نشریه علمی-پژوهشی مهندسی و مدیریت آبخیز
جلد ۷، شماره ۱، ۱۳۹۴، صفحات ۲۷-۴۱

Journal of Watershed Engineering and Management
Volume 7, Issue 1, 2015, Pages 27-41`

منطقه‌بندی حوزه‌های آبخیز با استفاده از ترکیب نگاشت‌های خود سازمانده و الگوریتم فازی c-means

علی آهنی*[1] و سید سعید موسوی ندوشنی[2]

[1] کارشناس ارشد، دانشکده کشاورزی، دانشگاه صنعتی شاهرود و [2] استادیار، دانشکده مهندسی آب و محیط زیست، پردیس فنی و مهندسی شهید عباسپور دانشگاه شهید بهشتی

تاریخ دریافت: ۱۳۹۳/۰۲/۰۱ تاریخ پذیرش: ۱۳۹۳/۰۹/۲۹

چکیده

روش‌های تحلیل خوشه‌ای گونه‌ای از کارآمدترین روش‌های منطقه‌بندی حوزه‌های آبخیز به‌منظور انجام تحلیل فراوانی منطقه‌ای سیلاب هستند. منطقه‌بندی فازی، نوعی منطقه‌بندی است که در آن هر یک از سایت‌های مورد مطالعه می‌تواند هم‌زمان به بیش از یک منطقه اختصاص یابد. به‌منظور اجرای منطقه‌بندی فازی، گونه‌ای از الگوریتم‌های تحلیل خوشه‌ای به‌نام الگوریتم‌های خوشه‌بندی فازی به‌کار گرفته می‌شوند که مرسوم‌ترین آن‌ها الگوریتم خوشه‌بندی فازی c-means است. همچنین، نگاشت‌های مشخصه خود سازمانده، یکی از انواع شبکه‌های عصبی مصنوعی هستند که در زمینه‌های تشخیص الگو کاربرد قابل توجهی پیدا کرده‌اند. قابلیت این نوع از شبکه‌ها در تشخیص الگو و خوشه‌بندی داده‌ها با استفاده از ویژگی‌های آن‌ها موجب شده است که عده‌ای از محققان علم هیدرولوژی به آزمودن توانایی این نگاشت‌ها در زمینه منطقه‌بندی حوزه‌های آبخیز به‌منظور اجرای تحلیل فراوانی منطقه‌ای روی آورند. در مطالعه حاضر از نگاشت‌های خود سازمانده برای تعیین مراکز اولیه خوشه‌ها در الگوریتم فازی c-means به‌منظور منطقه‌بندی حوزه آبخیز سفیدرود بزرگ استفاده شده است. نتایج به‌دست آمده در این پژوهش نشان داد که این روش در حوزه مورد مطالعه، از نظر تشکیل مناطق همگن و ارائه برآوردهای مناسب در تحلیل فراوانی منطقه‌ای سیلاب با استفاده از الگوریتم گشتاورهای خطی، عملکرد قابل قبولی دارد. افزون بر این، مشاهده شد که استفاده از خوشه‌بندی فازی می‌تواند برآوردهای قابل اعتماد سیلاب را برای دوره‌های بازگشت طولانی‌تر امکان‌پذیر کند. همچنین، بر اساس شاخص‌های صحت خوشه‌بندی فازی به نظر می‌رسد که تعداد دو یا سه منطقه برای اجرای تحلیل فراوانی منطقه‌ای در این آبخیز مناسب است.

واژه‌های کلیدی: تحلیل خوشه‌ای، تحلیل فراوانی منطقه‌ای سیلاب، شبکه‌های عصبی مصنوعی، گشتاورهای خطی، منطقه‌بندی

مقدمه

نگاشت‌های مشخصه خود سازمانده[1]، گونه‌ای از شبکه‌های یادگیرنده رقابتی هستند (Kohonen، ۱۹۸۲). عملکرد این شبکه‌ها مبتنی بر یاد گیرندگی

[1] Self Organizing Feature Maps (SOFM)

* مسئول مکاتبات: ali.ahani66@yahoo.com

بدون نظارت است، یعنی هیچ خروجی هدفی برای دسته‌بندی داده‌های موجود مورد نیاز نیست. این شبکه‌ها می‌کوشند تا به‌وسیله نگاشت داده‌های موجود بر یک لایه مشخصه، ساختاری توپولوژیکی در داده‌های ورودی بیابند.

