

# همبستگی کاذب و خاصیت بسته بودن داده‌های ترکیبی در علوم زمین

حمید قربانی<sup>(۱)</sup>

۱. استادیار گروه آمار، دانشکده علوم ریاضی، دانشگاه کاشان

تاریخ دریافت: ۱۴۰۳/۰۲/۲۶

تاریخ پذیرش: ۱۴۰۳/۰۵/۲۷

## چکیده

داده‌های ترکیبی که همواره 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 نرمال کردن داده‌ها استفاده می‌کنند در واقع وجود همبستگی کاذب در داده‌های ترکیبی را نادیده می‌گیرند که این باعث نتایج آماری نادرست می‌شود. در این پژوهش 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 به‌عنوان 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 لگاریتمی نسبتی، روش‌های آماری استوار، همبستگی کاذب.

## مقدمه

به‌جای اطلاعات مطلق، اطلاعات نسبی را در اختیار داریم. به‌عنوان 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  $10^6$  mg/kg می‌شود. واضح است که یک مجموعه داده، تنها زمانی می‌تواند ترکیبی باشد

در علم آمار، داده‌های ترکیبی<sup>۱</sup> 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 برای تحلیل اطلاعات

1. Compositional data

\* نویسنده مرتبط: hamidghorbani@kashanu.ac.ir

زمین‌شناسان متوجه همبستگی کاذب در داده‌های ترکیبی شدند (Chayes, 1960). هر چند بحث درباره این نوع همبستگی بارها در متون آماری مورد توجه قرار گرفته است (Pendleton et al., 1983) با وجود این در طول سال‌ها این خطرات، بیشتر فراموش و یا نادیده گرفته شده‌اند، تا اینکه اولین بار توسط ایتچیسون (Aitchison, 1986) برای تحلیل آماری داده‌های ترکیبی نظریه‌پردازی شد.

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 پژوهش 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 پژوهش روی این نکته تاکید شده است که داده‌های زمین‌شناسی 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 نشده و باعث نتایج آماری غیرقابل قبول می‌گردد. در این پژوهش، تبدیل‌های نسبت لگاریتمی معرفی می‌شوند که اعمال آنها 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) اندازه‌گیری شده‌اند، ارائه می‌دهد همبستگی کاذب، طبق تعریف یک همبستگی معنی‌دار آماری بین دو متغیر است که به دلیل نگرش تحلیل‌گر به متغیرها و نحوه مدیریت آنها و نه

که حداقل متشکل از دو جزء باشد. هنگام کار با داده‌های ترکیبی حتی در موردی که فقط یک جزء را گزارش می‌کنیم، در واقع به‌طور 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 بدون اینکه بدانیم کل بودجه چقدر بوده یا چقدر بودجه برای دیگر وزارتخانه‌ها در نظر گرفته شده، به‌خودی‌خود بی‌معنی است. یک بردار از مشاهدات ترکیبی،  $x$ ، که نمونه‌ای شامل  $n$  جزء است، به‌صورت زیر بیان می‌شود:

$$x = (x_1, x_2, \dots, x_n), x_i > 0, \sum_{i=1}^n x_i = \kappa, \quad (\text{رابطه ۱})$$

با این تعریف، از نظر آماری دو بردار  $ax$  و  $x$  به ازای هر ثابت مخالف صفر  $x$  دارای اطلاعات یکسانی هستند. به‌عبارت دیگر اطلاعات مورد نیاز نه در خود آنها، بلکه در نسبت بین اجزاء قرار دارد. خاصیت مجموع ثابت که در ادبیات داده‌های ترکیبی به خاصیت بسته‌بودن داده‌ها 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 به‌عنوان 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<sup>۲</sup> حتی از بیش از یک قرن پیش به‌خوبی شناخته شده‌اند. پیرسون (Pearson, 1987) 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 دو عنصر اندازه‌گیری شده). وی همبستگی حاصل بین چنین متغیرها یا به‌عبارت بهتر نسبت‌ها را همبستگی کاذب نامید. همچنین در این پژوهش کلاسیک، پیرسون به خطراتی که متوجه تحلیل‌گری که سعی در تفسیر همبستگی بین نسبت‌هایی که صورت‌ها یا مخرج‌های آن دارای اجزای مشترک هستند اشاره کرده است. زمان زیادی طول کشید تا

