همایش ملی مدیریت بحران آب
The National Conference on Water Crisis Managemen
دانشگاه آزاد اسلامی واحد مرودشت ،اسفندماه ۱۳۸۸

آنالیز مولفه‌های اصلی روشی برای بررسی نحوه اصلاح الگوی برداشت داده‌ها در مطالعات صحرائی

علیرضا زمانی نوری*

چکیده

در برداشت‌های صحرائی مثل نمونه‌برداری از خاک، نمونه‌برداری کیفی از آب‌های سطحی و زیرزمینی، اندازه‌گیری بارندگی(ایستگاههای باران سنج) و ... تعیین نقاط مهم جهت نمونه‌برداری بلحاظ کاهش حجم نمونه‌ها و صرفه‌جویی در هزینه و زمان ، بسیار مهم است. در این مقاله از آنالیز مولفه‌های اصلی جهت تعیین ایستگاههای (یا نقاط نمونه‌برداری) مهم برای نمونه‌برداری و حذف ایستگاههای کم اهمیت استفاده می‌شود. آنالیز مولفه‌های اصلی یکی از روش‌های آماری چند متغیره بوده که با توجه به ساختار کوواریانس داده‌ها، تغییرات موجود در داده‌های اصلی را توصیف می‌کند. این روش بطور موردی برای تعیین اهمیت نسبی چاهها در مطالعات سطح ایستابی چاه بکار گرفته می‌شود.

کلید واژه: آنالیز مولفه‌های اصلی، نمونه‌برداری، آمار چند متغیره

* عضو هیات علمی دانشگاه آزاد اسلامی واحدساوه Ar.zamani@iau-saveh.ac.ir

مقدمه

زمانی که مطالعه جامعه بعلت دردسترس نبودن و یا نامحدود بودن مقدور نباشد مجبور به نمونه‌گیری و تجزیه و تحلیل آن و در نهایت تعمیم نتایج آن به کل جامعه هستیم. یکی از هدف‌های نمونه‌گیری کاهش هزینه‌ها و صرفه جویی در زمان می‌باشد. اگر نمونه‌گیری از جامعه صحیح باشد، حجم نمونه انتخابی می‌تواند کوچک باشد و یا بعبارت دیگر اگر اهمیت هر نمونه بلحاظ میزان اطلاعاتی که در مورد جامعه ارائه می‌دهد مشخص گردد، می‌توان در نمونه‌گیری‌های بعدی از نمونه‌هایی که اطلاعات چندانی ارائه نمی‌دهد صرف نظر کرد که این صرفه جویی در زمان و هزینه را بسیار بالا می‌برد.

آنالیز مولفه‌های اصلی با توجه به میزان همبستگی یا کوواریانس بین ایستگاهها یا نقاط اندازه‌گیری، متغیرهای نهان یا مولفه‌های اصلی را تعریف می‌کند که این مولفه‌های اصلی توانایی تفسیر تغییرات موجود در داده‌ها می‌باشد. این مولفه‌های اصلی که از رابطه خطی متغیرهای تشکیل شده‌اند برای کل منطقه مورد مطالعه صادق می‌باشند. یعنی در این تحلیل بجای m ایستگاه که وابسته بهم هستند، m مولفه اصلی مستقل تعریف می‌شود که برای کل منطقه قابل بکارگیری می‌باشد. از بین این مولفه‌های اصلی آنهایی که دارای واریانس بیشتری هستند، دارای اطلاعات بیشتری از ساختار داده‌های اصلی بوده و از اهمیت بالایی برخوردارند. بهمین جهت تنها از مولفه‌های اصلی استفاده می‌شود که دارای واریانس زیادی باشند. اولین مولفه اصلی نسبت به مولفه‌های اصلی دیگر دارای بیشترین واریانس یا اطلاعات بوده و همینطور دومین و سومین مولفه‌اصلی بلحاظ اهمیت بترتیب در رتبه دوم و سوم قرار دارند. بنابراین بجای بررسی همه مولفه‌های اصلی می‌توان تنها از این سه مولفه اصلی اول استفاده کرد و بقیه مولفه‌های اصلی که اطلاعات ناچیزی دارند را حذف کرد. بنابراین بجای m مولفه اصلی تنها ۳ مولفه اصلی اول استفاده می‌شود که این همان کاهش داده می‌باشد[۱]. در این روش m متغیر اصلی(یا ایستگاه) به سه مولفه اصلی تبدیل می‌شود. هر ایستگاهی که دارای همبستگی بیشتری با این سه مولفه اصلی اول باشد، آن ایستگاه اهمیت بالایی برخوردار بوده و در نمونه‌گیری‌های بعدی باقی خواهد ماند ولی اگر همبستگی داده‌های یک ایستگاه با سه مولفه اصلی اول ناچیز باشد، آن ایستگاه از اهمیت پایینی برخوردار بوده و در نمونه‌گیری‌های بعدی جهت صرفه‌جویی در وقت و زمان حذف می‌شود. با حذف ایستگاههای کم اهمیت تنها اطلاعات بسیار ناچیزی که حداکثر می‌تواند ۵ درصد باشد، از بین می‌روند که

