مقایسه روش‌های مناسب جانهی مقادیر سانسورشده در داده‌های ژئوشیمیایی

سید علی حسینی [1]، سمانه افتخاری مهابادی [2] ، امید اصغری [3*]

1- دانشجوی دکتری، آزمایشگاه شبیه‌سازی و پردازش داده، دانشکده مهندسی معدن، دانشگاه تهران

2- استادیار آمار، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران

2- استادیار، آزمایشگاه شبیه‌سازی و پردازش داده، دانشکده مهندسی معدن، دانشگاه تهران

(دریافت: اسفند 1393 پذیرش: مهر 1394)

چکیده

در این تحقیق به بررسی روش‌های جانهی مقادیر سانسور شده در مجموعه داده‌های چند متغیره ژئوشیمیایی پرداخته شده است. وجود مقادیر گم‌شده باعث محدودیت در استفاده از اغلب روش‌های آماری همچون تحلیل مولفه‌های اصلی می‌شود. حذف نمونه‌های شامل داده‌های گم‌شده باعث اریب شدن نتایج و از دست دادن اطلاعات می‌شود به همین دلیل در نظر گرفتن رویکردی مناسب در مواجهه با داده‌های گم‌شده یک نیاز اساسی در تحلیل مجموعه داده‌های ناکامل است. در این مقاله، با توجه به ماهیت ترکیبی داده‌های ژئوشیمیایی، چند روش مناسب برای جانهی مقادیر گم‌شده که در چند سال اخیر ارائه شده‌اند و به سادگی در محیط نرم‌افزار آماری *R* قابل اجرا هستند، معرفی شده‌اند. در نهایت با استفاده از یک مجموعه داده کامل مربوط به منطقه ظفرقند، این روش‌ها با یکدیگر مقایسه شده‌اند. نتایج نشان می‌دهند که استفاده از روش‌های چند متغیره برای جانهی و به طور خاص روش *ilr-EM* نسبت به دیگر روش‌ها ارجحیت دارند.

واژگان کلیدی

داده‌های ژئوشیمیایی، مقادیر سانسور شده، روش‌های جانهی، ماهیت ترکیبی، روش *ilr-EM*

ارجاع به این مقاله:

حسینی، ع.، افتخاری مهابادی، س.، اصغری، الف.، (1394)، مقایسه روش‌های مناسب جانهی مقادیر سانسور شده در داده‌های ژئوشیمیایی، روش‌های تحلیلی و عددی در مهندسی معدن، 5(9)، 63-72.

http://dx.medra.org/10.17383/S2251-6565(15)940916-X

* عهده دار مکاتبات: o.asghari@ut.ac.ir

1- مقدمه

داده‌های ژئوشیمیایی به صورت گسترده در مطالعات اکتشافی و بررسی‌های محیط زیستی مورد استفاده قرار می‌گیرد. در عمل پس از جمع‌آوری این‌گونه داده‌ها، از تحلیل‌های آماری تک متغیره، چند متغیره و روش‌های آمار فضایی به منظور توصیف و تحلیل داده‌های ژئوشیمیایی استفاده می‌شود. اغلب در مجموعه داده‌های ژئوشیمیایی وجود داده‌های گم‌شده[1] موضوعی اجتناب ناپذیر است یکی از دلایل رخداد گم‌شدگی در اینگونه داده‌ها مقادیر کوچک‌تر از حد تشخیص دستگاه اندازه‌گیری[2] (داده‌های سانسور شده) است.

وجود مقادیر گم‌شده در یک مجموعه داده باعث محدودیت‌هایی در محاسبات و استفاده از روش‌های آماری همچون تحلیل مولفه‌های اصلی می‌شود چرا که استفاده این روش‌ها نیازمند وجود یک مجموعه داده کامل است. یک راه حل مرسوم برای این مشکل، حذف نمونه‌های دارای گم‌شدگی از مجموعه داده‌ها است که به آن تحلیل داده‌های کامل[3](CCA) گفته می‌شود. استفاده از این روش باعث انحراف در نتایج و از دست دادن اطلاعات می‌شود. می‌توان به‌جای حذف نمونه‌های دارای گم‌شدگی، سلول‌های دارای گم‌شدگی در مجموعه داده‌ها را با یک مقدار مناسب جانهی[4] نمود[1].

از طرفی، داده‌های ژئوشیمیایی به صورت سهمی از یک مقدار ثابت به عنوان مثال درصد و یا ppm گزارش می‌شوند. این مقادیر، منحصر به فرد، مثبت و مجموع آنها ثابت است (در صورتی که تمام عناصر موجود در نمونه اندازه‌گیری شوند). بنابراین داده‌های ژئوشیمیایی ماهیت ترکیبی[5] دارند[2]. این ویژگی داده‌های ژئوشیمیایی موجب محدودیت در استفاده از روش‌های آماری چون آنالیز فاکتوری و آنالیز مولفه اصلی می‌شود. برای تحلیل داده‌های ترکیبی روش‌های مختلفی ارائه شده است که بر پایه هندسه آجیسون[6] هستند[3]. بنابراین انتظار می‌رود یک روش جانهی مناسب، ماهیت ترکیبی داده‌های ژئوشیمیایی را نیز در نظر بگیرد.