1. Closed data

2. Spurious correlation

زیادی اولین حمله قلبی خود را در سنین پایین تری تجربه می‌کنند. همچنین افراد بزرگسال بیشتر از کودکان مطالعه می‌کنند، بنابراین سن بالاتر معادل اندازه کفش بالاتر و ساعت مطالعه بیشتر است. همواره در هوای گرم از یک طرف میزان مصرف بستنی و از طرف دیگر تعداد غرق شدگی افزایش می‌یابد. پس در مثال‌های بالا نقش سن و دمای هوا به عنوان متغیر میانجی نادیده گرفته شده است. باید به این نکته توجه شود که، 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 سروکار داریم، با استفاده از تبدیل مناسبی که در ادامه به آنها اشاره خواهیم کرد، ابتدا سیستم به اصطلاح بسته داده‌ها را باز، سپس تحلیل‌های آماری مورد نظر را روی داده‌های باز شده پیاده می‌شود.

### پیشینه تحقیق

از حدود شش دهه گذشته، پژوهشگران مشکلات مربوط به تحلیل آماری سیستم‌های 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 آماده‌سازی داده‌ها قبل از مدل‌بندی مبتنی بر ماتریس همبستگی که اساس بررسی ارتباط بین متغیرها است، نیاز به یک تبدیل احساس می‌شود. فیلتز موستر و همکاران در پژوهش (Filzmoser and Hron, 2009) توضیح داده‌اند که چرا در مرحله آماده‌سازی برای محاسبه ماتریس همبستگی، نیاز است که با یک تبدیل، داده‌های ترکیبی به یک فضای نمونه مناسب منتقل شود. برای فهم ضرورت این کار، توجه شود که در رابطه (۱) مقدار ثابت  $\kappa$  که باعث بسته بودن سیستم داده‌های ترکیبی است به‌خودی‌خود اهمیت ندارد (چون مقدار این ثابت با تغییر واحد اندازه‌گیری مشاهدات تغییر می‌یابد). آنچه تعیین‌کننده

وجود هرگونه رابطه ذاتی بین متغیرها به‌دست‌آمده است. در واقع این همبستگی معنادار 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 به هم مرتبط هستند. بسیاری از رویدادها وجود دارند که ارتباط آنها ضریب همبستگی بزرگی را ایجاد می‌کند، اما این نتیجه‌گیری که 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؟
- پژوهشگران 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 مرگ‌ومیر در اثر غرق‌شدگی باید مصرف بستنی را کاهش داد؟

چرا همبستگی‌های 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 نبود جمع‌آوری مناسب داده‌ها است. اغلب، دو متغیر به‌دلیل متغیر سوم که بر هر دو تأثیر می‌گذارد، با هم مرتبط هستند. به‌عبارت‌دیگر عامل یا متغیر دیگری که این دو متغیر را به هم مرتبط می‌کند، هنگام بررسی و اندازه‌گیری این دو متغیر نادیده گرفته شده است. به این متغیر که علت مشترک تغییرات همسو در دو متغیر است، متغیر میانجی<sup>۱</sup> 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 با شانس

