نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد 29، شماره 5، آذر – دی 1394، ص. 1218-1207

Journal of Water and Soil
Vol. 29, No. 5, Nov.-Dec. 2015, p. 1207-1218

کاربرد خوشه‌بندی سلسله مراتبی برای افزایش کارایی نگاشت ویژگی خود سامان در شناسایی مناطق همگن هیدرولوژیکی به منظور برآورد سیلاب

فرهاد فرسادنیا[1*] - بیژن قهرمان[2]

تاریخ دریافت: 1393/01/28
تاریخ پذیرش: 1394/05/28

چکیده

شناسایی گروه‌های همگن هیدرولوژیک یکی از مباحث بنیادی هیدرولوژی در دو بعد کاربردی و تحقیقاتی است. یکی از روش‌های معمول به منظور دستیابی به مناطق همگن هیدرولوژیکی برای برآورد منطقه‌ای سیلاب، استفاده از روش‌های خوشه‌بندی است. چندین تحقیق از نگاشت ویژگی خود سامان (SOM) در این خصوص استفاده شده است. تفسیر نقشه خروجی مشکل اصلی این روش است. هدف از این تحقیق کاربرد روش خوشه‌بندی دو مرحله‌ای نگاشت ویژگی خود سامان و سلسله مراتبی وارد (Ward) به منظور تعیین مناطق همگن هیدرولوژیک در حوضه‌های آبخیز استان‌های خراسان شمالی و رضوی است. ابتدا ابعاد ماتریس ورودی SOM با تحیل مولفه‌ی اصلی کاهش یافت. سپس از SOM برای تشکیل نقشه ویژگی دو بعدی استفاده شد. پس از آن گره‌های خروجی SOM بمنظور تعیین مناطق همگن در تحلیل فراوانی سیلاب به عنوان ورودی برای روش وارد به کار رفت. سپس توسط آزمون ناهمگنی هاسکینگ و والیس، پنج منطقه که از لحاظ هیدرولوژیکی از یک فرآیند سیلاب پیروی می‌کردند، شناسایی شدند. نتایج نشان داد که روش ترکیبی نگاشت ویژگی خود سامان و سلسله مراتبی وارد با ورودی مولفه‌های اصلی به مراتب کاراتر از روش‌های سلسله مراتبی تنها با ورودی‌های استاندارد شده یا مولفه‌های اصلی، در دستیابی به مناطق همگن هیدرولوژیک است.

واژه‌های کلیدی: تحلیل مولفه‌ی اصلی، تحلیل فراوانی منطقه‌ای سیلاب، خوشه‌بندی ترکیبی، گشتاورهای خطی

مقدمه

تحلیل فراوانی در هیدرولوژی مانند دبی اوج سیلاب، جریان کمینه و بارش‌های حدی اهمیت فراوانی دارد. به دلیل نادر بودن مقادیر حدی و کوتاه بودن طول دوره آماری، تخمین فراوانی وقوع این حوادث به صورت ایستگاهی دشوار است. تحلیل فراوانی منطقه‌ای سیلاب روشی مناسب برای افزایش طول دوره‌ی آماری بوده، عملکرد آن بهتر از برآورد نقطه‌ای بوده و میزان این برتری به برقراری فرضیات اساسی تحلیل منطقه‌ای سیلاب، یعنی همگنی ناحیه و استقلال داده‌ها، بستگی دارد. مطالعات زیادی به منظور برآورد شرط اول تحلیل فراوانی منطقه‌ای یعنی تعیین مناطق همگن انجام شده است. بیش‌تر این مطالعات برپایه‌ی روش‌های تحلیلی آماری چند - متغیره از جمله تحلیل خوشه‌ای هستند.

نگاشت ویژگی خود سامان (SOM[3]) نوعی شبکه عصبی غیرنظارت شونده است که در شناسایی گفتار، الگو سازی زیستی، فشرده سازی داده‌ها، پردازش سیگنال و داده کاوی کاربردی گسترده دارد (10-). شبکه‌های عصبی از نوع SOM یاد می‌گیرند که چگونه داده‌های ورودی را با شناسایی الگوهای متفاوت داده‌ها براساس شباهت آنها، به‌صورت کمی با متریک (مانند فاصله اقلیدسی) خوشه-بندی کنند. نگاشت ویژگی خود سامان در سالهای اخیر به عنوان ابزاری مفید در شناسایی مناطق همگن هیدرولوژیک به‌کار رفته است (1-، 2-، 13- و 14-).

دی‌پرینزو و همکاران (2-) با مطالعه روی 300 حوضه آبخیز در ایتالیا با استفاده از توصیف‌گرهای حوضه بعنوان ورودی SOM، به طبقه‌بندی حوضه‌های آبخیز پرداختند. نتایج آنها نشان داد که انجام تحلیل مولفه‌ی کانونی پیش از الگوریتم SOM نقش زیادی در بهبود نتایج خروجی دارد. لی و همکاران (13-) به بررسی شباهت هیدرولوژیکی 53 حوزه با شرایط آب و هوایی متفاوت در آلمان پرداختند. آنها نشان دادند که استفاده از خصوصیات فیزیوگرافی و

1 و 2- دانشجوی دکتری آبیاری و زهکشی و استاد گروه مهندسی آب، دانشکده کشاورزی، دانشگاه فردوسی مشهد
(*- نویسنده مسئول: Email: farhadfarsadnia@stu.um.ac.ir)

3- Self-Organizing Feature Maps

اقلیمی بعنوان ورودی خوشه‌بندی SOM باعث بهبود نتایج خوشه-بندی می‌شود.

هرچند SOM در خوشه‌بندی حوضه‌های آبخیز به نواحی همگن هیدرولوژیک با موفقیت استفاده شده است، اما تجزیه و تحلیل نقشه خروجی SOM دشوار است. به طور کلی این الگوریتم به ندرت خوشه‌های واضحی در خروجی نشان می‌دهد. پژوهشگران راهکارهای متفاوتی به منظور حل این مشکل ارائه کرده‌اند. لامپینن و اوجا (12-) یک الگوریتم خوشه‌بندی دو مرحله‌ای ارائه کردند که خروجی SOM اول را به‌عنوان ورودی SOM دوم استفاده می‌کند. آنها نشان دادند که الگوریتم SOM ترکیبی برای تفکیک رده‌های مختلف داده‌ها قابلیت بیشتری از الگوریتم k-میانگین کلاسیک و SOM معمولی دارد. وسانتو و الهونیمی (18-) برای خوشه‌بندی خروجی SOM از الگوریتم خوشه‌بندی سلسله مراتبی و k-میانگین استفاده کردند. آنها مهمترین برتری الگو ارائه شده را کاهش نسبتاً زیاد بارگذاری انجام شده توسط الگوریتم‌ها، امکان خوشه‌بندی مجموعه داده‌های بزرگ و بررسی چندین راهبرد پردازش متفاوت در زمان محدود اعلام کردند. فرسادنیا و مقدمنیا (3-) روش خوشه‌بندی دو مرحله‌ای را برای منطقه‌ای کردن حوضه‌های آبخیز استان مازندران استفاده کردند. مرحله اول SOM را برای شکل‌دهی نقشه ویژگی دو بعدی استفاده کردند. سپس گره‌های خروجی SOM را با الگوریتم c-میانگین فازی (FCM)[1]خوشه‌بندی نمودند.

