همبستگی نزدیک به‌هم (رنگ‌های مشابه)، طوری جابجا می‌شوند که بلوک‌هایی از رنگ‌های مشابه در نقشه گرمایی ایجاد شود تا ساختار همبستگی منفی درون داده‌ها، بهتر دریافت شود. ملاحظه می‌شود نقشه گرمایی شکل ۵ و ۶ تقریباً از نظر رنگ‌بندی، شبیه به‌نظر می‌رشد

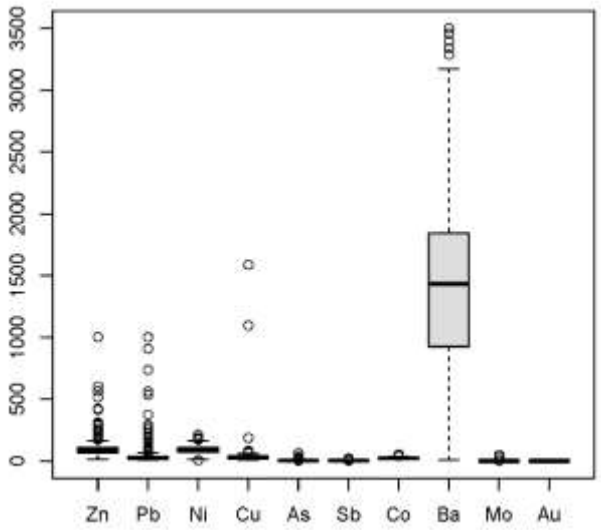
---

[12] Robust Statitical Methods
[13] Heat Map
[14] Dendrogram

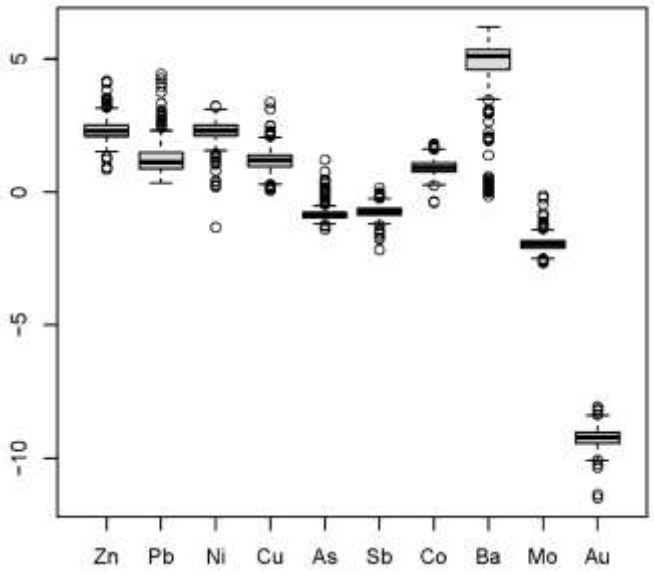
(کامل یکسان نیستند). این موضوع می‌تواند به این دلیل باشد که اولاً داده‌های پرت روی ضریب همبستگی اسپیرمن تأثیر ندارند. از طرفی برخلاف ضریب همبستگی پیرسون که فقط نسبت به تبدیلات خطی پایا است، تبدیل‌های یکنوا روی ضریب همبستگی اسپیرمن تأثیری ندارند (یعنی ضریب همبستگی اسپیرمن دو سری داده و ضریب همبستگی لگاریتم آنها با هم برابر است چون لگاریتم یک تابع صعودی است). لذا وقتی رو داده‌ها تبدیل clr زده‌ایم، این تبدیل روی ضریب همبستگی اسپیرمن تاثیر چندانی نداشته است.