1. Mediator variable

آبراهه‌ای بوده است. گستره مورد مطالعه در مختصات طول جغرافیایی  $30^{\circ} 51'$  تا  $52^{\circ}$  شرقی، عرض جغرافیایی  $30^{\circ} 33'$  تا  $34^{\circ}$  شمالی و  $120$  کیلومتری شمال شهر اصفهان در منطقه نطنز واقع شده و دارای مساحت  $2500$  کیلومترمربع بوده است. هدف ما در ادامه بررسی آماری این داده‌ها از منظر آمار توصیفی در راستای مطالب ذکر شده در بخش‌های قبلی بوده و بنابراین در این چارچوب موقعیت مکانی داده‌ها وارد تحلیل نشده است. با وجود این اطلاعات بیشتر در مورد زمین‌شناسی این داده‌ها در (اعلمی‌نیا و همکاران، ۱۳۹۷) آمده است. داده‌های ژئوشیمیایی متعارف به چهار دسته اصلی تقسیم می‌شوند که شامل عناصر اصلی، عناصر کمیاب، ایزوتوپ‌های پرتوزا و ایزوتوپ‌های پایدار هستند. به دلیل ماهیت پیچیده داده‌های ژئوشیمیایی یا به‌طور کلی داده‌های زمین‌شناسی در کاربست روش‌های معمول آماری برای تحلیل این داده‌ها محدودیت وجود دارد. دلیل این محدودیت علاوه بر ماهیت ترکیبی بودن این نوع داده‌ها که موضوع اصلی این پژوهش 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 به ذکر است حتی اگر از تبدیل لگاریتمی استفاده شود، این تبدیل سیستم بسته داده‌ها را باز نمی‌کند و همبستگی‌های به‌دست‌آمده بعد از اعمال تبدیل‌هایی غیر آنچه در ادامه به آن اشاره خواهد شد، می‌توانند به‌طور جدی گمراه‌کننده باشند، (Reimann and Filzmoser, 2000). در متون آماری مربوط

است، مقیاس اندازه‌گیری می‌باشد، موضوعی که هنگام استفاده از ضریب همبستگی مشکل‌ساز خواهد شد. توجه شود که چون  $\sum_{i=1}^n x_i = \kappa$ ، به ازای هر  $i$  داریم:

$$cov\left(\sum_{i=1}^n x_i, x_i\right) = cov(\kappa, x_i) = 0. \quad (\text{رابطه ۲})$$

به عبارت دیگر جمع هر سطر یا ستون ماتریس واریانس-کواریانس داده‌های ترکیبی برابر صفر می‌شود. به‌عنوان مثال برای سطر اول این ماتریس داریم:

$$cov(x_1, x_2) + cov(x_1, x_3) + \dots + cov(x_1, x_n) = -var(x_1). \quad (\text{رابطه ۳})$$

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 به دلیل خاصیت بسته بودن داده‌ها، منفی شود. این موضوع در تحلیل داده‌های ترکیبی به مسئله اریبی منفی<sup>۱</sup> معروف است که نخستین بار توسط پیرسون (Pearson, 1897) و بار دیگر توسط چایز بیان شد (Chayes, 1960).

به دلیل نقش اساسی ماتریس واریانس-کواریانس (یا همبستگی) که در روش‌های آماری چند متغیره، می‌توان انتظار داشت نادیده گرفتن خاصیت بسته بودن داده‌های ترکیبی، ممکن است تلاش بیهوده‌ای برای تفسیر آماری نتایج نادرست به دلیل همبستگی‌های بی‌دلیل، 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.

### روش مطالعه

#### 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  $\text{Co}$ ،  $\text{Ba}$ ،  $\text{Mo}$ ،  $\text{Pb}$ ،  $\text{Zn}$ ،  $\text{Cu}$ ،  $\text{As}$ ،  $\text{Sb}$  و  $\text{Au}$  و  $\text{Ni}$  حاصل از ۴۳۳ نمونه از رسوبات

1. Negative bias problem

داریم:

$$\sum \text{clr}(x) = \sum (\log(x_i) - \log(G(x))) = 0. \quad (\text{رابطه ۶})$$

مزیت اصلی این تبدیل این است که ارتباط متغیرهای جدید و قبلی حفظ و تفسیر نتایج به نسبت ساده است. تبدیل clr ایزومتريک است ولی در عين 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 ایزومتريک 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 به هم مربوط می‌شوند. هر چند برای این منظور در Egozcue and Pawlowsky (2009)؛ Filzmoser and Hron, 2009؛ Glahn, 2005 راه‌حلی مبتنی بر بالانس‌ها<sup>۱</sup> مطرح شده است. همچنین رایمن و همکاران نوع خاصی از تبدیل ilr بر اساس مختصات متقارن را برای تحلیل همبستگی داده‌های ترکیبی ارائه داده‌اند (Reimann et al., 2017).