تقریباً در تمام مقالات و کتاب‌ها به این نکته اشاره شده است که، روش‌های جانهی داده‌های گم‌شده نباید ساختار کلی داده‌ها را دچار تغییرات شدید کند. چندین روش برای جانهی مقادیر گم‌شده در مجموعه داده‌ها ترکیبی ارائه شده است که می‌توان از آنها برای کامل کردن مجموعه داده‌های ژئوشیمیایی و تجزیه و تحلیل آنها بهره گرفت. در این مقاله ابتدا چندین روش مناسب برای استفاده در تجزیه و تحلیل داده‌های ژئوشیمیایی مرور می‌شود، از آن جمله می‌توان به روش معرفی شده توسط هرون و همکاران[4] که اخیراً در تجزیه و تحلیل داده‌های ژئوشیمیایی نیز به کار گرفته شده است، روش جایگزینی ضربی ساده[7] [5، 6] و جایگزینی لاگ نرمال ضربی [8] [7] و الگوریتم EM[9] و DA[10] در فضای تبدیل یافته لگاریتمی [8، 9] اشاره نمود. سپس در ادامه با استفاده از یک مجموعه داده کامل ژئوشیمیایی مربوط به منطقه ظفرقند، مجموعه داده‌های جدید که شامل مقادیر سانسور شده با نرخ‌های متفاوت هستند، ایجاد شده و توسط هر کدام از روش‌ها عملیات جانهی صورت گرفته و در نهایت دقت هر کدام از روش‌ها با بررسی میزان تغییر در آماره‌های مهم توزیع آنها و مقایسه مقدار واقعی و جانهی شده، مورد ارزیابی قرار خواهد گرفت.

2- گم‌شدگی در مجموعه داده‌های ژئوشیمیایی

مشکلی که اغلب متخصصان تجزیه و تحلیل داده‌های حاصل از نمونه‌گیری تجربی و اندازه‌گیری عناصر با آن روبرو هستند، این است که غلظت برخی از عناصر در نمونه به صورت "کمتر از[11]" و یا به ندرت "بیشتر از[12]" یک مقدار مشخص گزارش شده است. در مطالعات ژئوشیمیایی معمولاً مجموعه داده‌های در دسترس شامل سلول‌هایی است که به صورت کمتر از حد تشخیص دستگاه اندازه‌گیری[13](DL) گزارش شده است. چگونگی رسیدگی و برخورد با مقادیر گم‌شده برای تجزیه و تحلیل داده‌ها از نگرانی اصلی متخصصان به شمار می‌رود[10]. مقادیر کمتر از حد تشخیص دستگاه و یا سانسور در واقع مقادیر گم‌شده هستند، چرا که اطلاعات کاملی در مورد آنها وجود ندارد. با توجه به این‌که سانسور شدن و یا عدم ثبت مقدار واقعی عنصر، وابسته به غلظت خود عنصر در نمونه است به این‌گونه گم‌شدگی مکانیزم گم‌شدگی غیر تصادفی $NMAR$[14] اتلاق می‌گردد[1]. گم‌شدگی در این مکانیزم غیر قابل چشم پوشی است و تجزیه و تحلیل داده‌ها به روش CCA باعث اریب شدن نتایج می‌شود به ویژه اگر درصد گم‌شدگی بیشتر از 5٪ باشد.

3- روش‌های جانهی مقادیر سانسورشده

3-1- جایگزینی ضربی ساده[15] (*SMR*)

یکی از روش‌های تک متغیره‌ای که اغلب برای جانهی مقادیر کمتر از حد تشخیص دستگاه در مطالعات زیست محیطی استفاده می‌شود روش جایگزینی ضربی ساده است [11] که شامل جایگزین کردن 50٪ و یا 70٪ حد تشخیص دستگاه اندازه‌گیری به جای مقادیر گم‌شده است. پالارا و همکاران[7] با بهره گرفتن از این ایده بر روی داده‌های ترکیبی، نشان داده‌اند که یک ضریب تعدیل در جایگزینی مقدار سانسور شده برای حفظ مبانی مربوط به داده‌های ترکیبی مانند قید مجموع ثابت نیاز است. فرض کنید مجموعه‌ی داده‌ها (ماتریس $x_{n\times D}$) شامل n نمونه ($i=1,...,n$) و D عنصر اندازه‌گیری شده ($j=1,...,D$)، باشد. مقدار جانهی برای x_{ij} که اندازه‌گیری مربوط به عنصر j ام در نمونه iام را نشان می‌دهد و به طور کامل مشاهده نشده است از رابطه (1) به دست می‌آید:

$$x_{ij}=\begin{cases}\delta_{ij} & if \quad x_{ij}<DL_{ij}\\ \left(1-\dfrac{\sum_{k|x_{ik}<DL_{ij}}\delta_{ik}}{c_i}\right)x_{ij} & if \quad x_{ij}\geq DL_{ij}\end{cases} \qquad (1)$$

که $c_i=\sum_{j=1}^{D}x_{ij}$ برابر مجموع عناصر برای نمونه i ام (به عنوان مثال 1,000,000 اگر مقادیر به صورت ppm گزارش شده باشد) و δ_{ij} درصدی از DL_{ij} ($\delta_{ij}=\alpha DL_{ij}$) است. توجه داشته باشید δ و DL به این خاطر اندیس ij گرفته اند و این امکان را فراهم می‌کنند که غلظت عناصر در نمونه‌ها از روش‌های مختلفی با DL متفاوت، اندازه‌گیری شده باشند. فرناندز و همکاران[12] نشان داده‌اند که اگر δ_{ij} برابر با 65٪ حد تشخیص دستگاه در نظر گرفته شود، ساختار کوواریانس داده‌ها تغییر نخواهد یافت البته اگر میزان گم‌شدگی کمتر از 10٪ باشد.