تحلیل مولفه‌ی اصلی (PCA[2]) یک روش آماری تحلیل چند متغیره است که به منظور کاهش ابعاد متغیرهای ورودی و استخراج متغییرهای غیر همبسته بکار می‌رود. محققان با بررسی نمودارهای مولفه‌های اصلی و متعامد که از داده‌های اصلی استخراج شده‌اند، درک عمیق‌تری نسبت به داده‌های اصلی به دست می‌آورند. مشکل اصلی در استفاده از PCA در مطالعات هیدرولوژی این است که با کاهش ابعاد داده‌های ورودی بخشی از اطلاعات (پراش) موجود در داده‌های اصلی از بین می‌رود. در این مطالعه به‌منظور بررسی این موضوع از هر دو داده‌های خام و مولفه‌های اصلی به صورت جداگانه بعنوان ورودی الگوریتم خوشه‌بندی ترکیبی SOM-Ward استفاده شده است. سپس از SOM برای شکل‌دهی نقشه ویژگی دو بعدی[3] استفاده شد. پس از آن، گره‌های خروجی از SOM با الگوریتم سلسله مراتبی وارد، خوشه‌بندی شد و در نهایت با بررسی و تعدیل همگنی مناطق حاصل از الگوریتم خوشه‌بندی ارائه شده، مناطق همگن هیدرولوژیک به منظور تخمین چندک‌های منطقه‌ای سیلاب به دست آمد.

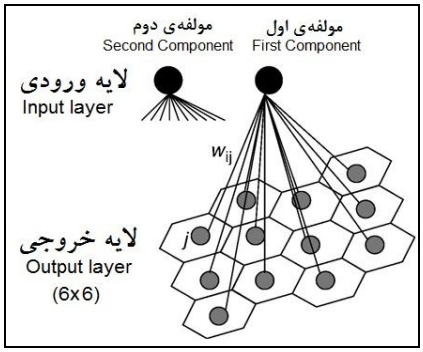
مواد و روش‌ها

منطقه مورد مطالعه و داده های مورد استفاده

منطقه مورد بررسی شامل زیرحوضه‌های سه حوضه آبخیز اصلی کشور، اترک، کشف‌رود (قره‌قوم) و کال‌شور، واقع در استان‌های خراسان شمالی و رضوی است. آمار ایستگاه‌های آب‌سنجی از شرکت مدیریت منابع آب ایران تهیه شد. اسامی ایستگاه‌های مورد استفاده در جدول 1 ذکر شده است.

نگاشت ویژگی خود سامان (SOM)

نگاشت ویژگی خود سامان، تابع چگالی احتمال از داده‌های ورودی تحت الگوریتم یادگیری غیرنظارت شونده است، که روشی موثر در خوشه‌بندی بوده و برای خلاصه سازی و بصری نمودن داده-ها به‌کار می‌رود (11-). این الگوریتم دارای خصوصیات حفاظت از همسایگی و تجزیه و تحلیل فضای ورودی متناسب با توزیع داده‌ها را دارا است (10-و 11-). SOM شامل دو لایه است: یک لایهٔ ورودی شکل گرفته از مجموعه گره‌ها[4] (یا نرون‌هایی که واحدهای[5] محاسباتی دارند) و یک لایه خروجی (لایه کوهونن) که توسط گره‌هایی که در شبکهٔ دو بعدی قرار گرفته‌اند، تشکیل شده است (شکل 1).



شکل 1- ساختار نگاشت ویژگی خود سامان

Figure 1- Schematic diagram of SOM

تعداد نرون‌های خروجی SOM (یا اندازه نقشه)[6] در یافتن انحراف داده‌ها مهم است. اگر اندازه نقشه خیلی کوچک باشد، نمی‌تواند تفاوت‌های مهم را آشکار کند. در مقابل، اگر اندازه نقشه خیلی بزرگ باشد، تفاوت‌ها خیلی کم می‌شود (21-). می‌توان تعداد نرون‌های خروجی در SOM را با استفاده از روش ابتکاری پیشنهاد شده توسط

1 - Fuzzy c-mean
2 -Principal Component Analysis
3 - SOM map

4 - Nodes
5 - Units
6 - Map size

وسانتو و همکاران (19-) انتخاب کرد. تقریباً تعداد بهینه‌ی واحدهای نقشه برابر $5\times\sqrt{N}$ می‌باشد که N تعداد نمونه‌ها در مجموعه داده‌ها هستند. لایه خروجی از S گره خروجی در شبکه‌ی شش وجهی برای فراهم کردن تجسم بهتر به دست می‌آید. به طور کلی شبکه‌ی شش وجهی نسبت به شبکه مستطیلی برتر است، زیرا هیچ یک از جهت‌های عمودی یا افقی بر دیگری برتری ندارند (11-).

هر گره در لایهٔ ورودی توسط شبکه‌ی سیناپسی به تمامی گره ها در لایه خروجی متصل است. هر گره خروجی دارای بردار ضرایب متصل به داده‌های ورودی است. بردار ضرایب وزنی (یا شدت اتصال) را با نام W بین ورودی و لایه خروجی باز می‌گرداند. وزن‌ها شبکه‌ای را بین واحدهای ورودی (بردار مشخصه‌ها) و واحدهای خروجی وابسته به آنها (گروه‌هایی از بردار مشخصه‌ها) برقرار می‌کند.

عملکرد الگوریتم به شرح زیر می‌باشد: زمانی که بردار مشخصه ورودی X' به SOM ارائه شد، گره‌ها در لایه خروجی با یکدیگر رقابت کرده و گره برنده (گره‌ای که فاصله‌ی تمامی وزن‌هایش از بردار ورودی در مقایسه با سایر گره ها کوچک تر است) انتخاب می‌شود. براساس قاعده یادگیری SOM، بردار وزن گره برنده و همسایه‌های از پیش تعریف شده‌اش در الگوریتم، براساس معادلهٔ زیر به روز رسانی می‌شوند:

$$W_{ij}(t+1)=W_{ij}+\alpha(t).h_{jc}(t)[X_i(t)-W_{ij}(t)] \qquad (1)$$

که در آن $w_{ij}(t)$ وزن بین گره i در لایه ورودی و گره j در لایه خروجی در زمان تکرار t است و $\alpha(t)$ فاکتور سرعت یادگیری است که تابعی نزولی از زمان تکرار t است و $h_{jc}(t)$ تابع همسایگی (هسته اصلی هموار سازی تعریف شده روی نقاط شبکه) است که مقدار همسایگی از گره برنده (c) در طی فرآیند یادگیری به روز رسانی می‌شود. فرآیند یادگیری تا زمانی که معیار توقف معرفی شود (معمولاً زمانی که بردار وزن ثابت شده یا زمانی که تعداد تکرارها کامل شود) ادامه می‌یابد. برای جزئیات بیشتر در مورد الگوریتم SOM می‌توان به هایکین (5-) مراجعه کرد.

ماتریس وزن‌های نهایی بعد از مرحلهٔ SOM، ماتریسی با بعد $n\times m'$ به نام ماتریس W' است.

$$W'=\begin{bmatrix} W_{11} & \dots & W_{1m'} \\ \vdots & & \vdots \\ W_{n1} & \cdots & W_{nm'} \end{bmatrix} \qquad (2)$$

از آنجا که مرز دقیق خوشه‌ها را در نقشه خروجی SOM مشخص نیست، برای مشخص کردن خوشه‌ها بر روی نقشهٔ آموزش یافته، از ماتریس وزن‌های خروجی SOM به‌عنوان ورودی در الگوریتم خوشه‌بندی سلسله مراتبی وارد استفاده شد.