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 پژوهش فیلترز ماستر و همکاران، نقل به مضمون می‌شود. در پژوهش‌های مرتبط، تبدیل‌های ذکر شده برای باز کردن داده‌های بسته به ندرت اعمال می‌شوند. این تبدیل‌ها پیچیده‌تر از تبدیل لگاریتمی هستند و به واسطه اعمال آنها ارتباط داده‌های تبدیل شده و داده‌های ابتدائی قطع می‌شود. ممکن است از نظر پژوهشگر، نتایج به دست آمده از اعمال روش‌های آماری متداول روی داده‌های بسته، منطقی به نظر برسد و در نتیجه پیامدهای احتمالی کار با داده‌های بسته به ندرت مورد سؤال قرار بگیرد. در این پژوهش نشان داده می‌شود که مشکل

به تحلیل داده‌های ترکیبی، برای باز کردن سیستم‌های بسته اعداد، سه تبدیل از نوع نسبت لگاریتمی<sup>۱</sup> یعنی تبدیل‌هایی بر حسب لگاریتم نسبت‌ها پیشنهاد شده است، که عبارت‌اند از تبدیل نسبت لگاریتمی جمعی<sup>۲</sup>، تبدیل نسبت لگاریتمی مرکزی شده<sup>۳</sup> و تبدیل نسبت لگاریتمی ایزومتريک<sup>۴</sup> که آنها را به ترتیب با تبدیل‌های alr، clr و ilr نمایش می‌دهیم. تبدیل alr و clr توسط ایتچیسون (Aitchison, 1986) و تبدیل ilr توسط اگزو (Egozcue et al., 2003) تعریف شده‌اند. تبدیل alr به این صورت تعریف می‌شود که یکی از متغیرها به دلخواه به عنوان 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 ایزومتريک ندارد، یعنی فاصله بین نقاط در فضای اولیه و فضای تبدیل شده یکسان نیستند. توجه شود 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<sup>۵</sup> 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 پژوهش (Filzmoser and Hron, 2008) مراجعه کند. نکته اینجاست که هنگام کار با داده، هرگاه تمایل به اعمال یکی از تبدیل‌ها وجود داشته باشد، می‌توان از امکانات موجود در نرم‌افزار R استفاده کرد (Gerald van den Boogaart and Tolosana-Delgado, 2013).

تبدیل clr روی بردار مشاهدات  $x = (x_1, x_2, \dots, x_n)$  به صورت زیر تعریف می‌شود:

$$\text{clr}(x) = \left\{ \log\left(\frac{x_1}{G(x)}\right), \dots, \log\left(\frac{x_n}{G(x)}\right) \right\} \quad (\text{رابطه ۴})$$

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

$$\log(G(x)) = \log\left(\exp\left[\frac{1}{n} \sum_{i=1}^n \log(x_i)\right]\right) = \frac{1}{n} \sum_{i=1}^n \log(x_i), \quad (\text{رابطه ۵})$$

1. Logratio transformation
2. Additive logratio transformation
3. Centered logratio transformation
4. Isometric logratio transformation
5. Pivot coordinate
6. Balances

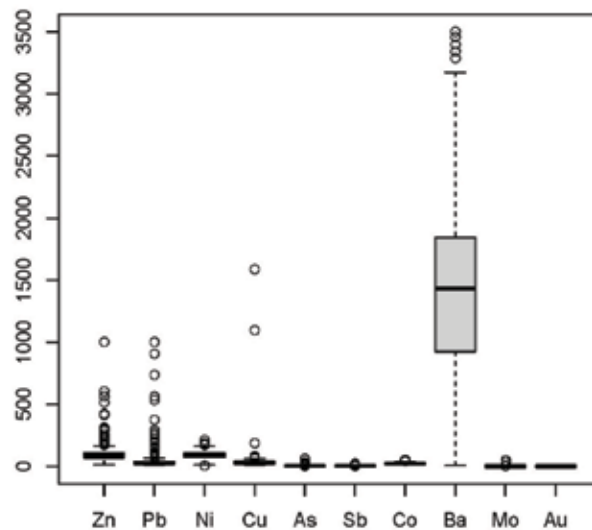
همبستگی اسپیرمن به‌عنوان ماتریس فاصله، انجام و دندروگرام نتایج به همراه نقشه گرمایی ماتریس فاصله ارائه شده است. هدف، مقایسه بررسی تاثیر اعمال تبدیل و انتخاب ماتریس فاصله مناسب بر نتایج حاصل از خوشه‌بندی و نشان دادن این موضوع بوده است که با اعمال تبدیل، نتایج تا چه میزان نسبت به وقتی تبدیل اعمال نشده، متفاوت بوده‌اند.