3-2- جایگزینی لاگ نرمال ضربی[16] (*MLNR*)

این روش از ویژگی‌های آماری توزیع داده‌ها با فرض یک مدل احتمالی خاص استفاده می‌کند. معمولاً از توزیع لاگ نرمال برای مدل کردن توزیع داده‌های با چولگی به راست مانند غلظت شیمیایی استفاده می‌شود[11]. در این روش به جای در نظر گرفتن درصدی از حد تشخیص دستگاه برای تعیین δ_{ij} در رابطه (1) از تخمینی از میانگین هندسی مقادیر کمتر از حد تشخیص دستگاه با فرض توزیع لاگ نرمال برای متغیر jام ترکیب استفاده می‌شود. رابطه مقدار جایگزین بر اساس این روش به صورت زیر است.

$$\hat{x}_{ij}=\begin{cases}\delta_{ij}-\exp\left(\hat{\mu}_j-\hat{\sigma}_j\hat{\lambda}_{ij}\right) & if \quad x_{ij}<DL_{ij}\\ \left(1-\dfrac{\sum_{k|x_{ik}<DL_{ij}}\delta_{ik}}{c_i}\right)x_{ij} & if \quad x_{ij}\geq DL_{ij}\end{cases} \qquad (2)$$

که در آن:

$$\hat{\lambda}_{ij}=\frac{\phi\left(\left(\ln DL_{ij}-\hat{\mu}_j\right)\Big/\hat{\sigma}_j\right)}{\Phi\left(\left(\ln DL_{ij}-\hat{\mu}_j\right)\Big/\hat{\sigma}_j\right)}$$

در رابطه فوق $\phi(.)$ و $\Phi(.)$ به ترتیب تابع چگالی و تابع تجمعی توزیع نرمال استاندارد، $\hat{\mu}_j$ و $\hat{\sigma}_j$ پارامترهای تخمین خورده از الگوریتم مبتنی بر احتمال[17] برای داده‌های سانسور شده در jامین عنصر هستند.

3-3- k- نزدیک‌ترین همسایگی[18] (*KNN*)

روش جانهی *KNN* بعنوان یکی از روش‌های موفق برای داده‌های چند متغیره استاندارد شناخته شده است[13]. در این روش از یک اندازه فاصله برای پیدا کردن k مشاهده مشابه برای ترکیب‌های دارای گم‌شدگی استفاده می‌شود و مقادیر گم‌شده را توسط اطلاعات متغیرهای مشاهده شده در نزدیک‌ترین همسایگی آن مقدار جانهی می‌کند. با توجه به ماهیت ترکیبی داده‌ها در این روش از اندازه فاصله آجیسون استفاده می‌شود. فرض کنید $M_i\subset\{1,...,D\}$ مجموعه‌ای از اندیس‌ها برای نشان دادن سلول‌های دارای گم‌شدگی در سطر iام ماتریس مشاهدات X، $X_i=(x_{i1},...,x_{iD})$، باشد. بنابراین $O_i=\{1,...,D\}\setminus M$ اندیس مقادیر مشاهده شده را در X_i نشان می‌دهد. برای جانهی سلول گم‌شده X_{ij} ، برای هر $j\in M_i$ ، تمام اجزاء باقی‌مانده ترکیب که در عنصر jام و اندیس‌های O_i مشاهده شده‌اند در نظر گرفته می‌شود و k همسایه $x_{i1},...,x_{ik}$ برای نمونه X_i با استفاده از فاصله آجیسون محاسبه می‌شود. ابتدا این مقادیر همسایه انتخاب شده توسط فاکتور زیر تعدیل می‌شوند.

$$f_{ii_l} = \frac{\underset{o\in O_i}{median}\ x_{io}}{\underset{o\in O_i}{median}\ x_{i_l o}} \quad for\ \ l = 1,..,k \tag{3}$$

با استفاده از این فاکتور تعدیل به عنوان وزن برای مشاهدات، k همسایه قابل مقایسه می‌شوند. در نهایت مقدار جانهی برای سلول X_{ij} برابر است با:

$$x_{ij}^{*} = median\left\{ f_{ii_1} x_{i_1 j},..., f_{ii_k} x_{i_k j} \right\} \tag{4}$$

با توجه به این که در ضریب تعدیل از میانه استفاده شده است این روش در صورت وجود مقادیر دور افتاده[19] مقاوم[20] است.

3-4- الگوریتم *EM* نسبت لگاریتمی جمعی (*alr-EM*[21])

در این روش ابتدا از تبدیل نسبت لگاریتمی جمعی[22] به عنوان بخشی از الگوریتم استفاده می‌شود. این تبدیل برای ترکیب D- جزئی ام $X_i = [x_{i1},..., x_{iD}]$ با در نظر گرفتن یک جزء k ام به عنوان مخرج به صورت زیر تعریف می‌شود:

$$y_{ij} = \ln\left(\frac{x_{ij}}{x_{ik}}\right), \qquad j \in \{1,..., D\} \tag{5}$$

از این تبدیل برای مدل کردن آماری بر اساس مدل جمعی لوژستیک نرمال[23] (ALN)، که یک مدل چند متغیره با بردار میانگین و ماتریس کوواریانس داده‌های تبدیل یافته توسط رابطه (5) است، استفاده می‌شود. در روش *alr-EM* طی یک فرآیند تکراری (الگوریتم EM) مقادیر گم‌شده و پارامترهای مدل ALN متعاقبا به روز می‌شوند تا الگوریتم به همگرایی برسد. برای شروع الگوریتم پارامترهای مدل باید مقدار دهی اولیه شوند که می‌توان آنها را از مقادیر مشاهده شده محاسبه کرد. در k امین تکرار یک مقدار گم‌شده y_{ij} توسط مقدار مورد انتظار تخمین شرطی $\hat{y}_{ij}$ با توجه به تخمین‌های $\hat{\mu}^{(k)}$ و $\hat{\Sigma}^{(k)}$ جانهی می‌شود:

$$\hat{y}_{ij} = y_i^t \hat{\beta}_j^{(k)} = \hat{\sigma}_j^{(k)} \frac{\phi\left(\frac{\psi_{ij} - y_{ij}^t \hat{\beta}_j^{(k)}}{\hat{\sigma}_j^{(k)}}\right)}{\Phi\left(\frac{\psi_{ij} - y_{ij}^t \hat{\beta}_j^{(k)}}{\hat{\sigma}_j^{(k)}}\right)} \tag{6}$$