خوشه بندی سلسله مراتبی وارد

الگوریتم خوشه‌بندی وارد (20-) فن تحلیل داده‌های چند متغیره‌ی پرکاربرد در مطالعات منطقه‌ای به شمار می‌رود. اساس این روش بر این فرض استوار است که اگر دو خوشه ترکیب شوند، تغییر در مقدار تابع هدف در نتیجه‌ی از دست دادن اطلاعات، تنها به رابطه‌ی بین دو خوشه‌ی ترکیب شده بستگی داشته و به رابطه‌ی خوشه‌های دیگر بستگی ندارد. جزئیات الگوریتم وارد را می‌توان در رائو و اسرینیواس (16-) جستجو کرد.

بررسی همگنی مناطق با استفاده از آزمون‌های گشتاورهای خطی

پس از شکل‌گیری خوشه‌ها، به منظور بررسی و اصلاح خوشه‌ها و یافتن مناطق همگن هیدرولوژیک از پرکاربردترین آزمون‌های بررسی همگنی براساس گشتاورهای خطی استفاده شد. دو آماره‌ی استفاده شده برای تشکیل مناطق همگن هیدرولوژیک عبارتند از: (الف) آماره ناهمنوائی[1]؛ (ب) آماره ناهمگنی[2] هاسکینگ و والیس (8-). آزمون ناهمنوائی، داده ناهمنوا را نسبت به کل گروه مشخص می‌کند و به تعداد ایستگاه در گروه وابسته است (9-). اگر مقدار آماره‌ی ناهمنوائی (D) بزرگ‌تر از 3 باشد، ایستگاه ناهمنوا بوده و از گروه حذف می‌شود.

اگر تغییر پذیری یا فضای پراکنش ایستگاه‌ها بزرگ باشد، احتمال تعلق این ایستگاه‌ها به مجموعه‌ای واحد را می‌توان آزمون ناهمگنی گشتاورهای خطی بررسی کرد که شامل سه آماره H_1, H_2, H_3 است. اگر مقدار این آماره‌ها کم‌تر از یک باشد، منطقه همگن، اگر بین 1 تا 2 باشد، منطقه احتمالاً ناهمگن و اگر بزرگ‌تر از 2 باشد، منطقه کاملاً ناهمگن است. در عمل معیار H_1 کفایت می‌کند (6-). برای جزئیات بیشتر به هاسکینگ و والیس (9-) مراجعه شود.

عملیات خوشه‌بندی توسط نرم‌افزار Matlab 2013 و SPSS 20 و محاسبه گشتاورهای خطی توسط الگوریتم نوشته شده توسط هاسکینگ (7-) به فورترن انجام شده است.

نتایج و بحث

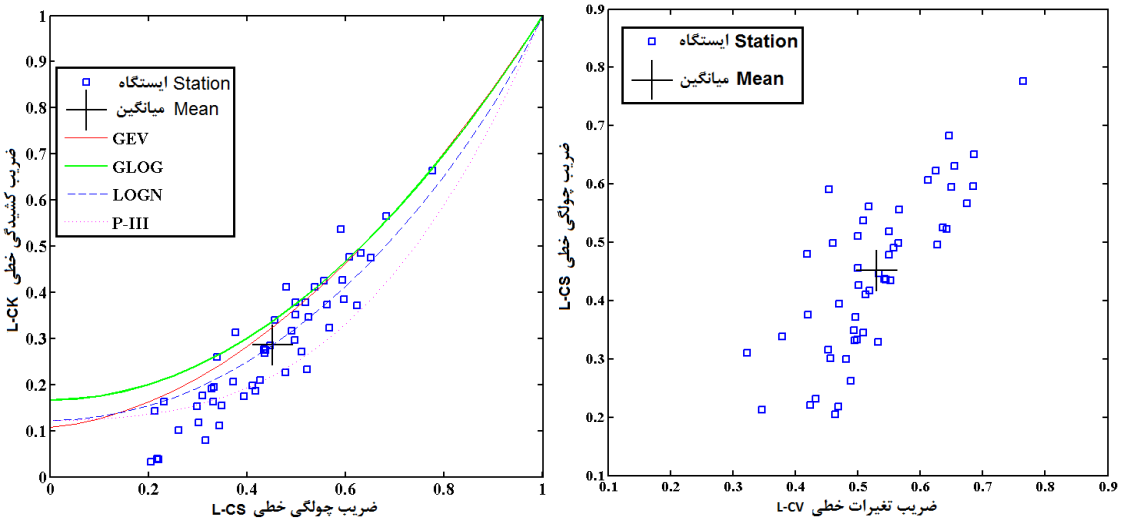
پس از اخذ آمار دبی اوج سالانه و خصوصیات فیزیوگرافی حوضه‌های آبریز، با توجه به حداقل طول دوره آماری توصیه شده توسط هاسکینگ و والیس (8-) از 68 ایستگاه موجود، ایستگاه‌های دارای آمار بالای 20 سال انتخاب گردیدند. جدول 1 مشخصات ایستگاه‌های منتخب را نشان می‌دهد.

1- Discordancy
2- Heterogeneity

جدول 1- ایستگاه های مورد مطالعه و برخی خصوصیات آنها

Table 1- Stations and some attributes in the study area

ردیف	نام ایستگاه		طول جغرافیایی	عرض جغرافیایی	میانگین دبی اوج(m3/s)	ردیف	نام ایستگاه		طول جغرافیایی	عرض جغرافیایی	میانگین دبی اوج(m3/s)
Row	Name		longitude	latitude	mean annual flood pead	Row	Name		longitude	latitude	mean annual flood pead
1	Hatam Gale	حاتم قلعه	59.37	37.30	77.20	26	Ayrghaye	آیرقایه	56.36	38.17	172.91
2	Bagh Abas	باغ عباس	59.73	35.58	52.88	27	Ghatlish	قتلیش	57.28	37.81	136.50
3	Timank	تیمنک	60.59	35.46	60.87	28	Dar Kesh	در کش	56.74	37.44	43.96
4	Darband Sangkhast	دربندسنگخواست	56.83	37.25	101.15	29	Hosein Abad Jangal	حسین آباد جنگل	58.38	36.05	70.25
5	Kapkan	کپکان	58.92	37.25	12.81	30	Beirot	بیروت	58.12	35.73	101.57
6	Mohamad Taghi Beik	محمد تقی بیک	58.64	37.62	45.66	31	Dahane Shor	دهنه شور	58.08	36.59	22.73
7	Emamzade Radkan	امامزاده رادکان	59.02	36.84	32.58	32	Erye	اریه	58.71	36.46	30.58
8	Golmakan	گلمکان	59.13	36.44	3.45	33	Taghon	طاغون	58.68	36.42	23.49
9	Dolat Abad	دولت آباد	59.16	36.43	7.34	34	Zarande	زرنده	58.50	36.47	57.40
10	Zoshk	زشک	59.20	36.33	7.85	35	Hatite	حطیطه	57.21	36.56	35.69
11	Agh Darband	آق دربند	60.86	36.01	189.16	36	Jafar Mashhadi	جعفر مشهدی	59.42	35.18	101.43
12	Olang Asadi	اولنگ اسدی	59.65	36.25	87.42	37	Senobar	صنوبر	59.13	35.42	39.25
13	Ghale Bar Bar	قلعه ببر	57.19	37.75	140.04	38	Sang Divar	سنگ دیوار	59.54	37.17	40.15
14	Darband Samlghan	دربند سملقان	56.98	37.60	49.84	39	Barzo	بارزو	57.95	37.60	125.30
15	Esfarayen	اسفراین	57.50	37.08	70.25	40	Pol Khaton	پل خاتون	61.11	35.98	225.59
16	Hey Hey Ghochan	هی هی قوچان	58.55	37.11	51.26	41	Kalate Rahman Jadid	کلاته رحمان جدید	59.92	35.56	25.05
17	Chekne Olia	چکنه علیا	58.48	36.84	21.65	42	Shir Abad	شیرآباد	56.93	37.50	54.09
18	Andrakh	اندرخ	59.63	36.60	47.41	43	Baba Aman	باباامان	57.44	37.50	66.13
19	Sar Asiab Shandiz	سر آسیاب شاندیز	59.34	36.40	22.22	44	Aghmazar	آغمزار	56.91	37.70	231.25
20	Hesar	حصار	59.40	36.31	11.50	45	Brdo-Ghar Shishe	بردو-غارشیشه	60.08	35.44	28.56
21	Kertyan	کرتیان	59.51	36.17	36.14	46	Karde Bala	کارده بالا	59.67	36.66	56.68
22	Chah Chahe	چهچهه	60.33	36.64	73.61	47	Ghare Khan Bandi	قره خان بندی	57.51	37.51	99.04
23	Majmoe Dorod	مجموع دورود	59.05	36.16	16.42	48	Moshang	موشنگ	59.03	36.51	40.10
24	Roh Abad	روح آباد	58.86	36.06	53.90	49	Golestan	گلستان	59.40	36.31	13.31
25	Ghare Tikan	قره تیکان	60.17	36.83	89.22	50	Iraj Abad	ایرج آباد	58.17	35.30	21.75