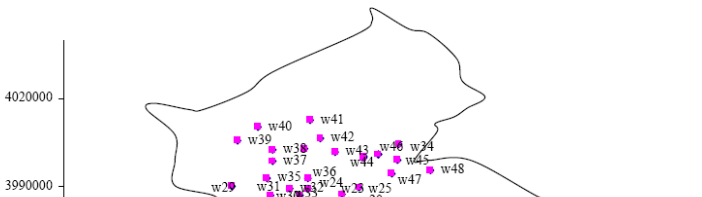
آن هم مقایسه با کاهش هزینه‌ها و صرفه‌جویی در زمان قابل صرفه‌نظر می‌باشد[۲].

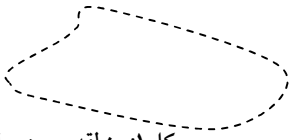
زمینه‌های کاربردی روش مولفه‌های اصلی وسیع بوده و مخصوصاً در آبهای سطحی و ، خاک استفاده‌های فراوانی دارد[۳]. از روش آنالیز مولفه‌های اصلی برای بررسی تغییرات مکانی و زمانی آلودگی آبهای زیرزمینی استفاده می‌گردد [٤، ٥ و ٦]. Wan (۲۰۰۹) از روش مولفه‌های اصلی برای بررسی روند تغییر اقلیم استفاد کرد و یک متغیر جدید برای بررسی روند تغییر اقلیم پیشنهاد کرد[۷]. Sauquet و همکارانش (۲۰۰۰) برای بررسی رژیم رودخانه و بررسی روند تغییر آن روش مولفه‌های اصلی را بکاربرند[۸]. Hisdal (۲۰۰۰) برای ناحیه‌بندی ایستگاههای هیدرومتری جهت آنالیز فراوانی منطقه‌ای سیلاب از آنالیز مولفه‌های اصلی استفاده کرد و در یک نمودار اولین مولفه‌اصلی ایستگاهها را در مقابل دومین مولفه اصلی ایستگاهها رسم کرد و به این نتیجه رسید که هرچقدر این نقاط نزدیک هم باشند نشان دهنده همگنی ایستگاهها می باشند[۹].