جدول ۱. آماره‌های توصیفی برای عیار عناصر بر حسب ppm

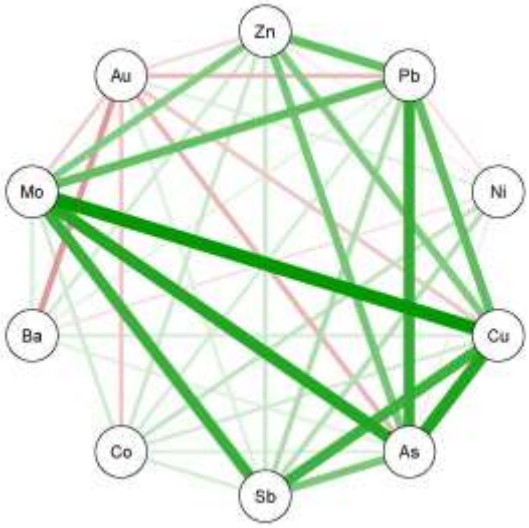
	روی	آنتیموان	سرب	نیکل	مولیبدن	مس	کبالت	باریم	طلا	آرسنیک
میانگین	۱۰۱.۷۴	۴.۱۹	۴۸.۶۷	۹۰.۲۳	۱.۴۶	۳۶.۹۲	۲۲.۸۱	۱۴۵۶.۴۶	۰	۴.۲۳
انحراف معیار	۸۵.۲۴	۰.۹۶	۱۰۸.۰۱	۳۰.۰۸	۲.۶۸	۹۱.۶	۶.۲۷	۸۴۵.۶۱	۰	۴.۲۹
کمینه	۱۳	۱	۱۰	۴.۹۸	۰.۵۲	۸.۱۹	۱۱	۶.۸۱	۰	۱.۸۳
چارک اول	۶۸	۳.۸۲	۲۰	۶۸.۹۲	۱.۰۴	۲۱.۹۷	۱۸	۹۲۳.۳۸	۰	۳.۲۸
میانه	۸۶	۴.۱۴	۲۵	۸۷.۱۸	۱.۲۳	۲۹.۸۶	۲۱	۱۴۳۱.۹۳	۰	۳.۵۷
چارک سوم	۱۰۸	۴.۴۸	۳۸	۱۰۷.۹۵	۱.۴۷	۳۷.۸۵	۲۶	۱۸۴۳.۹	۰	۴.۰۱
بیشینه	۱۰۰۰	۱۹.۴۳	۱۰۰۰	۲۱۰.۱۲	۴۷.۸۹	۱۵۸۵.۳۱	۵۲	۳۵۰۰	۰	۶۰.۲
بزرگی انحراف از میانگین	۲۸.۱۷	۰.۴۸	۱۰.۳۸	۲۹.۰۹	۰.۳۳	۱۱.۷۹	۵.۹۳	۶۸۲.۸۸	۰	۰.۵
دامنه میان چارکی	۴۰	۰.۶۶	۱۸	۳۹.۰۳	۰.۴۳	۱۵.۸۸	۸	۹۲۰.۵۲	۰	۰.۷۳
ضریب تغییرات	۰.۸۴	۰.۲۳	۲.۲۲	۰.۳۳	۱.۸۳	۲.۴۸	۰.۲۷	۰.۵۸	۰.۱۷	۱.۰۱
چولگی	۶.۸۹	۹.۳۱	۷.۰۸	۰.۳۹	۱۴.۹۶	۱۴.۷۱	۱.۳۲	۰.۴۸	-۰.۱۵	۱۰.۳۴
انحراف معیار چولگی	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲
کشیدگی	۶۱.۲۸	۱۴۲.۳۲	۵۴.۰۸	۰.۵۵	۲۳۶.۸۴	۲۲۶.۵۷	۲.۷	۰.۲۲	-۰.۵۷	۱۲۰.۹۵
تعداد داده معتبر	۴۳۳	۴۳۳	۴۳۳	۴۳۳	۴۳۳	۴۳۳	۴۳۳	۴۳۳	۴۳۳	۴۳۳
درصد معتبر	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰



