فصل چهارم

پردازش داده ها

۱-مقدمه

در۵ محدوده آنومال ۱:۲۰۰۰۰ در برگه های داویجان و ملایر برای هر نمونه ۴۸ عنصر اندازه گیری شده و سپس مورد پردازش کلی قرار گرفته است. برای پردازش داده ها ابتدا آنالیز شیمیایی رسوبات آبراهه ای در یک بانک اطلاعاتی وارد گردید. (این داده ها پس از اخذ، از طریق تایپ کامپیوتری و قرائت دوبل و کنترل خطاهای مربوطه در بانک اطلاعاتی وارد گردید.) علاوه بر داده های ژئوشیمیایی، شماره نمونه، اطلاعات لیتولوژی (بر مبنای نقشه زمین شناسی ۱:۱۰۰۰۰۰ ملایر) مربوط به سنگهای بالادست هر نمونه نیز در همان بانک ذخیره شده است. داده های خام مذکور در جدول ضمیمه آورده شده است.

بعد از این مرحله برای بخشی از داده ها، که به صورت سنسورد گزارش شده بود مقادیر جانشینی محاسبه و جایگزین مقادیر سنسورد گردید . در مرحله بعدی برای هرکدام از جوامع سنگی تعیین شده براساس نقشه زمین شناسی۱:۱۰۰۰۰۰ ملایر که دارای بیش از ۴ نمونه بوده اند، و نیز جوامعی که از طریق آنالیز کلاستر تفکیک شده اند ضرایب غنی شدگی محاسبه گردید و در نهایت جامعه کلی ضرایب غنی شدگی از اختلاط جوامع مذکور تشکیل شد و این جامعه کلی برای انجام عملیات آماری و رسم نقشه ها مورد استفاده قرار گرفت.

۲-پردازش داده های سنسورد

داده های ژئوشیمیایی معمولاً دارای مقادیر سنسورد هستند، یک مقدار سنسورد داده ای است که به صورت کوچکتر و یا بزرگتر از یک مقدار معین گزارش می شود. برای داده های ژئوشیمیایی، مقدار سنسورد به طور تیپیک در حد قابل ثبت آنالیزهای شیمیایی قرار دارد. داده های سنسورد زمانی ایجاد می شوند که یا تکنیک های آنالیز برای ثبت مقادیرکوچک یک عنصر به اندازه کافی حساس نیستند و یا تکنیک ها بسیار حساس بوده و قابلیت ثبت تمرکزهای بالای عناصر را در نمونه ها ندارد. داده های سنسورد در کار آنالیز آماری اختلال ایجاد می نمایند، چراکه اغلب تکنیک های آماری مهم نیازمند یک مجموعه کامل از داده های غیر سنسورد می باشند. در مورد تخمین مقادیر سنسورد روشهای مختلفی بکار می رود. از جمله این روشها قرار دادن ۳/۴ حد قابل ثبت برای مقادیر کوچکتر از حد قابل ثبت و ۴/۳ حد بالایی برای مقادیر بزرگتر از حد قابل ثبت می باشد. در بعضی موارد به جای این مقادیر عدد صفر قرار می دهند. مسئله ای که تصمیم گیرنده با آن مواجه است آن است که چه درصدی از جانشینی ها، بدون ایجاد خطاهای معنی دار، قابل توجیه است؟ در اینجا یک روش علمی برای تعیین مقدار جانشینی را نشان می دهیم. فرض براین است که مقدار جانشینی باید برابر باشد با میانگین مقادیر واقعی داده هایی که بصورت سنسورد گزارش شده است. در این پروژه روش بیشترین درست نمایی کوهن جهت تخمین این میانگین استفاده شده است.

گرایش داده های ژئوشیمیایی به پیروی از توزیع لاگ نرمال امری شناخته شده است. در حقیقت این روش شامل تخمین میانگین جامعه لاگ نرمال با استفاده از بیشترین درست نمایی است. سپس این میانگین تخمینی، برای محاسبه یک مقدار جانشینی تخمینی برای مقادیر سنسورد بکار می رود. برای روشن شدن بحث، ما چند عبارت و علائم مربوطه را بکار می بریم. در اینجا غلظت بوسیله X و حد قابل ثبت یا نقطه سنسورد بوسیله X_d نمایش داده می شود. مقدار جانشینی X_r، عددی است که باید جانشین هر مقدار سنسورد گردد. فاکتور جانشینی R_x نسبت مقدار جانشینی به حد قابل ثبت برای یک عنصر مشخص شده است:

$$R_x = \frac{X_r}{X_d} \qquad (۱)$$

بعنوان مثال ۳/۴ یک فاکتور جانشینی و ۳/۴ حد قابل ثبت، مقدار جانشینی مربوطه است. پس از تعیین اینکه لگاریتم غلظتها توزیع نرمال تری نسبت به داده های اولیه دارند، داده ها را برای عناصر انتخاب شده به Log_{10} تبدیل می کنیم. تبدیلات بین داده های لگاریتمی(Y) و داده های اولیه(X) بصورت زیر است:

برای هر X: $Y = Log_{10}X$ و $X = 10^y$ (۲)