مواد و روش‌ها

۱.۲ منطقه مورد مطالعه

منطقه مورد مطالعه دشت قیدار می‌باشد که حدود ۹۲۰ کیلومتر مربع از آن توسط پهنه دشت پوشیده شده و بالغ بر۱۵٤۰ کیلومتر مربع نیز توسط ارتفاعات دربر گرفته شده است. این ناحیه در حد فاصل عرضهای جغرافیایی ۳۵ درجه و ۳۵ دقیقه تا ۳٦ درجه و ۱۸ دقیقه شمالی و طولهای جغرافیایی ٤۸ درجه و ۳۰ دقیقه تا ٤۹ درجه و ۱۸ دقیقه شرقی قرار گرفته است. منطقه قیدار از جمله نواحی کوهستانی ایران به شمار می‌آید که از منابع غنی آبهای سطحی برخوردار است که در سالهای اخیر به علت خشکسالی و بهره‌برداری بیش از حد مجاز افت زیادی در سطح آب زیرزمینی منطقه مشاهده گردیده است. در شکل ۱ محدوده مورد مطالعه و موقعیت چاههای منطقه نمایش داده شده است. در این تحقیق از بین ٤۸ چاه نظارت شده دشت قیدار، ۱۳ چاه $W_{۷}$، $W_{٦}$، $W_{۹}$، $W_{۱۰}$، $W_{۱۱}$، $W_{۱٤}$، $W_{۱٥}$، $W_{۱٦}$، $W_{۱}$، $W_{٥}$، $W_{۲}$، $W_{٤}$ و $W_{۳}$ برای مطالعه موردی انتخاب شده است. که موقعیت آنها بلحاظ موقعیت جغرافیایی در شکل ۱ نمایش داده شده است. برای آنالیز مولفه‌های اصلی از داده‌های سالانه سطح آب زیرزمینی این چاهها که از سالهای ۱۳۷۳ تا ۱۳۸۵ ثبت گردیده استفاده شده است.





کل۱: منطقه مورد مطالعه و موقعیت چاهها

(محدوده چاههایی که در این تحقیق مورد مطالعه می‌باشد با خط چین مشخص شده است)

۲.۲ روش آنالیز مولفه‌های اصلی

فرض شود ماتریس X یک ماتریس n×p باشد که n تعداد مشاهدات برای p متغیر است. در این تحقیق n تعداد سالهای آماری است که عمق آب زیرزمینی اندازه‌گیری شده و p تعداد چاهها یا تعداد ایستگاهها می‌باشد. با استفاده از همبستگی موجود در p چاه مجاور هم، به کمک روش آنالیز مولفه‌های اصلی اهمیت نسبی هر چاه(یا هر نموته) در نمایش تغییرات عمق آب زیرزمینی آبخوان(یا جامعه) تعیین می‌گردد. در روش آنالیز مولفه‌های اصلی، مولفه اصلی های بصورت توابع خطی زیر تعریف می‌شوند[٦].

$$
\begin{aligned}
z_1 &= Xa_1 = a_{1,1}x_1 + a_{2,1}x_2 + \ldots + a_{p,1}x_p \\
z_2 &= Xa_2 = a_{1,2}x_1 + a_{2,2}x_2 + \ldots + a_{p,2}x_p \\
&\vdots \\
z_p &= Xa_p = a_{1,p}x_1 + a_{2,p}x_2 + \ldots + a_{p,p}x_p
\end{aligned}
\qquad (۱)
$$

که در آن $a_{i,j}$ عنصر i ام از مولفه اصلی j ام بوده و a_j ضریب تبدیل متغیرهای اصلی(X) به j امین مولفه‌های اصلی(z_j) می‌باشد. با استفاده از خواص ماتریس ها می‌توان ثابت کرد که ضرایب مولفه‌های اصلی(a_j)، بردارهای ویژهٔ مربوط به ماتریس کوواریانس، S ، می‌باشند. مقدار و بردار ویژه ماتریس S از روابط زیر محاسبه می‌گردد.

$$|S - \lambda I| = 0 \qquad (۲)$$

در رابطه فوق I یک ماتریس واحد P×۱ بوده و S ماتریس کوارایانس مرتبه P است که از رابطه زیر قابل محاسبه است.

$$S = X^T X / n - 1 \qquad (۳)$$

در رابطه فوق T علامت ترانهاده می‌باشد. محدودیت‌های حل معادله ۲ عبارتند از:

- عمود بودن بردارهای ویژه $(i \neq j, a_j^T a_i = a_i^T a_j = 0)$
- نرمال یا یکه بودن بردارهای ویژه $(a_j^T a_j = 1)$

محدودیت‌های فوق باعث می‌گردد جواب‌های معادله ۲ یگانه بوده و مولفه‌های اصلی،z_j، مستقل باشند. اگر $a_1, a_2, ..., a_p$ بترتیب بردارهای ویژه مربوط به مقادیر ویژه $\lambda_1, \lambda_2, ..., \lambda_p$ باشند(بطوریکه برای $i < j$ ، $\lambda_i > \lambda_j$ است) آنگاه معادله ۱ بصورت زیر نمایش داده می‌شود.

$$Z = X A \qquad (٤)$$

که در آن $Z = (z_1, z_2, ..., z_p)$ و $A = (a_1, a_2, ..., a_p)$ است.