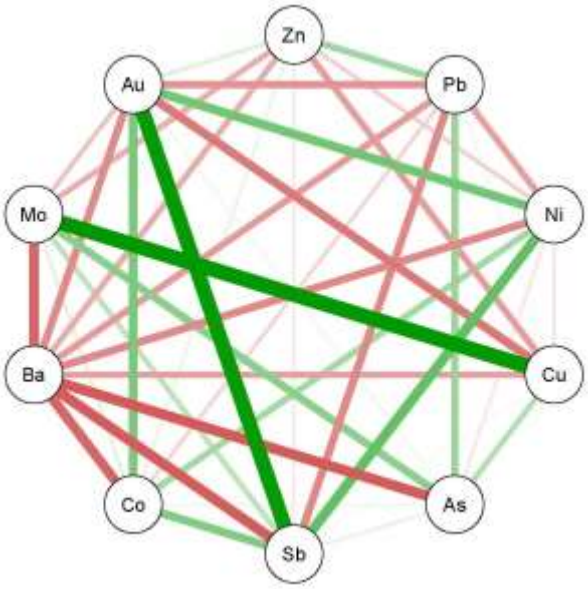
شکل ۱. نمودار جعبه‌ای عیارهای تک عنصری. برای برخی عناصر چولگی به سمت راست و نقاط پرت متعدد مشاهده می‌شود



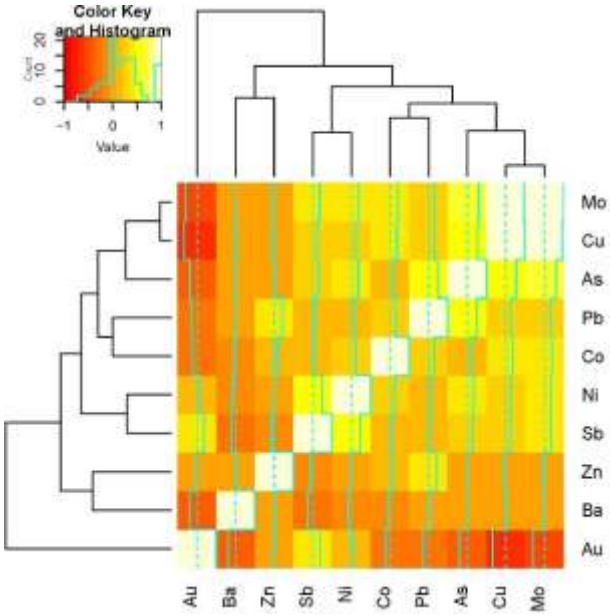
شکل ۲. نمودار جعبه‌ای عیارهای تک عنصری بعد از تبدیل clr. میزان چولگی عناصر کمتر شده و نقاط پرت همچنان حاضرند



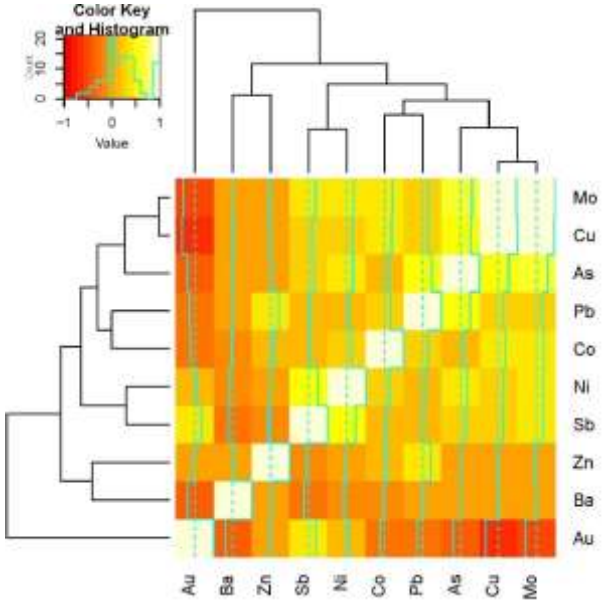
شکل ۳. گراف شبکه‌ای از ماتریس همبستگی عیارهای تک عنصری. همبستگی کاذب بین عناصر مشاهده می‌شود