که در آن y_i^t بردار ترانهاده مقادیر مشاهده شده (بعد از تبدیل *alr*)، $\hat{\beta}_j^{(k)}$ بردار ضرایب رگرسیونی برآورد شده در k امین تکرار، $\hat{\sigma}_j^2$ برآورد واریانس شرطی متغیر y_j در kامین تکرار، ψ بردار مربوط به مقدار حد تشخیص دستگاه بعد از تبدیل *alr*، ϕ و Φ به ترتیب تابع چگالی و تابع توزیع تجمعی نرمال استاندارد هستند. بعد از همگرایی الگوریتم مقادیر به دست آمده در آخرین تکرار با استفاده از رابطه (7) تبدیل معکوس می‌شوند:

$$\hat{X}_i = \begin{cases} \hat{x}_{ij} = \dfrac{\exp(\hat{y}_{ij})}{\sum_{j\neq D}\exp(\hat{y}_{ij})+1} \\ \hat{x}_{iD} = \dfrac{1}{\sum_{j\neq D}\exp(\hat{y}_{ij})+1} \end{cases} \tag{7}$$

بعد از اجرای تبدل معکوس *alr* مجموعه داده کامل شده در دست خواهد بود.

3-5- الگوریتم *DA* نسبت لگاریتمی جمعی (*alr-DA*)[24]

این روش شامل الگوریتم داده افزایی(DA) بر اساس مدل ALN بریده شده[25] نرمال است. این الگوریتم یک حلقه تکراری بر پایه شبیه‌سازی است که مقادیری را از توزیع پسین[26] مقادیر سانسور شده و پارامترهای مدل ALN تولید می‌کند. مدلسازی ALN بر اساس معادلات رگرسیون در فضای تبدیل یافته *alr* خواهد بود. بنابراین در k امین تکرار مقدار سانسور شده در فضای *alr* به صورت تصادفی از توزیع زیر جانهی می‌شود:

$$\hat{y}_{ij}^{(k)} \sim TruncNormal\left(y_i^t \hat{\beta}_j^{(k)}, \hat{\sigma}_j^{2(k)}, y_{ij} < \psi_{ij} \mid \hat{\mu}^{(k)}, \hat{\Sigma}^{(k)}\right) \tag{8}$$

این طرح شباهت زیادی به الگوریتم بر پایه EM دارد (بخش 3-4). تفاوت اصلی آنها این است که در روش *alr-DA* تخمین پارامترها توسط شبیه‌سازی صورت می‌گیرد. بعد از همگرایی الگوریتم با استفاده از رابطه (7) مقادیر تبدیل معکوس و به واحد اصلی برگردانده می‌شوند.

3-6- الگوریتم *EM* نسبت لگاریتمی ایزومتریک(*ilr-EM*[27])

در آخر روش ارائه شده توسط فرناندز و همکاران[9] معرفی می‌شود. در این روش هدف کاهش تأثیر مقادیر دور افتاده بر روی مقادیر جانهی شده است. این روش بر پایه یک مدل نرمال چند متغیره بر روی داده‌های تبدیل یافته *ilr* هستند. تبدیل *ilr* برای یک ترکیب D-جزئی $X_i = [x_{i1},..., x_{iD}], i = 1,..., n$ به صورت زیر تعریف می‌شود:

$$ilr(x_{ij}) = z_{ij} = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{\sqrt[D-j]{\prod_{l=j+1}^{D} x_{il}}}{x_{ij}}, \quad for \quad j = 1,\ldots,D-1 \qquad (9)$$

و تبدیل معکوس ilr به صورت زیر است:

$$\begin{cases} x_{i1} = \exp\left(-\sqrt{\frac{D-1}{D}} z_{i1}\right) \\ x_{ij} = \exp\left(\sum_{l=1}^{j-1} \frac{1}{\sqrt{(D-l+1)(D-l)}} z_{il} - \sqrt{\frac{D-j}{D-j+1}} z_{ij}\right) \\ x_{iD} = \exp\left(\sum_{l=1}^{D-1} \frac{1}{\sqrt{(D-l+1)(D-l)}} z_{il}\right) \end{cases} \qquad (10)$$

این روش بر اساس یک مدل رگرسیون مقاوم سانسور شده بنا گشته است. مراحل اجرای این الگوریتم را می‌توان به صورت زیر خلاصه کرد.

1. انتخاب یک عنصر (x_j) که دارای مقادیر گم‌شده است و اجرای تبدیل ilr آن بر اساس رابطه (9).
2. استفاده از رابطه (6) بر اساس رگرسیون مقاوم سانسور شده z_1 بر روی z_2، ...، z_{D-1} به منظور تخمین سلول‌های گم‌شده و تکرار طرح EM تا زمان رسیدن به همگرایی.
3. اجرای تبدیل معکوس با استفاده از رابطه (10).
4. تکرار مراحل 1-3 برای تمام متغیرهای دارای گم‌شدگی.

4- نحوه اجرای روش‌های جانهی

تمام روش‌های ذکر شده توسط نرم‌افزار R که یک نرم‌افزار محبوب در محاسبات آماری است به سادگی قابل اجرا است. توابع مورد نیاز برای اجرای روش‌ها در دو بسته[28] *zcompositions* و *robcomsitions* در این نرم‌افزار در دسترس و قابل استفاده است. توابع مربوط به اجرای روش‌ها به ترتیب آورده شده در بالا عبارتاند از: *multRepl(.)*، *multLN(.)*، *impKNNa(.)*، *lrEM(.)*، *lrDA(.)*، *impRZilr(.)*.

5- مطالعه موردی (ناحیه ظفرقند)

در این تحقیق به منظور اجرای روش‌های مختلف جانهی برای داده ژئوشیمیائی از داده‌های ناحیه ظفرقند که در جنوب شرقی اردستان در 110 کیلومتری شمال شرق اصفهان است، استفاده شده است. از نظر موقعیت زمین‌شناسی بخشی از زون ایران مرکزی محسوب شده ودر نوار ولکانوپلوتونیک ارومیه- دختر واقع است. مطالعات ژئوشمیایی کل سنگ نشان می‌دهد که سنگ‌های اصلی تشکیل دهنده منطقه اکثرا از نوع بازالت-آندزیت هستند. مشاهدات صحرایی ومیکروسکپی نشان می‌دهند که در این منطقه تناوبی از ماگماتیسم بی مدال اسیدی و بازی وجود دارد[14].