شکل 2- نمودارهای نسبت گشتاورهای خطی برای 50 ایستگاه آب‌سنجی در منطقه مورد مطالعه

Figure 2- L moment ratio diagram for 50 stations in the study area

بمنظور بررسی لزوم استفاده از تحلیل خوشه‌ای و بررسی مقدماتی ایستگاه‌ها، نمودار ضریب چولگی خطی نسبت به ضریب تغییرات خطی و ضریب کشیدگی خطی (شکل 2) رسم شد. شکل 2 نشان داد که گشتاورهای خطی پراکنده بوده و حوضه‌های مورد مطالعه نمی‌توانند بعنوان یک منطقه‌ی همگن درنظر گرفته شوند. به این دلیل باید از تحلیل خوشه‌ای برای دستیابی به مناطق همگن هیدرولوژیکی استفاده شود.

به‌منظور انتخاب خصوصیات تاثیرگذار در تحلیل خوشه‌ای و تشکیل مناطق همگن هیدرولوژیکی، همبستگی بین میانگین سیلاب سالانه (به عنوان متغیر وابسته) و شش خصوصیت فیزیوگرافی (طول

آبراهه اصلی L، مساحت حوضه A، متوسط شیب حوضه S، میانگین ارتفاع حوضهEL، ضریب شکل FF و ضریب شکل C)، یک پارامتر هواشناسی (میانگین بارندگی سالانه در محل ایستگاه؛ P) و موقعیت جغرافیایی مرکز ثقل (X,Y)حوضه‌های مورد مطالعه (به عنوان متغیرهای مستقل)، محاسبه شد (شکل 3). همانطور که در شکل 3 مشاهده می‌شود ضریب همبستگی بین میانگین سیلاب سالانه و پارامتر شکل حوضه (FF) ناچیز (r^2=0/0005) می‌باشد. بنابراین این پارامتر از محاسبات حذف شد. طول و عرض جغرافیایی به این دلیل در تحلیل‌ها مورد استفاده قرار گرفتند تا تضمین کننده نزدیکی جغرافیایی حوضه‌های آبخیز واقع در مناطق شکل گرفته پس از تحلیل خوشه‌ای باشند (شباهت بیشتر در فواصل کمتر). هرچند میانگین بارندگی سالانه عاملی مهم در مطالعات برآورد سیلاب می‌باشد (9-)، اما بیشتر ایستگاه‌های باران سنجی در منطقه مورد مطالعه در خروجی حوضه‌ها قرار گرفته‌اند و نمی‌توانند معرف میانگین بارندگی سالانه در کل حوضه‌ها باشند و همبستگی ناچیزی (0/054 =r^2)با میانگین سیلاب سالانه دارد. بنابراین میانگین بارندگی سالانه نیز از ورودی تحلیل خوشه‌ای حذف می‌گردد. همچنین از شکل (3) آشکار است که بین برخی متغیرهای ورودی مانند مساحت حوضه آبخیز و طول آبراهه اصلی همبستگی بالایی وجود دارد (16-). در نتیجه پیش از تحلیل خوشه‌ای به منظور کاهش بعد متغیرهای ورودی با همبستگی بالا (15-) از تحلیل مولفه اصلی استفاده شد.

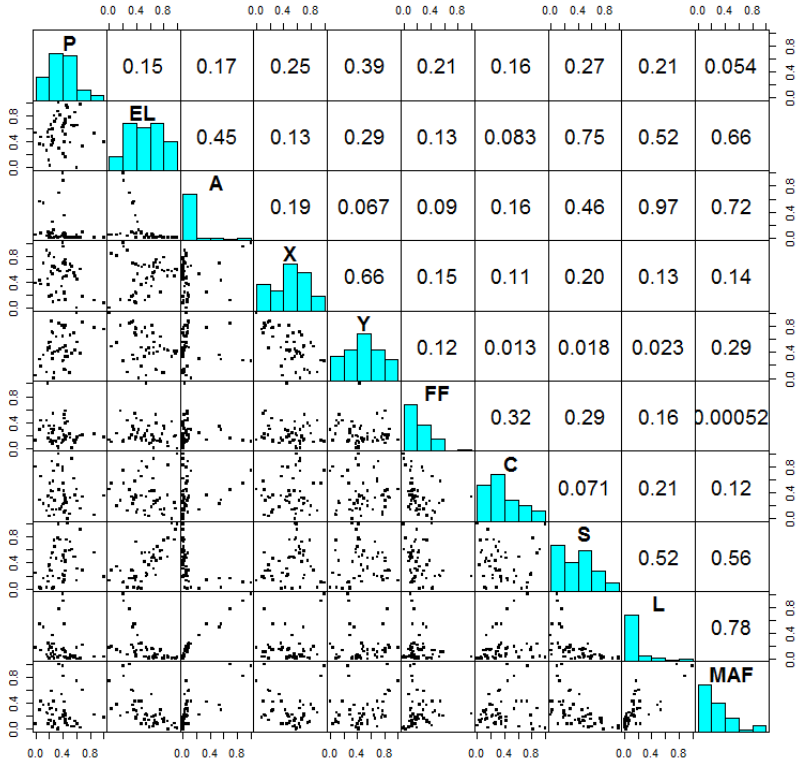
برای تعیین تعداد مولفه‌های اصلی ورودی الگوریتم خوشه‌بندی ارائه شده و کاهش بعد داده‌ها، درصد پراش هر یک از مولفه‌ها در نمودار سنگ ریزه‌ای[1] (شکل 4-الف) رسم شد. همانطور که مشاهده می‌شود، بیشترین پراش توسط دو مولفه‌ی اول استخراج شده و سهم پراش تجمعی استخراج شده توسط دو مولفه‌ی اول (28/1+42/4) 70/5 درصد است. بنابراین تعداد مولفه‌های اصلی برابر 2 انتخاب می‌گردد. به منظور بررسی سهم هر مولفه از بردار خصوصیات ورودی و تفسیر مولفه‌های استخراج شده، نمودار بای پلات[2] (شکل 4-ب) رسم گردید. همانطور که در شکل (4-ب) نشان داده شده است طول آبراهه اصلی (L) و مساحت حوضه (A) بیشترین سهم از مولفه‌ی اول را در جهت مثبت و متوسط شیب حوضه (S) و میانگین ارتفاع حوضه (EL) بیشترین سهم از مولفه‌ی اول در جهت منفی را دارا هستند. از طرف دیگر تاثیرگذارترین خصوصیات تشکیل دهنده مولفه دوم، طول و عرض جغرافیایی هستند.