شکل ۴. گراف شبکه‌ای از ماتریس همبستگی عیارهای تک عنصری بعد از تبدیل clr



شکل ۵. نقشه گرمایی ماتریس همبستگی اسپیرمن عیار عنصر و دندروگرام حاصل از خوشه‌بندی سلسه مراتبی وارد



شکل ۶ نقشه گرمایی ماتریس همبستگی اسپیرمن عیار عنصر بعد از تبدیل clr و دندروگرام حاصل از خوشه‌بندی سلسه مراتبی

نتیجه‌گیری

تحلیل آماری داده‌های ترکیبی به دلیل مشکلات پیچیده این داده‌ها از قبیل پیروی نکردن آنها از توزیع نرمال، چولگی شدید، وجود نقاط پرت متعدد، مقادیر گمشده، کار پیچیده‌ای است. به همه این موارد خاصیت بسته بودن این داده‌ها نیز اضافه می‌شود. سوال اینجاست که به‌هنگام آمده‌سازی داده‌ها برای تحلیل آماری، اولویت با برطرف کردن کدام یک از مشکلات بالاست. موضوعی که در برخی متون علمی مشاهده می‌شود، نادیده‌گرفتن خاصیت بسته بودن و دادن اولویت به مباحثی همچون کاهش چولگی یا نرمال کردن داده‌هاست.

این کار اغلب با اعمال تبدیل لگاریتمی روی داده‌ها انجام می‌شود. از طرفی خاصیت بسته بودن داده‌های ترکیبی یک نتیجه مستقیم به نام همبستگی کاذب دارد. از آنجا که بررسی روابط دومتغیره نخستین گام در تحلیل داده‌هاست و از آنجایی که ماتریس همبستگی نقش مهمی در بسیاری از روش‌های آماری دارد، نادیده گرفتن همبستگی کاذب می‌تواند منجر به نتایج گمراه کننده آماری شود. در این مقاله سه نوع تبدیل برای باز کردن سیستم بسته داده‌های ترکیبی مرور شد. تاکید بر این است که اولویت اول آماده‌سازی داده‌ها اعمال یکی از این تبدیل‌ها برای باز کردن داده‌هاست. در اینجا دو مشکل خودنمایی می‌کند. اول اینکه هر کدام از این تبدیل‌ها در یک بستر تاریخی معرفی شده‌اند و طبیعی است هر کدام با وجود مفید بودن نسبت به هم مزایا و عیوبی نیز دارند. از طرف دیگر بعضاً به‌دلیل پیچیدگی نسبی فرمول آنها ممکن است یک پژوهشگر مایل به استفاده از آنها نباشد. خصوصاً اینکه ممکن است بر اساس تجربه پژوهشگر، نتایج حاصل از اعمال تبدیل لگاریتمی به عنوان یک روش متداول تاکنون قابل قبول بوده است. هدف این مقاله پرداختن به موضوع همبستگی کاذب و ضرورت رفع آن با باز کردن سیستم بسته داده‌ها به عنوان گام اول علیرغم مشکلات بالاست. لازم است پژوهشگر در زمینه پیشرفت‌های موجود بروز باشد. در زمینه مشکلات محاسباتی، خوشبختانه امروزه نرم‌افزارهای آماری نظیر R روش‌های آماری مختص تحلیل داده‌های ترکیبی را به‌طور قابل توجه‌ای توسعه داده‌اند. در مجموع امید است، با مجهز شدن به نظریه داده‌های ترکیبی، استفاده از روش‌های آماری مناسب و نرم‌افزارها بتوانیم داده‌ها را از نظر آماری دقیق‌تر تحلیل کنیم.

منابع

- اعلمی نیا، ز.، منصوری اصفهانی، م.، طباطبایی، س. ح. و بختیاری، ن. م.، ۱۳۹۷. شناسایی و پی‌جویی ناهنجاری‌های زمین‌شناسی همراه با کانی‌سازی مس در چهارگوش ۱:۱۰۰۰۰۰ نطنز (شمال اصفهان)، ایران. بلورشناسی و کانی‌شناسی ایران، (۳)۲۶، ۶۲۵-۶۳۴.