5-1- داده‌های اصلی

مجموعه داده اصلی شامل نتایج آنالیز 250 نمونه جمع‌آوری شده از قسمت‌های مختلف ناحیه بوده است که به روش *MASS-ICP* صورت گرفته است. این مجموعه داده کامل بوده و دارای مقادیر سانسور شده نیست. نمونه‌ها برای 43 عنصر آنالیز شده است و در جدول 1 عناصر و حد تشخیص دستگاه اندازه‌گیری برای هر کدام آورده شده است.

5-2- ایجاد داده‌های سانسور به طور مصنوعی در داده‌های اصلی

به منظور ارزیابی روش‌های اشاره شده برای جانهی داده‌های سانسور شده، بر اساس مکانیزم زیر در مجموعه داده‌های اصلی داده‌های سانسور شده ایجاد شده است:

$$CM = \begin{cases} Censored & if \ x_{ij} \le n \times DL_{x_j} \\ Observed & otherwise \end{cases} \qquad (11)$$

در حالت اول با استفاده از رابطه (12) چهار مجموعه داده تولید شده است که فقط عنصر بریلیوم دارای داده‌های سانسور است. به ازای مقادیر $n = \{2.5, 4.25, 5, 5.5\}$ در رابطه (12) در داده‌های عنصر بریلوم ٪5/7، 15٪، 25% و 40٪ داده سانسور به طور مصنوعی ایجاد شده است.

در حالت دوم 13 عنصر از 43 عنصر بطور تصادفی انتخاب شده است و با استفاده از رابطه (12) درصدهای متفاوتی از سانسور در آنها ایجاد شده است و یک مجموعه داده‌ای جدید ساخته شده است. 13 عنصری که از 43 عنصر دارای داده سانسور هستند و همچنین درصد سانسور در هر کدام در جدول 2 آورده شده است. در این حالت میانگین درصد سانسور در کل مجموعه داده برابر 15% است.

جدول 1: لیست عناصر آنالیز شده و حد تشخیص دستگاه (در واحد *ppm*) متناظر هر عنصر

عنصر	حد تشخیص	عنصر	حد تشخیص	عنصر	حد تشخیص	عنصر	حد تشخیص
Au	0/001	*Bi*	0/2	*Nb*	1	*Th*	0/5
Al	100	*Cd*	0/1	*Ni*	1	*Ti*	10
Ca	100	*Ce*	1	*P*	10	*Tl*	0/2
Fe	100	*Co*	1	*Pb*	1	*U*	0/5
K	100	*Cr*	1	*Rb*	1	*V*	2
Mg	100	*Cs*	0/5	*S*	50	*W*	0/5
Na	100	*Cu*	1	*Sb*	0/5	*Y*	0/5
Ag	0/1	*La*	1	*Sc*	0/5	*Yb*	0/2
As	0/5	*Li*	1	*Sn*	0/5	*Zn*	1
Ba	2	*Mn*	5	*Sr*	2	*Zr*	5
Be	0/2	*Mo*	0/5	*Te*	0/1		

جدول 2: عناصر دارای گم‌شدگی در حالت دوم و درصد سانسور هر کدام

عنصر	درصد سانسور
Ag	44/6
Au	14/34
Bi	16/73
Co	34/66
Cu	30/67
Li	15/14
Mo	27/5
Na	1/6
Nb	0/4
Te	6/7
U	39/44
W	61/75
Zn	16/73

6- مقایسه روش‌های مختلف جانهی

برای ارزیابی روش‌های مختلف اشاره شده برای جانهی داده‌های سانسور شده معیاری برای مقایسه مقدار واقعی و مقدار جانهی شده به عنوان ریشه دوم نرمال‌شده میانگین مربع خطا[29] (*NRMSE*) در نظر گرفته شده است که به صورت زیر تعریف می‌شود:

$$NRMSE = \left(\frac{mean((X_{true} - X_{imp})^2)}{\mathrm{var}(X_{true})} \right)^{1/2} \qquad (12)$$

که در آن X_{true} ماتریس داده‌های اصلی، X_{imp} ماتریس داده‌های جانهی شده هستند.

در ابتدا مجموعه داده‌های ساخته شده در دو حالت که نحوه تولید آنها در بالا توضیح داده شد مورد استفاده قرار گرفت و با استفاده از روش‌های گفته شده داده‌های سانسور شده در آنها جانهی شدند. سپس میزان *NRMSE* برای هر روش در مجموعه داده‌های مختلف محاسبه شد که در جدول 3 آورده شده است.

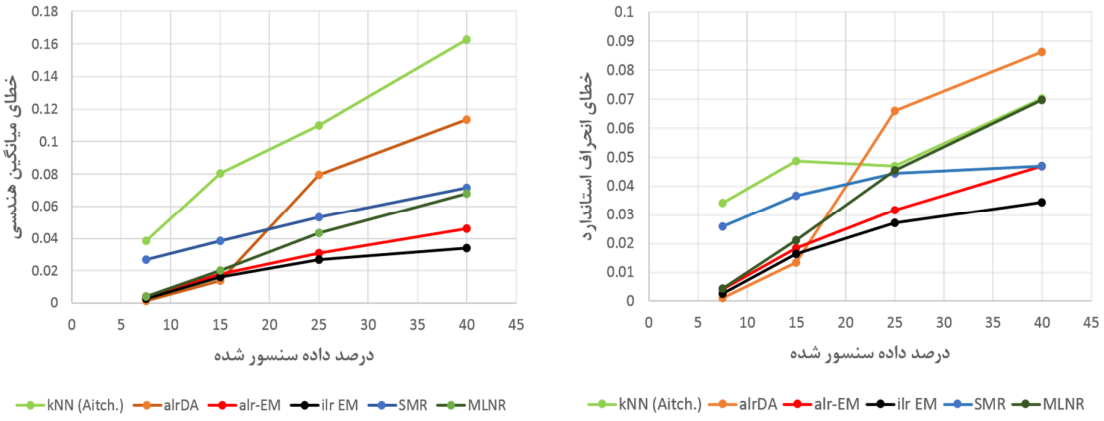
روش *KNN* در مقایسه با سایر روش‌ها نتایج خوبی ارائه نکرده است که احتمالا مربوط به مکانیزم گم‌شدگی است و این روش برای جانهی زمانی که مکانیزم گم‌شدگی *NMAR* است مناسب نیست. روش‌های تک متغیره جایگزینی ضربی ساده و نرمال تقریبا نتایج مشابهی ارائه کرده‌اند.