در سال‌های اخیر نگاشت‌های خود سازمانده کوهونن با لایه یک بعدی (که شبکه کوهونن خطی نیز خوانده می‌شود) در مطالعات منطقه‌بندی مورد استفاده قرار گرفته است. Hall و Minns (۱۹۹۹) استفاده از شبکه کوهونن را برای منطقه‌بندی با اعمال آن بر یک نمونه ۱۰۱تایی از سایت‌های مجهز به تجهیزات هیدرومتری در جنوب غربی انگلیس و ولز در انگلستان مورد آزمایش قرار دادند. Hall و همکاران (۲۰۰۲) SOFM یک بعدی را برای سه مجموعه از جنوب‌غربی انگلیس و ولز، ولز و اسکاتلند و جزایر جاوه و سوماترا در اندونزی به‌کار گرفتند. این مطالعات در مورد صحت-سنجی مناطق تشکیل شده با استفاده از شاخص‌های ناهمگنی، چیزی گزارش نکردند. Jingyi و Hall (۲۰۰۴) شاخص‌های ناهمگنی معرفی شده توسط Hosking و Wallis (۱۹۹۷) را برای ارزیابی همگنی مناطق مشخص شده به‌وسیله SOFM یک بعدی از ۸۶ ایستگاه هیدرومتری در استان‌های جیانگژی و فوجیان چین مورد استفاده قرار دادند.

تحلیل خوشه‌ای نام گونه‌ای از روش‌های آماری چند متغیری است که به‌منظور دسته‌بندی داده‌های موجود در گروه‌های مشابه مورد استفاده قرار می‌گیرند. نقاط معرف داده‌ها در یک خوشه باید تا حد امکان مشابه و در خوشه‌های مختلف حتی‌الامکان متفاوت باشند. بیشتر الگوریتم‌های خوشه‌بندی موجود می‌توانند در دو گروهخوشه‌بندی سخت[1] و فازی[2] طبقه‌بندی شوند (Rao و Srinivas، ۲۰۰۸).

در منطقه‌بندی به‌وسیله خوشه‌بندی سخت، یک سایت بر اساس اختصاص یا عدم اختصاص مطلق به یک خوشه دسته‌بندی می‌شود. اما در واقعیت، اکثر سایت‌ها دارای شباهت‌های جزئی به خوشه‌های متعدد هستند. در مقابل، خوشه‌بندی فازی به یک سایت اجازه داشتن عضویت‌های جزئی یا توزیعی در تمام خوشه‌ها را می-دهد. به‌عبارت دیگر، یک سایت می‌تواند هم‌زمان به بیش از یک خوشه اختصاص یابد. خوشه‌بندی فازی به یک بردار مشخصه[3] اجازه می‌دهد که هم‌زمان با یک درجه عضویت مشخص در فاصله [۰،۱] به تمام خوشه-ها اختصاص یابد (Rao و Srinivas، ۲۰۰۶b).

در هیدرولوژی، پژوهش‌های متعددی در زمینه منطقه‌بندی حوزه‌های آبخیز به‌صورت گروه‌هایی همگن از ایستگاه‌های هیدرومتری یا واکنش سیلابی مشابه با استفاده از خوشه‌بندی سخت انجام گرفته است (Tasker، ۱۹۸۲؛ Burn، ۱۹۸۹؛ Nathan و McMahon، ۱۹۹۰؛ Hosking و Wallis، ۱۹۹۷؛ Burn و Goel، ۲۰۰۰؛ Rao و Srinivas، ۲۰۰۶a). با این حال تلاش-های محدودی در زمینه بررسی توانایی خوشه‌بندی فازی برای منطقه‌بندی حوزه‌های آبخیز صورت گرفته است.