برای مقدار جانشینی X_r: $Y_r = Log_{10}X_r$ و $X_r = 10^{Y_r}$ (۳)

برای حد قابل ثبت X_d: $Y_d = Log_{10}X_d$ و $X_d = 10^{Y_d}$ (۴)

با گرفتن لگاریتم از طرفین معادله (۱) فاکتور جانشینی تبدیل شده r_y را بدست می دهد:

$$r_y = Log_{10}X_r - Log_{10}X_d = y_r - y_d$$

و

$$r_x = 10^{r_y} \quad (۵)$$

تبدیلات مختلف دیگری نیز می تواند به جای Log_{10} بکار رود ولی در اینجا به علت سهولت آن در محاسبه و مزیت آن نسبت به روشهای جانشینی ساده قراردادی از آن استفاده شده است. ما از روش بیشترین درست نمایی کوهن (Cohen) جهت تخمین میانگین واقعی مجموعه داده ها استفاده کرده و سپس از نتیجه آن برای تخمین میانگین واقعی داده های سنسورد استفاده می کنیم. با استفاده از این روش میانگین کل مجموعه(μ) را تخمین می زنیم. همچنین میانگین داده های غیرسنسورد را نیز تخمین می زنیم (μ_u). حاصلضرب میانگین مجموعه داده ها، μ که با استفاده از روش کوهن(۱۹۶۱) تخمین زده می شود، در کل تعداد نمونه ها n، برابر با حاصلضرب میانگین داده های سنسورد، μ_d (نامشخص) و تعداد نمونه های سنسورد n_q، بعلاوه حاصلضرب میانگین داده های غیرسنسورد، μ_u (نامشخص)، در تعداد نمونه های غیرسنسورد، n_u می باشد. یعنی:

$$\mu = n_q\mu_q + n_u\mu_u \qquad (۶)$$

از حل معادله فوق مقدار μ_q که برای تخمین میانگین داده های سنسورد می باشد، بصورت زیر بدست می آید:

$$\mu_q = \frac{n\mu - n_u.\mu_n}{n_q} \quad (۷)$$

فرض ما بر این بوده است که میانگین تخمینی داده های سنسورد بهترین مقدار جانشینی می باشد یعنی:

$$r = \mu_q \quad (۸)$$

با استفاده از معادله (۳) و جایگزینی مقادیر با واحد اصلی آنها خواهیم داشت:

$$X = 10\mu_q \quad (۹)$$

تنها مجهول در معادله (۹) مقدار μ است که با استفاده از روش بیشترین درست نمایی کوهن بدست مس آید. در این محاسبات؛ N تعداد کل داده ها، n تعداد داده های غیرسنسورد X_0 حد قابل ثبت و یا مقدار سنسورد می باشد. مقدار میانگین کل و واریانس کل از روابط زیر محاسبه می شود:

$$\mu = X - \lambda(X - X_0) \quad (۱۰)$$

$$\delta^2 = \lambda(X - X_0)^2 + S^2 \quad (۱۱)$$

در معادلات بالا S^2 و X به ترتیب میانگین و پراش داده های غیرسنسورد هستند و λ تابع تخمین کمکی است که با در دست داشتن γ و h بدست می آید. مقادیر γ و h از روابط زیر بدست می آیند:

$$\gamma = \frac{S^2}{(X - X_0)^2} \quad (۱۲)$$

$$h = \frac{(N-n)}{N} \quad (۱۳)$$

با در دست داشتن γ و h عدد خوانده شده از روی این جدول یعنی λ بدست می آید. با جایگزینی این مقدار در معادله (۱۰) مقدار میانگین کل (μ) و سپس با استفاده از رابطه (۷) مقدار μ_q و سپس مقدار جانشینی بدست می آید.

در این پروژه عملیات فوق بر روی عناصر Ag,As,Au,Ge,Hg,Re,Se,Te که بخشی از داده های آنها به صورت سنسورد As(5), Ag(0.01), Au(0.001), Ge(0.05), Hg(0.01), Re(0.002), Se(1),

Te(0.05) گزارش شده بود، انجام گردید و مقدار جانشینی برای آنها بدست آمد. مقادیر بدست آمده و مقدار جانشینی برای هر عنصر به شرح جدول (۴-۲) می باشد. در این جدول X_0 مقدار سنسورد (حد قابل ثبت) و ، n تعداد نمونه های سنسورد، n_t تعداد کل نمونه ها، m_u میانگین بخش غیرسنسورد جامعه، $Slog$ انحراف معیار داده های لگاریتمی، γ و h مقادیر لازم برای بدست آوردن λ که طبق فرمول محاسبه می شود و λ تابع تخمین کمکی، m_t میانگین کل، m_c میانگین بخش سنسورد داده ها و در نهایت X_r مقدار جانشینی می باشد.

مقدار جانشینی X_r یک مقدار عددی است که پس از تبدیل بدست آمده است. نتایج نشان داده اند که مجموعه ای که دارای ۴۰٪ جانشینی است، نتایج صحیحی با ۹۰٪ اطمینان و مجموعه با ۸۰٪ جانشینی، نتایجی با حدود اطمینان ۶۰٪ بدست می دهند.