واریانس مولفه‌های اصلی، z_j، همان مقادیر ویژه متناظر آنها می‌باشد. یعنی واریانس اولین مولفه اصلی($z_۱$) برابر $\lambda_۱$ است. پس اولین مولفه اصلی بیشترین واریانس را داشته که نشان دهنده بالا بودن قابلیت اولین مولفه اصلی در شناسایی تغییرات سطح آب زیرزمینی می‌باشد. اولین مولفه اصلی خطی است که امتداد آن منطبق با بیشترین پراکندگی قابل مشاهد در داده‌های اصلی است. دومین مولفه اصلی(z_j) دارای واریانس $\lambda_۲$ بوده که از لحاظ مقدار واریانس در رتبه دوم قرار دارد و امتداد آن در راستایی است که پراکندگی قابل مشاهده داده‌ها در رتبه دوم قرار دارد. بقیه مولفه‌های اصلی نیز به همین ترتیب توصیف می‌گردند. مولفه‌های اصلی از مرکز داده‌های اصلی عبور کرده و دو به دو برهم عمود هستند.

برای محاسبه اهمیت نسبی هر چاه(یا هر نمونه)، از آماره t که دارای توزیع tاستیودنت با درجه آزادی n-۲ است استفاده می‌گردد. این آماره بصورت زیر تعریف می‌شود.

$$t = r\sqrt{\frac{n-2}{1-r^2}} \qquad (۵)$$

در رابطه فوق n تعداد داده‌ها و r ضریب همبستگی بین متغیر اصلی و مولفه اصلی می‌باشد. ضریب همبستگی بین چاه i (x_i) و مولفه اصلی jام (z_j) از رابطه زیر قابل محاسبه است.

$$r(z_j, x_i) = \lambda_j^{1/2}\, a_{i,j} \qquad (٦)$$

در رابطه فوق $a_{i,j}$ عنصر i ام از مولفه اصلی j ام می‌باشد. هرچقدرمقدار آماره t برای یک ایستگاه بالا باشد نشان‌دهنده بالا بودن اهمیت نسبی آن ایستگاه(یا چاه) است و ایستگاههای که مقدار آماره t از حد بحرانی t استیودنت در سطح معنی دار ۵ درصد و با درجه آزادی n–۲، کمتر باشد آنگاه می‌توان دریافت که اهمیت نسبی آن چاه‌ها و یا ایستگاه‌ها پایین بوده و می‌توان آنها را حذف نمود.

بحث و نتایج

برای آنالیز مولفه‌های اصلی جهت تعیین اهمیت نسبی هر چاه، ابتدا ماتریس داده‌ها(X) تشکیل شده و با رابطه ٤ ماتریس کوواریانس ایستگاهها محاسبه گردید. در این تحقیق ۱۳ چاه $w_۷$، $w_٦$، $w_۹$، $w_۱۰$، $w_۱۱$، $w_۱٤$، $w_۱٥$، $w_۱٦$، $w_۱$، $w_٥$، $w_۲$، $w_٤$ و $w_۳$ ، برای مطالعه موردی انتخاب شده که داده‌های مورد استفاده از آنها متوسط سالانه سطح آب زیرزمینی در ۱۳ سال گذشته است. بنابراین ماتریس داده‌های اصلی (X) و ماتریس کوواریانس آنها یک ماتریس ۱۳ ۱۳ می‌باشند. برای ماتریس کوواریانس تشکیل شده مقادیر و بردارهای ویژه بکمک رابطه ۲ استخراج گردید. این بردارهای ویژه و مقادیر ویژه بترتیب همان ضرایب و واریانس مولفه‌های اصلی می‌باشند. اهمیت نسبی و یا میزان اطلاعاتی که هر مولفه اصلی از داده‌های اصلی دارد از نسبت واریانس هر مولفه اصلی به مجموع واریانس مولفه‌های اصلی بدست می‌آید. نتایج بدست آمده نشان می‌دهد که سه مولفه اصلی اول بترتیب دارای ٦٨/٢، ٢١/٦ و ٤/٨ از اطلاعات داده‌های اصلی بوده و بقیه ۱۰مولفه‌های اصلی تنها شامل ٥/٤ درصد از اطلاعات داده‌های اصلی می‌باشند که با حذف آنها اطلاعات ناچیزی از بین می‌رود. در جدول ۱ ضرایب همبستگی بین هر