در میان روش‌های خوشه‌بندی فازی موجود، الگوریتم فازی c-means (FCM) مطرح شده توسط Bezdek (۱۹۸۱) ساده‌ترین و رایج‌ترین روش خوشه-بندی است که گونه توسعه یافته‌ای از الگوریتم خوشه-بندی سخت K-means در چارچوب فازی است.

در این زمینه، Hall و Minns (۱۹۹۹) برای منطقه-بندی ۱۰۱ سایت هیدرومتری از دو منطقه مشخص شده در گزارش مطالعات سیلاب انگلستان از الگوریتم فازی c-means استفاده کردند. Rao و Srinivas (۲۰۰۶b) منطقه‌بندی حوزه‌های آبخیز با استفاده از الگوریتم خوشه‌بندی فازی c-means را برای ایالت ایندیانا ارائه کردند. آن‌ها تأثیر این روش را در تشکیل مناطق همگن مثبت ارزیابی کردند. همچنین، منطقه-بندی ایالت ایندیانا با استفاده از شبکه کوهونن و ترکیب آن با الگوریتم خوشه‌بندی فازی توسط Srinivas و همکاران (۲۰۰۸) ارائه شد. آن‌ها از شاخص‌های صحت خوشه‌بندی و شاخص‌های ناهمگنی هاسکینگ و والیس برای قضاوت در مورد نتایج منطقه-بندی استفاده کردند. Sadri و Burn (۲۰۱۱) نیز از الگوریتم c-means برای منطقه‌بندی ۳۶ ایستگاه هیدرومتری در سه ایالت همجوار آلبرتا، ساسکاچوان و مانیتوبا به‌منظور تحلیل فراوانی منطقه‌ای خشکسالی

[1] Hard clustering
[2] Fuzzy clustering
[3] Feature vector

استفاده کردند. آن‌ها روشی مبتنی بر به‌کارگیری گشتاورهای خطی چند متغیره را برای بررسی ناجوری سایت‌های مورد مطالعه و همگنی مناطق تشکیل شده پیشنهاد کردند.

در مطالعه حاضر، هدف بررسی عملکرد روشی حاصل از ترکیب نگاشت‌های خود سازمانده کوهونن و الگوریتم خوشه‌بندی فازی c-means برای منطقه‌بندی حوزه آبخیز سفیدرود بزرگ به‌منظور اجرای تحلیل فراوانی منطقه‌ای سیلاب است.

مواد و روش‌ها

انتخاب ویژگی‌ها و آماده‌سازی بردارهای مشخصه: در گام نخست ویژگی‌های مؤثر بر پاسخ سیلابی حوضه در منطقه مورد مطالعه انتخاب می‌شوند. این ویژگی‌ها می‌توانند از میان مشخصات فیزیوگرافی، جغرافیایی، هواشناسی، زمین‌شناسی و مانند آن‌ها انتخاب شوند. از آماره‌های درون ایستگاهی سیلاب[1] (مانند میانگین، انحراف معیار و ضریب تغییرات داده‌های دبی سیلاب) نباید برای تشکیل مناطق، به‌منظور تحلیل فراوانی سیلاب استفاده شود، زیرا از آن‌ها در مراحل بعد به‌عنوان مبنای یک آزمون مستقل همگنی مناطق استفاده می‌شود (Rao و Srinivas، ۲۰۰۸).

داده‌های موجود برای هر ویژگی، به‌منظور خنثی کردن تفاوت‌ها در واریانس‌ها و بزرگی نسبی آن‌ها، تغییر مقیاس[2] می‌یابند. تغییر مقیاس ممکن است شامل شامل تبدیل مقادیر ویژگی‌ها با استفاده از تابع تبدیل مناسب و تقسیم مقادیر تبدیل یافته بر انحراف معیار و یا استانداردسازی آن‌ها باشد. هر بردار مشخصه از ویژگی‌های تغییر مقیاس یافته (بدون بعد) یک سایت تشکیل می‌شود.