بنابراین، 2 مولفه اصلی به عنوان ورودی SOM مورد استفاده قرار گرفت. بواسطه‌ی فرآیند یادگیری SOM، بردار وزن متناسب با چگالی احتمال داده‌ها (برای چگالی احتمال داده ها، شکل 3 را ببینید) تخمین زده شد. شکل (5) توزیع دو مولفه‌ی اصلی را در نقشهٔ آموزش یافتهٔ SOM نشان می‌دهد. مقدار هر مولفه در واحدهای خروجی SOM (هر یک از شش ضلعی‌های شکل 5) اهمیت آنرا در هر واحد[3] نشان می‌دهد. در شکل (5) رنگ تیره نشان دهنده‌ی مقایر زیاد هر مولفه‌ی اصلی و رنگ سفید کمترین مقدار را نشان می‌دهد. بعنوان نمونه مولفه‌ی اصلی اول در قسمت چپ-بالای شکل (5-الف) بیشترین مقادیر و در قسمت پایین-راست، کمترین مقادیر را داد. این موضوع قابلیت SOM در خوشه‌بندی و بصری نمودن نتایج را نشان می‌دهد. بعبارت دیگر، الگوریتم SOM مقادیر مشابه را در نقشه‌ی خروجی SOM کنار هم قرار می‌دهد.

بعد از آموزش SOM و به دست آمدن وزن‌های نهایی (ماتریس W' در معادله 2)، برای خوشه‌بندی واحدها توسط الگوریتم خوشه‌بندی سلسله مراتبی وارد، از بردار وزن‌های خروجی SOM استفاده شد. بر پایه‌ی نتایج دندوگرام (شکل 6)، ایستگاه‌ها بر روی نقشه SOM به 6 گروه تقسیم شدند (شکل 7). اعداد نوشته شده در هر شش ضلعی در شکل 7 تعداد ایستگاه اختصاص داده شده به هر واحد SOM را نشان می‌دهد. با توجه به شکل‌های (4-ب و 5 و 7) می‌توان دریافت که در خوشه‌ی دوم مولفه‌ی اول کم‌ترین مقادیر را دارا بوده و حوضه‌های این خوشه دارای مساحت‌های کم و طول آبراهه کوتاه می‌باشند. در ضمن ارتفاع و شیب زیاد تاثیرگذارترین خصوصیات فیزیوگرافی در ایجاد این خوشه می‌باشند. در خوشه چهارم مولفه دوم حداقل مقادیر را دارا بوده، شامل ایستگاه‌های شمال غربی منطقه مورد مطالعه بوده و با مقادیر میانگین دبی اوج بالای این ایستگاه‌ها تطابق دارد، هر چند که برای تایید این موضوع انجام آزمون همگنی نیاز است. خوشه‌های پنجم و ششم نیز مقادیر متوسط هر دو مولفه‌ی اول و دوم را در بر می‌گیرند. با توجه به بای-پلات (شکل 4-ب) و مقادیر شیب متوسط حوضه در ایستگاه‌های واقع در خوشه پنجم به‌نظر می‌رسد که تاثیرگذارترین عامل، شیب متوسط حوضه‌ها می‌باشد. مشخصه‌ی خوشه‌ی ششم نیز مساحت و طول آبراهه اصلی زیاد آنهاست که نتیجه آن قرار گرفتن ایستگاه‌هایی با میانگین دبی اوج لحظه‌ای زیاد در این خوشه است.

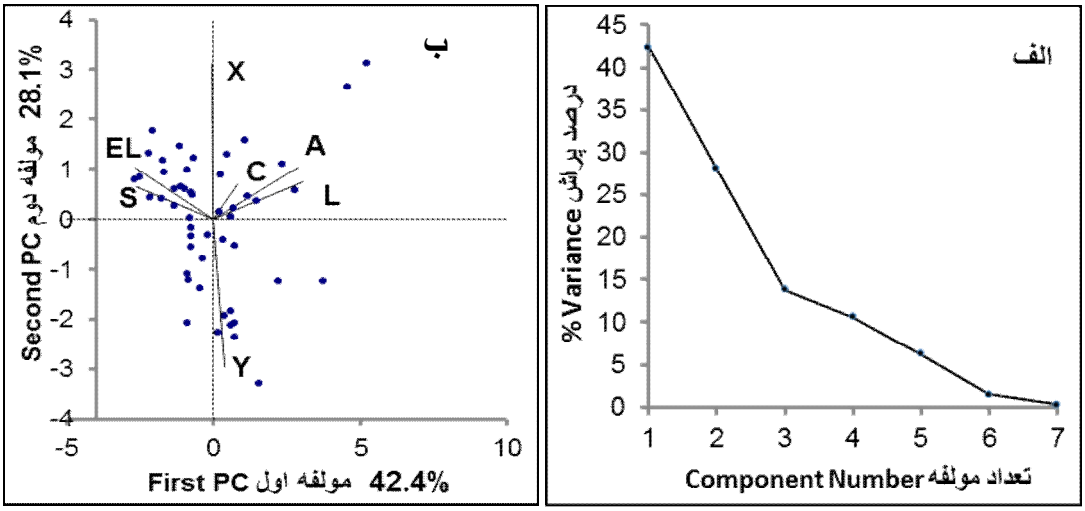
به منظور مقایسه‌ی تاثیر نوع ورودی بر نتایج الگوریتم‌های خوشه‌بندی، از خصوصیات حوضه‌های آبخیز نیز بعنوان ورودی استفاده شد. به‌دلیل تفاوت در پراش و بزرگی و اهمیت نسبی خصوصیات ورودی در تحلیل خوشه‌ای، خصوصیات ورودی تغییر مقیاس یافتند. سپس با انجام آزمون ناهمگنی براساس گشتاورهای خطی، همگنی مناطق به دست آمده با نتایج الگوریتم خوشه‌بندی با استفاده از ورودی مولفه‌های اصلی مقایسه شد (جدول 2).

1 -Scree plot
2 -Biplot
3 -Unit



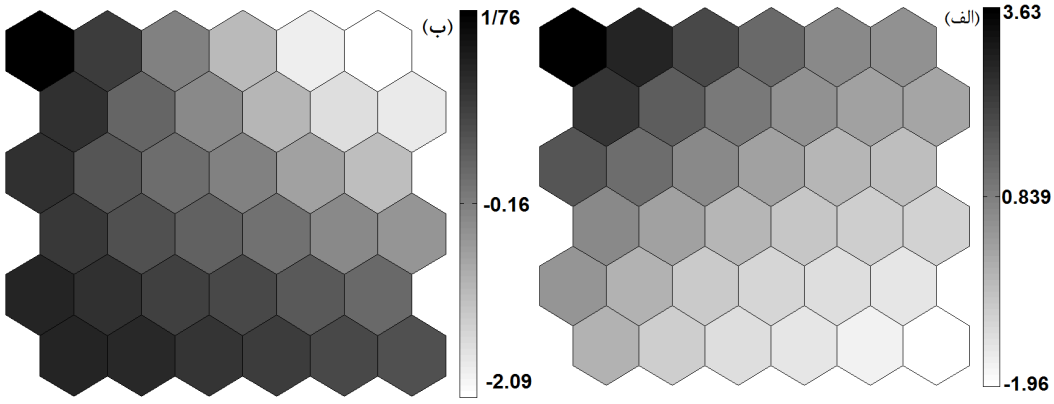
شکل 3- ماتریس ضرایب همبستگی، هیستوگرام و نمودار نقطه‌ای خصوصیات حوضه‌های آبخیز مورد مطالعه (طول آبراهه اصلی L، مساحت حوضه A، متوسط شیب حوضه S، میانگین ارتفاع حوضه EL، ضریب شکل FF و ضریب شکل C ، میانگین بارندگی سالانه در محل ایستگاه P و موقعیت جغرافیایی مرکز ثقل X,Y)