### بحث

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

بسته‌بودن داده‌ها باید حتی قبل از معیارهای آماری ساده مانند میانگین یا انحراف معیار یا رسم نمودارهای توزیع داده‌ها مانند بافت‌نگار 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 (Filzmoser et al., 2009).

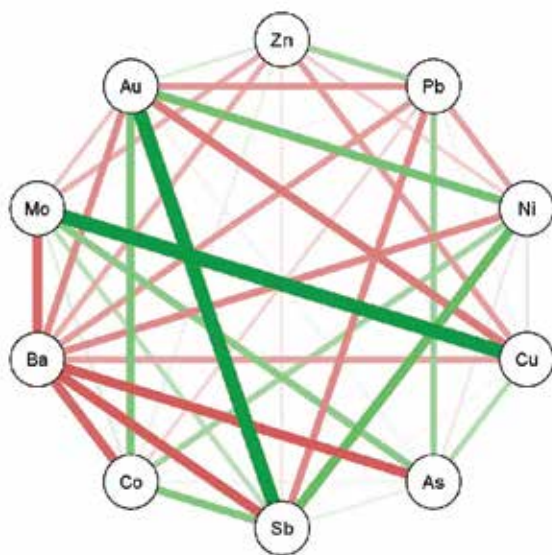
در بخش بعد، وجود همبستگی کاذب، چولگی شدید داده‌ها و وجود داده‌های پرت با استفاده از نمودارهای آماری بررسی شده است. همچنین نشان داده می‌شود که با اعمال تبدیل clr روی داده‌ها، علاوه بر حذف همبستگی‌های کاذب، دو مشکل دیگر یعنی چولگی شدید داده‌ها و نقاط پرت تا حدی اصلاح شده‌اند. همچنین یک تحلیل خوشه‌ای روی داده‌های اصلی و تبدیل شده با استفاده از ماتریس



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

قبل تبدیل مثبت و ضعیف و بعد از آن شدت همبستگی همچنان مثبت ولی شدت آن افزایش پیدا کرده است. توجه شود که نقاط پرت روی ساختار ضریب همبستگی بین دو متغیر تأثیر مخرب دارد. بنابراین در شکل ۴ علیرغم انتظار وجود همبستگی مثبت بین برخی متغیرها، ممکن است برخی همبستگی‌ها ضعیف یا حتی در مواردی همبستگی منفی باشد. همان‌طور که از شکل ۲ مشاهده می‌شود دلیل

در گراف داده‌شده رنگ سبز و قرمز به‌ترتیب نشان‌دهنده 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 می‌رفت به دلیل وجود همبستگی کاذب، ساختار همبستگی دو متغیره بین عیارها قبل و بعد از تبدیل بسیار متفاوت است. به‌عنوان مثال همبستگی دو متغیر Au-Sb