چاه و سه مولفه اصلی اول نشان داده شده است که با استفاده از رابطه ٦ محاسبه گردیده است که از آنها برای تعیین اهمیت نسبی هر چاه استفاده می‌شود. برای محاسبه اهمیت نسبی هر چاه از آماره t که در رابطه ٥ تعریف شده استفاده می‌شود. هر چاهی که آماره t در آن بالا باشد اهمیت نسبی آن بالا خواهد بود. در سطح معنی‌داری ٥ درصد مقدار بحرانی آماری t برابر ۲/۱۱ می‌باشد. بنابراین چاههایی که مقدار آماره t در آنها کمتر از ۲/۱۱ است از اهمیت کمتری برخوردار می‌باشد. همانطور که در جدول ۱ نشان داده شده است چاههای W٦، W۲ و W٤ از اهمیت پایینی برخوردار بوده و می‌توان جهت صرفه جویی در هزینه‌ها و زمان از نمونه‌برداری بعدی حذف گردند که با حذف آنها اطلاعات بسیار ناچیزی حذف می‌گردد. این روش را می‌توان برای تمامی چاههای منطقه انجام داد و از چاههایی که دارای اهمیت پایینی هستند صرفنظر کرد.

جدول ۱: ضرایب همبستگی بین چاهها و سه مولفه اصلی اول و مقدار آماره t برای چاههای مختلف

چاهها(x_i)	ضریب همبستگی بین هر چاه با سه مولفه اصلی اول			آماره t		
	r_1	r_2	r_3	t_1	t_2	t_3
W_{10}	۰.۲٦٤	۰.٦۱۲	۰.۰۷۹	۰.۹۰۷*	۲.٥٦٦	۰.۲٦۲
W_9	۰.٦۷	۰.۱۲۲	-۰.۱۳۸	۲.۹۹۳	۰.٤۰۷	-۰.٤٦
W_6	۰.۰۱	۰.٤٥۱	۰.۱٦۸	۰.۰۳۳	۱.٦۷٥	۰.٥٦٥
W_{10}	۰.٦۲٥	-۰.۰۸۷	۰.۱۷٦	۲.٦٥٥	-۰.۲۸	۰.٥۹۲
W_7	۰.۳٤٥	۰.٥٤٥	-۰.۰۳۷	۱.۲۱۹	۲.۱٥٥	-۰.۱۲۲
W_{11}	۰.٥٦۲	۰.۱۸٦	۰.٤٤٥	۲.۲٥۳	۰.٦۲۷	۱.٦٤۸
W_{14}	۰.٦۸٥	۰.٤۱۲	-۰.۲۰٥	۳.۱۱۸	۱.٤۹۹	-۰.٦۹٤
W_5	۰.۱۳٦	۰.٦۱۸	۰.۱۲٦	۰.٤٥٥	۲.٦۰۷	۰.٤۲۱۲
W_2	۰.۲۳۲	۰.۳۳٥	۰.٦٦۳	۰.۷۹۱	۱.۱۷۹	۲.۹۳۷
W_4	۰.٤۹۸	۰.۳۹٥	۰.۲٦٤	۱.۹۰٤	۱.٤۲٦	۰.۹۰۷
W_3	۰.۰٤۸	۰.٥۹٦	۰.٥٤٥	۰.۱٥۹	۲.٤٦۱	۲.۱٥٥
W_1	۰.٦۸٦	۰.۰۸۲	-۰.۲۳٦	۳.۱۲٦	۰.۲۷۲	-۰.۸۰٥

W۱۶	۰.۶۷۲	۰.۱۷۲	-۰.۱۲٤	۳.۰۰۹	۰.۵۷۹	-۰.٤۱٤

*: در سطح معنی داری ۵ درصد معنی دار است یعنی چاه از اهمیت بالایی برخوردار است.