Figure 3- Correlation matrix, histogram and scatter plot of catchments attributes in the study area (Channel Length, Basin Area, Basin Mean Elevation, Shape Factor, Mean annual rainfall, Longitude, Latitude)



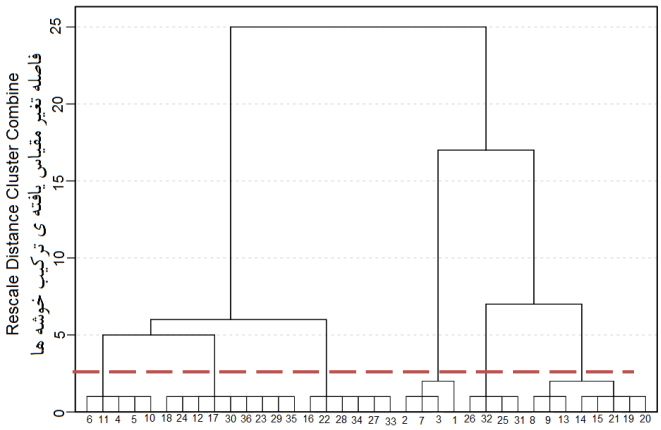
شکل 4- نمودار سنگ ریزه‌ای (الف) و بای پلات (ب) مولفه‌های اصلی (بردارها نشان دهنده‌ی بردار ویژه‌ی متغیرهای ورودی و نقاط نشانگر مقادیر مولفه‌های اصلی می‌باشند).

Figure 4-Scree plot and Biplot (arrows are shown specific vectors of input variable and points are shown principal components)



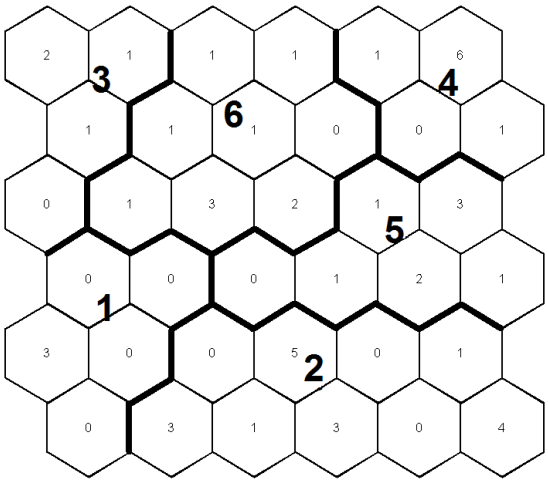
شکل 5- توزیع مولفه اول (الف) و مولفه دوم (ب) در نقشهٔ آموزش یافتهٔ SOM (رنگ تیره نشان دهنده‌ی مقایر زیاد هر مولفه و رنگ سفید کمترین مقدار را نشان می‌دهد.)

Figure 5- Gradient distributions of PCs in the SOM map trained with 2 PCs (Dark color represents a high value of PCs in their given scale bar, whereas white color is a low value)



شکل 6- دندوگرام حاصل از خوشه‌بندی لایه‌ی کوهونن با استفاده از الگوریتم وارد و فاصله اقلیدسی

Figure 6- Dendrogram are obtained by clustering of the Kohonen layer by Ward's algorithm and Euclidean distance



شکل 7- خوشه‌بندی نقشه خروجی SOM توسط الگوریتم خوشه‌بندی سلسله مراتبی وارد (اعداد پر رنگ نشان دهنده‌ی شماره خوشه و اعداد داخل هر یک از شش ضلعی‌ها نشان دهنده‌ی تعداد ایستگاه اختصاص یافته به آن واحد است.)

Figure 7- SOM output classified by Ward's hierarchical clustering (numbers in the hexagons represent the number of station assigned in each SOM unit and bold numbers shows cluster number)

همانطور که در جدول 2 مشاهده می‌شود، تعداد خوشه‌ها با استفاده از خصوصیات حوضه‌های آبخیز بعنوان ورودی، به 7 افزایش یافته و همگنی مناطق کاهش یافت. بهترین نتایج با الگوریتم ترکیبی نگاشت ویژگی خود سامان و سلسله مراتبی وارد (SOM+Ward) و با استفاده از دو مولفه‌ی اصلی بعنوان ورودی به دست آمد. این موضوع با نتایج دی پرینزو (2-) مبنی بر بهبود نتایج SOM با استفاده از مولفه‌های اصلی، همخوانی دارد. بنابراین روش SOM+Ward با ورودی PCA برای ادامه محاسبات انتخاب شد.

جدول 2- مقادیر آماره ناهمگنی (H_1) برای مناطق حاصل از الگوریتم های خوشه‌بندی با ورودی‌های مختلف

Table 2- Homogeneity measures for each region formed by clustering algorithms with different inputs

مناطق Regions			1	2	3	4	5	6	7
نوع الگوریتم خوشه بندی Clustering algorithm	SOM+Ward	PCA	-1.3	2.49**	-0.5	1.75*	0.72	0.75	-
		RAR	-1.3	2.14**	-0.8	2.36**	1.32*	1.17*	-0.2
	Ward	PCA	-0.1	3.16**	0.2	1.91*	0.05	0.46	-
		RAR	0.31	3.02**	-0.1	1.91*	0.84	-0.26	0.49

الگوریتم سلسله مراتبی. Hierarchical Algorithm: Ward الگوریتم ترکیبی; SOM+Ward: Hybrid Clusterin

.خصوصیات حوضه های آبخیز بعنوان ورودیRAR: Catchment Attributes as Input مولفه های اصلی بعنوان ورودی؛PCA: Principal Component Analysis as Input

تقریبا همگن Possibly Homogeneous ناهمگن؛ *Heterogeneous**

زمانی که مجموعه جامعی از متغیرهای تاثیرگذار در تحلیل فراوانی منطقه‌ای سیلاب وجود ندارد، معمولا مناطق شکل گرفته برای تحلیل فراوانی منطقه‌ای سیلاب همگن نبوده و نیاز به تعدیل برای بهبود همگنی آن وجود دارد (9-). این حقیقت در هیدرولوژی کاملاً شناخته شده است، از این رو هیدرولوژیست‌ها پیشنهاد می‌کنند با تعدیل، این مناطق به گروه‌هایی همگن اصلاح شوند. گزینه‌های پیشنهاد شده توسط هاسکینگ و والیس (9-) برای اصلاح مناطق شکل گرفته با تحلیل خوشه‌ای عبارتند از:

(1) حذف یک یا چند ایستگاه از مجموعه داده‌ها؛ (2) انتقال یک یا چند ایستگاه از یک منطقه به منطقه دیگر؛ (3) تفکیک منطقه به دو یا چند خوشهٔ جدید؛ (4) امکان مشارکت یک ایستگاه در دو یا چند منطقه؛ (5) انحلال مناطق و انتقال ایستگاه‌های آن به دیگر مناطق؛ (6) الحاق یک منطقه به منطقه (مناطق) دیگر؛ (7) الحاق دو یا چند منطقه و تعریف دوباره گروه‌ها؛ (8) کسب داده‌های بیشتر و تعریف دوباره گروه‌ها.