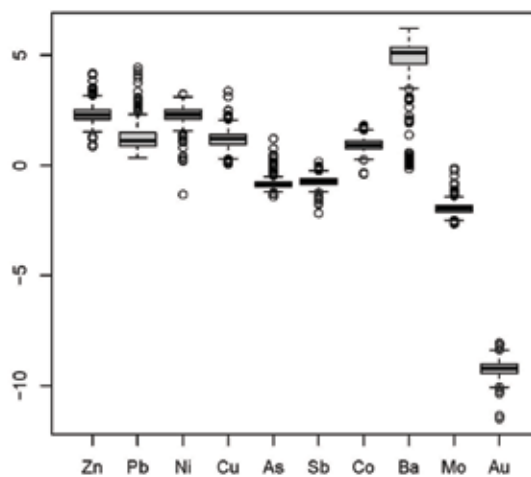


شکل ۴. گراف شبکه‌ای از ماتریس همبستگی عیارهای تک عنصری بعد از تبدیل clr

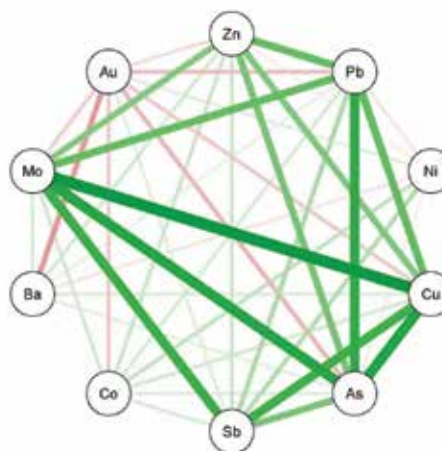
در چنین مواردی محاسبه ضریب همبستگی مقاوم به نقطه پرت می‌تواند اثر نقاط پرت را تعدیل کند. توجه شود که معرفی و استفاده از روش‌های آماری استوار<sup>۱</sup> موضوع این پژوهش نبوده است. با این وجود چون ضریب همبستگی رتبه‌ای اسپیرمن تحت تأثیر نقاط پرت نیست، در شکل‌های ۵ و ۶، نقشه گرمایی<sup>۲</sup> ضریب همبستگی اسپیرمن داده‌ها به همراه دندروگرام<sup>۳</sup> حاصل از خوشه‌بندی سلسه‌مراتبی نمایش داده شده است. توجه شود که 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؛ به‌این ترتیب که با افزایش شدت همبستگی (قدر مطلق) رنگ‌ها اشباع‌تر می‌شوند. هر سلول در ماتریس همبستگی با یک مربع در نقشه گرمایی نشان داده می‌شود و نقشه گرمایی همانند ماتریس همبستگی حول قطر اصلی اش متقارن است. توجه شود که به‌دلیل اینکه نقشه گرمایی خواناتر شود، ترتیب سطر و ستون‌های این نقشه متفاوت از ترتیب سطر و ستون‌های ماتریس داده‌هاست. به‌خصوص

1. Robust statistical methods
2. Heat map
3. Dendrogram

این موضوع به وجود داده‌های پرت است. 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 نرم‌افزارهای 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.



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



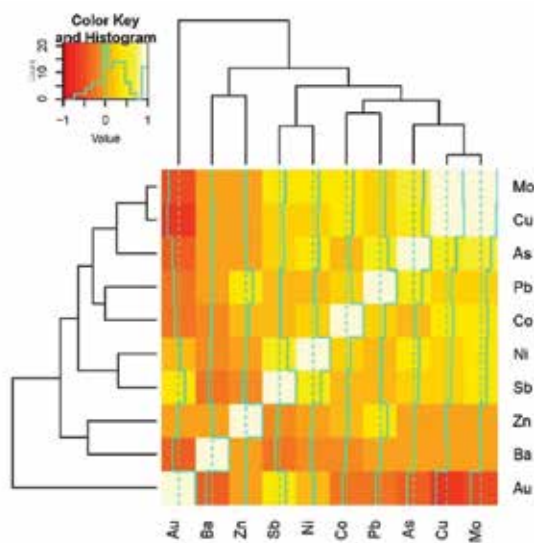
شکل ۳. گراف شبکه‌ای از ماتریس همبستگی عیارهای تک عنصری. همبستگی کاذب بین عناصر مشاهده می‌شود

که فقط نسبت به تبدیل 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). بنابراین وقتی رو داده‌ها تبدیل clr استفاده کرده‌ایم، این تبدیل روی ضریب همبستگی اسپیرمن تأثیر چندانی نداشته است.