در بین روش‌های چند متغیره روش *ilr-EM* دقت بیشتری دارد. به منظور مقایسه تغییرات ایجاد شده در آماره‌های توزیع، در شکل 1 میزان خطای برآورد میانگین هندسی و انحراف استاندارد در روش‌های مختلف جانهی برای عنصر بریلیوم محاسبه شده و نمایش داده شده است. زمانی که درصد داده‌های سانسور شده پایین است روش‌های جانشینی ضربی لاگ نرمال، *alr-EM* و *ilr-EM* نتایج مشابهی ارائه کرده‌اند اما زمانی که درصد داده‌های سانسور بالا باشد روش *ilr-EM* بهترین دقت را دارد. با افزایش میزان داده‌های سانسور شده از دقت روش‌ها نیز کاسته می‌شود اما میزان کاهش دقت در روش‌های مختلف

متفاوت است (شکل 1). در شکل‌های 2 هیستوگرام عنصر بریلیوم با داده‌های اصلی و بعد از جانهی مقادیر سانسور شده توسط روش‌های مختلف به منظور مقایسه آورده شده است. همانطور که مشاده می‌شود روش‌های تک متغیره با جانهی یک مقدار ثابت، باعث تغییر زیادی در هیستوگرام به ویژه زمانی که نرخ سانسور بالا است، می‌شود. در شکل 3، *Q-Q plot* برای بریلیوم داده‌های اصلی و بعد از جانهی مقادیر سانسور شده (در حالتی که 25٪ داده سانسور شده است) به منظور مشاهده بهتر این تغییرات نشان داده شده است. قابل توجه است که از نظر زمان انجام محاسبات رایانه‌ای روش‌های تک متغیره زمان کمتری نیاز دارند.

جدول 3: *NRSME* برای روش‌های مختلف جانهی در دو حالت در نظر سانسور

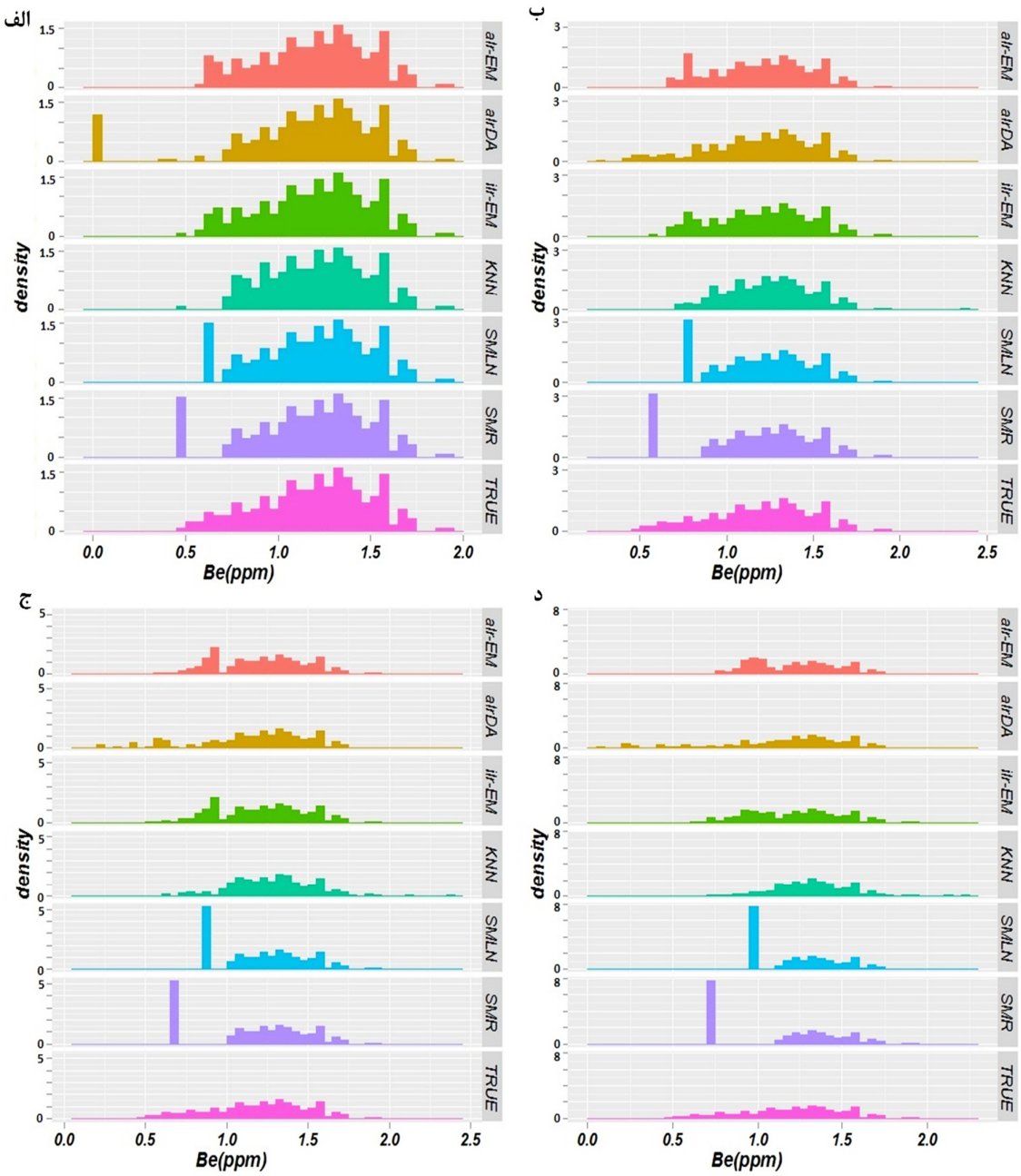
روش جانهی	*NRSME*				
	حالت اول				حالت دوم
	7/5%	%15	%25	%40	
kNN (Aitch.)	6/053	5/034	3/510	2/738	50/516
lrDA	1/303	1/780	2/482	2/180	1/040
alr-EM	1/054	1/047	1/072	0/950	0/530
ilr EM	1/083	0/993	1/035	0/873	0/504.
SMR	2/525	1/714	1/410	1/320	0/616
MLNR	1/082	1/218	1/231	1/210	0/550