اگرچه هاسکینگ و والیس (8-) مقادیر بحرانی آماره ناهمنوائی ($D>3$) را برای شناسایی ایستگاه‌های غیرهمگن معرفی کردند، اما بهتر است با شناسایی تمام ایستگاه‌ها با مقادیر آماره ناهمنوائی بالا شناسایی شوند.

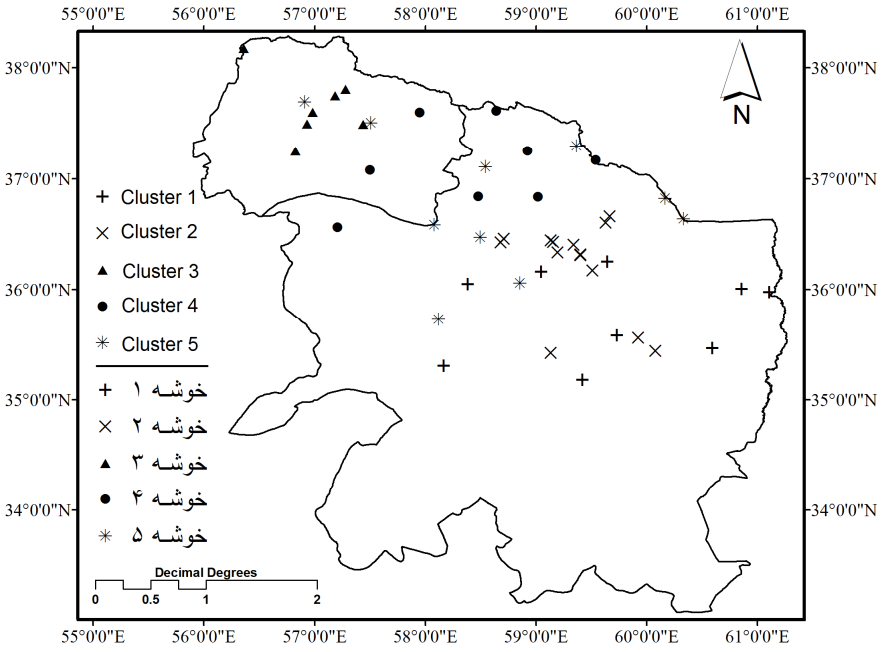
جدول 3- آماره ناهمگنی برای هر یک از مناطق

Table 3- Homogeneity measures for each region

Regions	قبل از تعدیل Before adjustment				بعد از تعدیل After adjustment			
مناطق	N	H_1	H_2	H_3	N	H_1	H_2	H_3
1	3	-1.28	0.17	0.76	9	0.81	0.66	0.02
2	17	2.49**	1.29*	0.12	14	0.99	-0.35	-1.08
3	4	-0.53	-0.84	-0.87	-	-	-	-
4	8	1.75*	1.14*	-0.06	7	1.28*	0.77	-0.28
5	8	0.72	0.61	0.09	8	0.72	0.61	0.09
6	10	0.75	0.33	-0.16	10	0.75	0.33	-0.16

شاخص همگنی Hi: Homogeneity measures, تعداد ایستگاه ها N: Number of stations

ناهمگن Heterogeneous ** ; تقریباً همگن Possibly Homogeneous *



شکل 8- مناطق همگن هیدرولوژیکی

Figure 8- Hydrologic homogeneous regions

سپس آماره ناهمگنی منطقه‌ای (H) با حذف یا تعویض ایستگاه‌ها تعدیل شود. در ادامه، ایستگاه‌های ناهمگون باغ عباس و مجموع دورود که به میزان معنی داری از کم شدن آماره ناهمگنی منطقه‌ای (H) در خوشه‌ی 2 جلوگیری می‌کنند، شناسایی شدند و بعد از بررسی، به خوشه‌ی 1 انتقال یافتند و بدلیل نیافتن نتیجه مطلوب، ایستگاه‌های درکش و موشنگ از خوشه‌ی 4 حذف گردیدند. با توجه به اینکه هدف از تحلیل منطقه‌ای دستیابی به بزرگترین مناطق همگن هیدرولوژیک است، با ترکیب خوشه های 1 و 3 (که تعداد ایستگاه‌های کمی را شامل می‌شدند؛ به ترتیب 3 و 4 ایستگاه) و بررسی همگنی آن، منطقه‌ای جدید تشکیل شد. مقادیر شاخص ناهمگنی قبل و بعد از تعدیل در جدول (3) ارائه شده است. باید این نکته مد نظر قرار گیرد که با توجه به همگنی نسبی منطقه‌ی 4، برای استفاده از تابع توزیع منطقه‌ای آن برای برآورد سیلاب احتیاط شود.

مناطق همگن هیدرولوژیکی پس از تعدیل، در شکل (8) نشان داده شده است. همانطور که در شکل (8) دیده می‌شود منطقه مورد مطالعه بصورت تقریبی از شمال به جنوب به پنج منطقه تقسیم شده است که با نتایج شامکوییان و همکاران (17-) همخوانی نشان می‌دهد.

نتیجه‌گیری کلی

در این مطالعه تاثیر نوع ورودی بر الگوریتم خوشه‌بندی دو مرحله‌ای SOM و سلسله مراتبی برای شناسایی مناطق همگن هیدرولوژیکی در تحلیل فراوانی منطقه‌ای سیلاب در حوضه‌های آبخیز استان‌های خراسان شمالی و رضوی در شمال شرق ایران به کار رفت. همگنی مناطق به دست آمده از الگوریتم خوشه‌بندی توسط آزمون ناهمگنی براساس گشتاورهای خطی مورد بررسی قرار گرفت. نتایج نشان داد که مناطق شکل گرفته با الگوریتم خوشه‌بندی ترکیبی SOM+Ward با ورودی PCA در مقایسه با ورودی‌های استاندارد شده نیاز کمتری به تعدیل بمنظور دستیابی به مناطق همگن دارد و موجب تفکیک منطقه مورد مطالعه به 5 ناحیه‌ی همگن هیدرولوژیک شد. همچنین الگوریتم خوشه‌بندی ترکیبی SOM+Ward تعداد خوشه‌های کمتر و همگن‌تری نسبت به الگوریتم خوشه‌بندی سلسله مراتبی وارد ایجاد کرد که نشان از کارایی بهتر این روش دارد.

هرچند استفاده از تحلیل مولفه‌ی اصلی موجب بهبود نتایج و کاهش ابعاد ورودی به الگوریتم خوشه‌بندی دو مرحله‌ای SOM و سلسله مراتبی شد، اما باعث دشوار شدن تفسیر نقشه‌ی خروجی SOM و عوامل موثر بر شکل‌گیری خوشه‌ها شد. بنابراین توصیه می‌شود در مواردی که همبستگی کمتری بین ورودی‌ها وجود دارد از این روش استفاده نشود و برای خوشه‌بندی تنها از ورودی‌های استاندار شده استفاده شود.

با توجه به همگنی نسبی منطقه‌ی 4، باید برای استفاده از تابع توزیع منطقه‌ای آن برای برآورد سیلاب احتیاط شود و برآوردهای منطقه‌ای سیلاب با برآوردهای ایستگاهی سیلاب در این منطقه مقایسه شود.

منابع

1- Chavoshi S., Azmin Sulaiman W.N., Saghafian B., Sulaiman MD. N.B. and Latifah A.M. 2012. Soft and hard clustering methods for delineation of hydrological homogeneous regions in the southern strip of the Caspian Sea Watershed, Journal of Flood Risk Management, 5 (4): 282–294.

2- Di Prinzio M., Castellarin A., and Toth E. 2011. Data-driven catchment classification: application to the pub problem. Hydrol, Earth System Sci, 15 (6): 1921–1935.