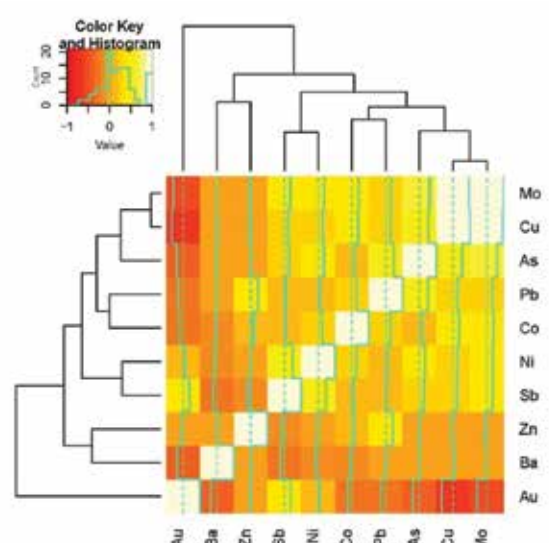
متغیرها با ضریب همبستگی نزدیک به هم (رنگ‌های مشابه)، طوری جابجا می‌شوند که بلوک‌هایی از رنگ‌های مشابه در نقشه گرمایی ایجاد شود تا ساختار همبستگی منفی درون داده‌ها، بهتر دریافت شود. ملاحظه می‌شود نقشه گرمایی شکل ۵ و ۶ کم‌وبیش از نظر رنگ‌بندی، شبیه به نظر می‌رسد (کامل یکسان نیستند). این موضوع می‌تواند به این دلیل باشد که داده‌های پرت روی ضریب همبستگی اسپیرمن تأثیر ندارند و همچنین برخلاف ضریب همبستگی پیرسون

جدول ۱. آماره‌های توصیفی برای عیار عناصر برحسب ppm

روى	آنتیموان	سرب	نیکل	مولیبدن	مس	کبالت	باریم	طلا	آرسنیک	
۱۰۱.۷۴	۴.۱۹	۴۸.۶۷	۹۰.۲۳	۱.۴۶	۳۶.۹۲	۲۲.۸۱	۱۴۵۶.۴۶	۰	۴.۲۳	میانگین
۸۵.۲۴	۰.۹۶	۱۰۸.۰۱	۳۰.۰۸	۲.۶۸	۹۱.۶	۶.۲۷	۸۴۵.۶۱	۰	۴.۲۹	انحراف معیار
۱۳	۱	۱۰	۴.۹۸	۰.۵۲	۸.۱۹	۱۱	۶.۸۱	۰	۱.۸۳	کمینه
۶۸	۳.۸۲	۲۰	۶۸.۹۲	۱.۰۴	۲۱.۹۷	۱۸	۹۲۳.۳۸	۰	۳.۲۸	چارک اول
۸۶	۴.۱۴	۲۵	۱۸۸۷.	۱.۲۳	۲۹.۸۶	۲۱	۱۴۳۱.۹۳	۰	۳.۵۷	میانه
۱۰۸	۴.۴۸	۳۸	۱۰۷.۹۵	۱.۴۷	۳۷.۸۵	۲۶	۱۸۴۳.۹	۰	۴.۰۱	چارک سوم
۱۰۰	۱۹.۴۳	۱۰۰	۲۱۰.۱۲	۴۷.۸۹	۱۵۸۵.۳۱	۵۲	۳۵۰۰	۰	۶۰.۲	بیشینه
۲۸.۱۷	۰.۴۸	۱۰.۳۸	۲۹.۰۹	۰.۳۳	۱۱.۷۹	۵.۹۳	۶۸۲.۸۸	۰	۰.۵	بزرگی انحراف از میانگین
۴۰	۰.۶۶	۱۸	۳۹.۰۳	۰.۴۳	۱۵.۸۸	۸	۹۲۰.۵۲	۰	۰.۷۳	دامنه میان چارکی
۰.۸۴	۰.۲۳	۲.۲۲	۰.۳۳	۱.۸۳	۲.۴۸	۰.۲۷	۰.۵۸	۰.۱۷	۱.۰۱	ضریب تغییرات
۶.۸۹	۹.۳۱	۷.۰۸	۰.۳۹	۱۴.۹۶	۱۴.۷۱	۱.۳۲	۰.۴۸	۰.۱۵	۱۰.۳۴	چولگی
۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	انحراف معیار چولگی
۶۱.۲۸	۳۲۱۴۲.	۵۴.۰۸	۰.۵۵	۲۳۶.۸۴	۲۲۶.۵۷	۲.۷	۰.۲۲	۰.۵۷	۱۲۰.۹۵	کشیدگی
۴۳۳	۴۳۳	۴۳۳	۴۳۳	۴۳۳	۴۳۳	۴۳۳	۴۳۳	۴۳۳	۴۳۳	تعداد داده معتبر
۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	درصد معتبر