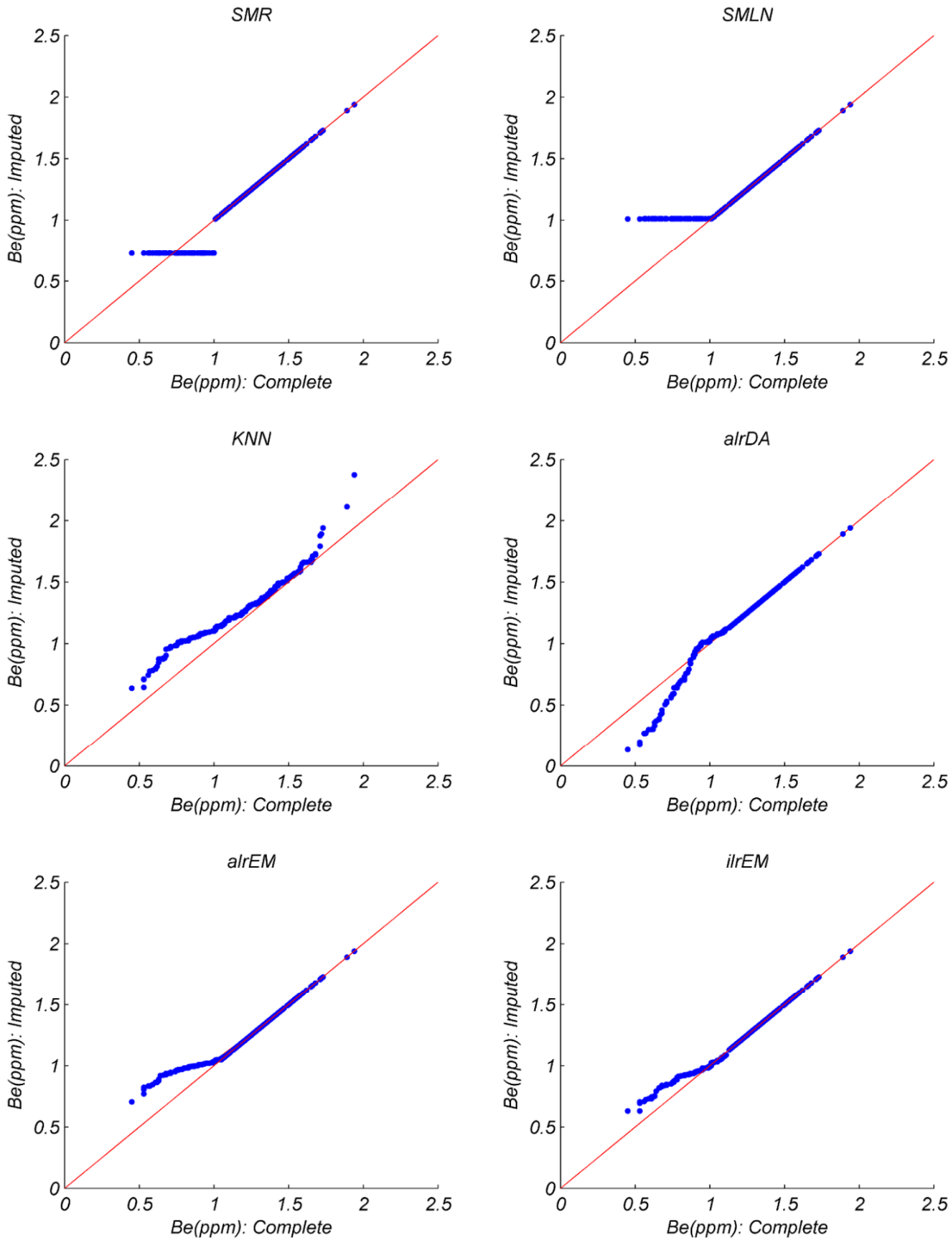
شکل 1: خطای تخمین میانگین هندسی (چپ) و انحراف استاندارد (راست) برای روش‌های مختلف جانهی در حالت اول

7- نتیجه‌گیری

اغلب مجموعه داده‌های ژئوشیمیایی شامل داده‌های سانسور یا گم‌شده هستند که در برخی مواقع درصد گم‌شدگی باعث محدودیت استفاده از روش‌های آماری و اریب شدن نتایج تجزیه و تحلیل آنها می‌شود. از طرفی داده‌های ژئوشیمیایی ماهیت ترکیبی دارند که در روش‌های جانهی باید مورد توجه باشد. روش‌های مختلفی برای جانهی داده‌های گم‌شده برای داده‌هایی که ماهیت ترکیبی دارند ارائه شده است که می‌توان از آنها در داده‌های ژئوشیمیایی استفاده کرد. روش‌های تک متغیره مانند جایگزینی ضربی لاگ نرمال زمانی که نسبت داده‌های سانسور پایین باشد (کمتر از 10٪) مناسب است و از نظر محاسبات کامپیوتری با توجه به حجم زیاد داده‌ها زمان کمتری نیاز دارد. بطور کلی روش‌های چند متغیره مدل پایه نتایج بهتری ارائه می‌کنند. استفاده از روش‌های چند متغیره با توجه به این که از روابط بین متغیرها را در نظر می‌گیرند، برای حفظ ساختار کوواریانس بین داده‌ها مناسب‌تر است. در بین روش‌های چند متغیره روش *ilr-EM* نتایج بهتری ارائه می‌دهد.