نتیجه‌گیری

آنالیز مولفه‌های اصلی یک روش آماری چند متغیره جهت کاهش حجم داده‌ها می‌باشد. در این تحقیق بجای بررسی داده‌های ۱۳ چاه مورد مطالعه تنها از سه متغیر نهان یا همان سه مولفه‌اصلی اول استفاده شد. این سه مولفه اصلی اول حدود ۹٤/٦ درصد از اطلاعات چاهها را شامل می‌شود یعنی اگر از بجای استفاده از ۱۳ چاه مورد بررسی این سه مولفه اصلی اول بکارگرفته شود تنها ۵/٦ درصد از اطلاعات از بین می‌رود. اگر هر کدام از متغیرهای اصلی(یا چاهها) دارای ضریب همبستگی بالایی با سه مولفه اصلی اول باشد آنگاه آن متغیر(یا چاه) از اهمیت بالایی برخوردار بوده و در نمونه‌گیری‌های بعدی باید سعی شود، داده‌های این متغیرها با دقت بالایی اندازه‌گیری گردد. اما اگر برای یک چاه یا یک ایستگاه ، ضریب همبستگی با مولفه‌های اصلی اول کم باشد آنگاه آن چاه یا ایستگاه از اهمیت پایینی برخوردار بوده و می‌توان جهت صرفه جویی در هزینه‌های نمونه‌برداری و کاهش زمان اندازه گیری و محاسبات از نمونه‌برداری های بعدی کنار گذاشت. کاهش حجم داده‌ها و افزایش دقت داده‌ها، میزان صحت مدل‌سازی آب زیرزمینی را بالا برده و زمان محاسبات کامپیوتری را نیز کاهش می‌دهد.

منابع

[۱] *Pop H.F., Principal components analysis versus fuzzy principal component analysis a case study: the quality of Danube water (۱۹۸۵-۱۹۹٦), Talanta ٦٥, ۱۲۱۵-۱۲۲۰.*

[۲] *Lucas L. and Jauzein M. ,Use of principal component analysis to profile temporal and spatial variations of chlorinated solvent concentration in groundwater, Environmental Pollutio ۱۵۱, ۲۰۰۸, ۲۰۵-۲۱۲.*

[۳] *Siyue L. et al. , Water quality in the upper Han River, China: The impacts of land use/land cover in riparian buffer zone, Journal of Hazardous Materials ۱٦٥, ۲۰۰۹, ۳۱۷-۳۲٤.*

[٤] *Giussani B. et al., Three-way principal component analysis of chemical data from Lake Como watershed, Microchemical Journal ۸۸, ۲۰۰۸ ۱٦۰-۱٦٦.*

[۵] Ouyang Y., Evaluation of river water quality monitoring stationa by principal component analysus, Water Research ۳۹, ۲۰۰۵, ۲۶۲۱-۲۶۳۵.

[۶] Petersen W. et al., Process identification by principal component analysis of river water-quality data, Ecological Modeling ۱۳۸, ۲۰۰۱, ۱۹۳-۲۱۳.

[۷] Wan K.L. et al, A new variable for climate change study and implications for the built environment, Renewable Energy ۳٤, ۲۰۰۹, ۹۱٦-۹۱۹.

[۸] Sauquet E. et al, Mapping mean monthly runoff pattern using EOF analysis, Hydrology and Earth System Sciences, ٤(۱), ۲۰۰۰, ۷۹-۹۳.

[۹] Hisdal.H, Methods for Regional Classification of Steamflow Drought Series: The EOF Method and L-moments، Assessment of the Regional Impact of Drought in Europe, Conceptual plan, ۲۰۰۰.