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



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



## نتیجه‌گیری

است، با مجهز شدن به نظریه داده‌های ترکیبی، استفاده از روش‌های آماری مناسب و نرم‌افزارها بتوان داده‌ها را از نظر آماری دقیق‌تر تحلیل کرد.

### 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. و زرگری، پ.، ۱۴۰۰. تحلیل پتانسیل کانیزایی با استفاده از روش تحلیل فاکتوری مرحله‌ای (SFA) در گستره خوشنامه، هسجین، استان اردبیل. فصلنامه زمین‌شناسی ایران، ۵۷، ۱۳-۱.
- حیدریان دهکردی، ن.، توکل، م. ح. و 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. و 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. and 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. and Barceló-Vidal, C., 2003. Isometric logratio transformations for composi-

تحلیل آماری داده‌های ترکیبی به دلیل مشکلات پیچیده این داده‌ها از قبیل پیرونی 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 آماده‌سازی داده‌ها برای تحلیل آماری، اولویت با برطرف کردن کدام‌یک 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 آنجاکه 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 گمراه‌کننده آماری شود. در این پژوهش 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 گاهی به‌دلیل پیچیدگی نسبی فرمول آنها ممکن است یک پژوهشگر مایل به استفاده از آنها نباشد. به‌ویژه اینکه ممکن است بر اساس تجربه پژوهشگر، نتایج حاصل از اعمال تبدیل لگاریتمی به‌عنوان یک روش متداول تاکنون قابل قبول بوده است. هدف این پژوهش پرداختن به موضوع همبستگی کاذب و ضرورت رفع آن با باز کردن سیستم بسته داده‌ها به‌عنوان 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

- tional data analysis. *Mathematical Geology*, 35, 279-300.
- Filzmoser, P. and Hron, K., 2008. Outlier detection for compositional data using robust methods. *Mathematical Geosciences*, 40, 233-248.
  - Filzmoser, P. and Hron, K., 2009. Correlation analysis for compositional data. *Mathematical Geosciences*, 41(9), 905-919.
  - Filzmoser, P., Hron, K. and Reimann, C., 2009. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. *Science of the Total Environment*, 407, 6100-6108.
  - Filzmoser, P., Horn, K. and Templ, M., 2018. *Applied Compositional Data Analysis with Worked Examples in R*. Springer, Switzerland.
  - Gerald van den Boogaart, K. and Tolosana-Delgado, R., 2013. *Analyzing Compositional Data with R*. Springer, New York.
  - Miesch, A.T. and 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. *Proceedings of the Royal Society of London*, 60, 489-498.
  - Pendleton, B. F., Newman, I. and Marshall, R. S., 1983. A Monte Carlo approach to correlation spuriousness and ratio variables. *Statist Comput Simul*, 18, 93-124.
  - Reimann, C. and Filzmoser, P., 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. *Environmental Geology*, 39, 1001-1014.
  - Reimann, C., Filzmoser, P., Garrett, R. and Dutter, R., 2008. *Statistical Data Analysis Explained-Applied Environmental Statistics with R*. John Wiley and Sons, London.
  - Reimann, C., Filzmoser, 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.