شکل 2: هیستوگرام بریلیوم با داده‌های اصلی و بعد از جانهی مقادیر سانسور شده با روش‌های مختلف. الف) ٪**7/5** درصد داده سانسور شده، ب) **15**٪ ج) **25**٪ د) **40**٪ داده سانسور شده.



شکل **3**: *Q-Q plot* بریلیوم داده‌های اصلی و بعد از جانهی مقادیر سانسور شده با روش‌های مختلف در حالتی که **25**٪ داده سانسور شده است.

7- مراجع

[1] Rubin, D. B., & Little, R. J. (2002). Statistical analysis with missing data. Hoboken, NJ: J Wiley & Sons.
[2] Buccianti, A., & Grunsky, E. (2014). Compositional data analysis in geochemistry: Are we sure to see what really occurs during natural processes?. Journal of Geochemical Exploration, 141, 1-5.
[3] Aitchison, J. (1986). The Statistical Analysis of Compositional Data. New York: Chapman and Hall. 416p.
[4] Hron, K., Templ, M., & Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. Computational Statistics & Data Analysis, 54(12), 3095-3107.
[5] de Caritat, P., & Grunsky, E. C. (2013). Defining element associations and inferring geological processes from total element concentrations in Australian catchment outlet sediments: multivariate analysis of continental-scale geochemical data. Applied Geochemistry, 33, 104-126.
[6] Carranza, E. J. M. (2011). Analysis and mapping of geochemical anomalies using logratio-transformed stream sediment data with censored values. Journal of Geochemical Exploration, 110(2), 167-185.
[7] Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2013). Values below detection limit in compositional chemical data. Analytica chimica acta, 764, 32-43.
[8] Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. Computers & Geosciences, 34(8), 902-917.
[9] Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., & Palarea-Albaladejo, J. (2012). Model-based replacement of rounded zeros in compositional data: classical and robust approaches. Computational Statistics & Data Analysis, 56(9), 2688-2704.
[10] Palarea-Albaladejo, J., Martín-Fernández, J. A., & Buccianti, A. (2014). Compositional methods for estimating elemental concentrations below the limit of detection in practice using R. Journal of Geochemical Exploration, 141, 71-77.
[11] Helsel, D. R. (2011). Statistics for censored environmental data using Minitab and R (Vol. 77). John Wiley & Sons.
[12] Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Mathematical Geology, 35(3), 253-278.
[13] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520-525.
[14] Sadeghian, M., & Ghaffary, M. (2011). The Petrogenesis of Zafarghand Granitoid Pluton (Se of Ardestan). Petrology, (6), 47-70. (In Persian)

1- Missing data
2- Below Detection limit
3- Complete Case Analysis
4- Impute
5- Compositional
6- Aitchison
7- Simple multiplicative Replacement
8- Lognormal multiplicative replacement
9- Expectation Maximization
10- Data augmentation
11- Less than
12- Greater than
13- Detection limit
14- Not missing at random
15- Simple multiplicative replacement
16- Multiplicative lognormal replacement
17- likelihood-based
18- k-nearest neighbor
19- Outlier
20- Robust
21- additive logratio Expectation Maximization
22- additive logratio
23- additive logistic normal
24- additive logratio Data augmentation
25- Truncated
26- Posterior distribution
27- isometric logratio Expectation Maximization
28- package
29- Normalized Root Mean Squared Error

The Comparison of Appropriate Methods in Imputation of The Censored Values in the Geochemical Datasets

S. A. Hosseini[1], S. Eftekhari Mahabadi[2], O. Asghari[3]*

1- PhD student, Simulation and Data Processing Laboratory, Dept. of Mining, University of Tehran, Iran
2- Assistant Professor, Dept. of Mathematics, Statistics and Computer Science, University of Tehran., Iran
3- Assistant Professor, Simulation and Data Processing Laboratory, Dept. of Mining, University of Tehran, Iran

* Corresponding Author: o.asghari@ut.ac.ir
(Received: March 2015, Accepted: October 2015)

Abstract

This study deals with the imputation methods of censored values in the multivariable geochemical data. Presence of the missing values causes limitation in the use of most of statistical methods, e.g. principle component analysis. Excluding the samples which include missing values bias the results and leads to the loss of information. Due to this, consideration of an appropriate approach to deal with missing values is necessary in the analysis of incomplete datasets. In this paper considering the nature of geochemical data, various approaches for imputing the missing values, which have been suggested in the recent years and are easy to be used in the R statistic software, are introduced. Finally, using the complete dataset of the Zafarghand region, these methods are compared with each other. Results show that the application of the multivariable methods in the imputation and particularly the ilr-EM method is preferable to the other methods.

Keywords

Geochemical Data, Censored Values, Imputation Methods, Compositional Nature, ilr-Em

Cite This Paper:
Hosseini, S. A., Eftekhari Mahabadi, S,. Asghari, O. (2015). "The Comparison of Appropriate Methods in Imputation of the Censored Values in the Geochemical Datasets." Journal of Analytical and Numerical Methods in Mining Engineering **5**(9): 63-72. *http://dx.medra.org/10.17383/S2251-6565(15)940916-X*