3- Farsadnia F., and Moghaddamnia A. 2014. Regional Flood Frequency Analysis by Self-Organizing Feature Maps and Fuzzy Clustering Approach, Iran-Water Resources Research, 9 (3): 24-36.

4- Farsadnia F., Rostami Kamrood M., Moghaddam Nia A., Modarres R., Bray M.T., Han D., and Sadatinejad, J. 2014. Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps, Journal of Hydrology, 509: 387–397.

5- Haykin S. 2003. Neural Networks: A Comprehensive Foundation. Fourth Indian Reprint, Pearson Education, Singapore.

6- Hosking J.R.M. 1986. The theory of probability weighted moments. Res. Rep. RC 12210, IBM Research Division, Yorktown Heights, NY.

7- Hosking J.R.M. 1991. Fortran routines for use with the method of L-moments, Version 2, Res. Rep. RC 17097, IBM Research Division, York Town Heights, NY 10598.

8- Hosking J.R.M. and Wallis J.R. 1993. Some statistics useful in regional frequency analysis, Water Resources Research, 29: 271–281.

9- Hosking J.R.M. and Wallis J.R. 1997. Regional frequency analysis: An approach based on L-moments, Cambridge University Press, New York.

10- Kohonen T. 1982. Self-organized formation of topologically correct feature maps, Biological Cybernetics, 43: 59–69.

11- Kohonen T. 2001. Self-Organizing Maps. Springer, Berlin, Germany.

12- Lampinen J. and Oja E. 1992. Clustering properties of hierarchical self-organizing maps, Journal of Mathematical Imaging and Vision, 2: 261–272.

13- Ley R., Casper M.C., Hellebrand H., and Merz R. 2011. Catchment classification by runoff behavior with self-organizing maps (SOM), Hydrology and Earth System Sciences, 15(9): 2947-2962.

14- Lin G. and Wang C. 2006. Performing cluster analysis and discrimination analysis of hydrological factors in one step, Advances in Water Resources, 29: 1573-1585.

15- Niromand H. 1999. Multivariate statistical analysis. Ferdowsi university of Mashhad Press, Mashhad.

16- Roa A.R., and Srinivas V.V. 2008. Regionalization of Watersheds (an approach based on cluster analysis), Speringer.

17- Shamkoueyan H., Ghahraman B., Davary K., and Sarmad M. 2009. Flood frequency analysis using linear moment and flood index method in Khorasan provinces, Journal of Water and Soil, 23(1): 31-43.

18- Vesanto J. and Alhoniemi R. 2000. Clustering of the self organizing map, IEEE Trans. Neural, Netw, 11 (3): 586-600.

19- Vesanto J., Himberg J., Alhoniemi E., and Parhankangas J. 2000. SOFM Toolbox for Matlab 5, Technical Report A57, Neural Networks Research Centre, Helsinki University of Technology, Helsinki, Finland.

20- Ward Jr. 1963. Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 58 (301): 236–244.

21- Wilppu R. 1997. The Visualisation Capability of Self Organizing Maps to Detect Deviation in Distribution Control. TUCS Technical Report No. 153, Turku Centre for Computer Science, Finland.

Journal of Water and Soil
Vol. 29, No. 5, Nov.-Dec. 2015, p. 1207-1218

نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد 29، شماره 5، آذر – دی 1394، ص. 1218-1207

Using Hierarchical Clustering in Order to Increase Efficiency of Self-Organizing Feature Map to Identify Hydrological Homogeneous Regions for Flood Estimation

F. Farsadnia[1*] - B. Ghahreman[2]

Received: 04-17-2014
Accepted: 08-19-2015

Introduction: Hydrologic homogeneous group identification is considered both fundamental and applied research in hydrology. Clustering methods are among conventional methods to assess the hydrological homogeneous regions. Recently, Self Organizing feature Map (SOM) method has been applied in some studies. However, the main problem of this method is the interpretation on the output map of this approach. Therefore, SOM is used as input to other clustering algorithms. The aim of this study is to apply a two-level Self-Organizing feature map and Ward hierarchical clustering method to determine the hydrologic homogenous regions in North and Razavi Khorasan provinces.

Materials and Methods: SOM approximates the probability density function of input data through an unsupervised learning algorithm, and is not only an effective method for clustering, but also for the visualization and abstraction of complex data. The algorithm has properties of neighborhood preservation and local resolution of the input space proportional to the data distribution. A SOM consists of two layers: an input layer formed by a set of nodes and an output layer formed by nodes arranged in a two-dimensional grid. In this study we used SOM for visualization and clustering of watersheds based on physiographical data in North and Razavi Khorasan provinces. In the next step, SOM weight vectors were used to classify the units by Ward's Agglomerative hierarchical clustering (Ward) methods. Ward's algorithm is a frequently used technique for regionalization studies in hydrology and climatology. It is based on the assumption that if two clusters are merged, the resulting loss of information, or change in the value of objective function, will depend only on the relationship between the two merged clusters and not on the relationships with any other clusters. After the formation of clusters by SOM and Ward, the most frequently applied tests of regional homogeneity based on the theory of L-moments are used to compare and modify the clusters which are formed by clustering algorithms and find the best clustering method to achieve hydrologically homogeneous regions. Two statistical measures are used to form a homogeneous region, (i) discordancy measure and (ii) heterogeneity measure. The discordancy measure, D_i, is used to find out unusual sites from the pooling group (i.e., the sites whose at-site sample L moments are markedly different from the other sites). Generally, any site with $D_i>3$ is considered as discordant. The homogeneity of the region is evaluated using homogeneity measures which are based on sample L-moments (LCv, LCs and LCk), respectively. The homogeneity measures are based on the simulation of 500 homogeneous regions with population parameters equal to the regional average sample l-moment ratios. The value of the H-statistic indicates that the region under consideration is acceptably homogeneous when $H<1$, possibly heterogeneous when $1\leq H<2$, and definitely heterogeneous when $H \geq 2$.

Results and Discussion Conclusions: At first by principal component analysis we reduced SOM input matrix dimension, then the SOM was used to form a two-dimensional features map. Then to determine homogeneous regions for flood frequency analysis, SOM output nodes were used as input into the Ward method. The regions identified by the clustering algorithms are, in general, not statistically homogeneous. Consequently, they have to be adjusted to improve their homogeneity. The sites that were flagged discordant by the discordancy measure were first identified. Secondly, the heterogeneity measures of the adjusted region were examined as they change with exclusion of each site from the region. Thirdly, the discordant site, whose exclusion reduces the heterogeneity measures of the region by a significant amount, was identified and removed. After adjustment of homogeneity regions by L-moment tests, five hydrologic homogeneous regions were identified. Finally adjusted regions are created by a two-level SOM and then the best regional distribution function and associated parameters are selected by the L-moment approach. The main results of this study are briefly mentioned:

The results showed that the combination of self-organizing maps and Ward hierarchical clustering by principal components as input is more effective than the hierarchical method, by principal components or

1,2 - PhD Student of Irrigation and Drainage and Professor, Department of Water Engineering, Faculty of Agriculture, Ferdowsi University of Mashhad
(*-Corresponding Author Email: farhadfarsadnia@stu.um.ac.ir)

standardized inputs to achieve hydrologic homogeneous regions.

SOM is a useful method to achieve homogeneous regions, because SOM has shown a high performance for visualization and abstraction of attributes, and displayed a distribution of each component. It is found that Ward's algorithm is an easy way to cluster SOM units because Ward's algorithm does not need to determine the optimum number of clusters before calculations.

Keywords: Principal Component Analysis, Regional flood frequency analysis, Hybrid clustering, linear moments