- حسین پور نجاتی، س.، سیاه چشم، ک.، علوی، س. غ.، زرگری، پ.، ۱۴۰۰. تحلیل پتانسیل کانیزایی با استفاده از روش تحلیل فاکتوری مرحلهای (SFA) در گستره خوشنامه، هشجین، استان اردبیل. فصلنامه زمین‌شناسی ایران، ۵۷، ۱۳-۱.

-حیدریان دهکردی، ن.، توکل، م. ح.، پورمحمدی، س.، ۱۳۹۶. پتانسیل سنجی رسوبات آبراهه‌ای منجیل با استفاده از GIS . فصلنامه زمین‌شناسی ایران، ۴۳، ۱۰۸-۹۵.

-محمدی اصل، ز.، سعیدی، ع.، آرین، م.، سلگی ع.، فرهادی نژاد، ط.، ۱۳۹۹. جداسازی آنومالی‌های ژئوشیمیایی از زمینه با استفاده از روش فرکتالی عیار-تعداد در محدوده وشنوه (جنوب قم). فصلنامه زمین‌شناسی ایران، ۵۳، ۷۳-۶۱.

- Aitchison, J., 1986. The Statistical Analysis of Compositional Data, Chapman and Hall/CRC, New York.

- Chayes, F., 1960. On correlation between variables of constant sum. Journal of Geophysical Research, 65(12), 4185–4193.

- Egozcue, J.J., Pawlowsky-Glahn, V. 2005. Groups of parts and their balances in compositional data analysis. Mathematical Geology, 37, 795–828.

- Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology, 35, 279-300.

- Filzmoser, P., Hron. K., 2008. Outlier detection for compositional data using robust methods. Mathematical Geosciences, 40, 233-248.
- Filzmoser, P., Hron, K, 2009. Correlation analysis for compositional data. Mathematical Geosciences, 41(9), 905-919.
- Filzmoser, P., Hron, K., Reimann, C., 2009. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Science of the Total Environmen, 407, 6100–6108.
- Filzmoser, P., Horn, K., Templ, M., 2018. Applied Compositional Data Analysis with Worked Examples in R. Springer, Switzerland.
- Gerald van den Boogaart, K., Tolosana-Delgado, R., 2013. Analyzing Compositional Data with R. Springer, New York.
- Miesch, A.T., Chapman, R. P. 1977. Log-transformation in geochemistry. Mathematical Geology, 9(2), 191-194.
- Pearson, K., 1897. Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc. R. Soc. Lond., 60, 489-498.
- Pendleton, B. F., Newman, I., Marshall, R. S., 1983. A Monte Carlo approach to correlation spuriousness and ratio variables. Statist Comput Simul, 18, 93-124.
- Reimann, C., Filzmoser, P., 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environmental Geology, 39, 1001–1014.
- Reimann, C., and Filzmoser, P., Garrett, R., Dutter, R., 2008. Statistical Data Analysis Explained - Applied Environmental Statistics with R. John Wiley and Sons, London.
- Reimann, C., Filzmoserand, P., Hron, K., Kynčlová P., and Garrett, R., 2017. A new method for correlation analysis of compositional (environmental) data – a worked example. Science of the Total Environment, 607–608, 965–971.

Spurious Correlation and the Closure Property of Compositional Data in Geological Sciences

Abstract

In earth sciences, measurements usually produce compositional data with a property called closedness. Researchers who use common statistical methods on compositional data ignore spurious correlations, which causes incorrect results. This article introduces transformations for opening closed system of compositional data. These transformations include the additive logarithmic ratio (alr), the centred logarithmic ratio (clr), and the isometric logarithmic ratio (ilr). They are all defined in terms of logarithms of ratios. We then applied the clr transformation to a soil chemical data set. We also analysed the results of applying cluster analysis on the clr transformed data using Spearman's correlation coefficient matrix as distance. We also investigated how applying clr transformation affects spurious correlation, skewness and outliers in the data using R statistical software.
Key words: Closure Property, Compositional data, Log-ratio transformations, Robust statistical method, Spurious correlation.