در این پژوهش، از میان ویژگی‌های مؤثر بر پاسخ سیلابی در ایستگاه‌های مورد نظر، با توجه به آمار و اطلاعات در دسترس و نتایج مطالعات پیشین، طول و عرض جغرافیایی، مساحت زهکشی، ضریب رواناب، متوسط بارندگی سالانه و ارتفاع از سطح دریا برای استفاده در تحلیل خوشه‌ای انتخاب شدند. در میان ویژگی‌های منتخب در این مطالعه، تنها مساحت زهکشی ایستگاه‌ها با لگاریتم آن‌ها جایگزین شد.

نگاشت‌های مشخصه خود سازمانده: نگاشت‌های مشخصه خود سازمانده یکی از پرکاربردترین شبکه‌های عصبی مصنوعی برای تشخیص ساختار توپولوژیک در داده‌ها هستند. SOFM دارای یک لایه ورودی و یک لایه خروجی است که هر یک شامل تعدادی گره است. تعداد گره‌ها در لایه ورودی SOFM برابر ابعاد بردار مشخصه است. لایه خروجی که لایه رقابتی[3] یا لایه کوهونن نیز خوانده می‌شود، دارای m گره است که در یک شبکه که معمولاً یک یا دو بعدی است، سازماندهی شده است. مقدار m می‌تواند به‌صورت بیشینه تعداد مورد نظر برای تشکیل خوشه‌ها انتخاب شود (Fausett، ۱۹۹۴).

در حالت یاد گیرنده، SOFM بردارهای مشخصه را که نمایش‌دهنده خصوصیات حوزه آبخیز هستند، به گره‌های خروجی مختلف اختصاص می‌دهد. اگر گروه-بندی‌های طبیعی خوش‌تعریف به‌صورت ذاتی در مجموعه‌ی داده‌ها وجود داشته باشند، بردارهای مشخصه حول گره‌های خروجی که کاملاً از یکدیگر جدا شده‌اند، مجتمع می‌شوند. این امر نشان می‌دهد که SOFMها تعداد بهینه خوشه‌ها را به‌صورت خودکار تشخیص می‌دهند. این ویژگی SOFMها یک مزیت نسبت به روش‌های خوشه‌بندی سخت و فازی که تنها قادر به تقسیم‌بندی مجموعه داده‌های موجود در تعداد مشخصی از خوشه‌ها هستند، است. البته در غیاب الگو-های قابل تمیز آشکار در داده‌های موجود باشند، تفسیر خوشه‌ها از خروجی SOFM، صرف نظر از اندازه و ابعاد آن، به‌ندرت امکان‌پذیر است. با این حال، در چنین موقعیت‌هایی، SOFMها می‌توانند به‌عنوان روشی مفید در بین الگوریتم‌های خوشه‌بندی منظور شوند.

الگوریتم خوشه‌بندی فازی c-means: یکی از مرسوم‌ترین روش‌های خوشه‌بندی فازی است که مبتنی بر بهینه‌سازی عددی یک تابع هدف فازی برای تقسیم N سایت در یک ناحیه به c خوشه فازی است.

اگر y_k، معرف kامین بردار مشخصه که نماینده kامین حوضه در فضای مختصات n بعدی با محورهای مختصات $(y_1, \dots, y_n)$ مانند $y_k = [y_{1k}, \dots, y_{nk}] \in$

[1] At-site flood statistics
[2] Rescaled
[3] Competitive layer

R^n باشد که در آن y_{ik} مقدار ویژگی i در y_k را نشان می‌دهد. ویژگی‌های بردار مشخصه y_k به‌صورت رابطه (۱) تغییر مقیاس می‌یابند.

$$x_{ik} = \frac{w_i}{\sigma_i}[f(y_{ik})] \quad for \quad 1 \le i \le n, \quad 1 \le k \le N \tag{۱}$$

که در آن، x_{ik} نشان‌دهنده مقدار تغییر مقیاس یافته y_{ik}، w_i وزن اختصاص یافته به ویژگی i، σ_i معرف انحراف معیار ویژگی i، $f(.)$ نماینده تابع تبدیل و N نشان‌دهنده تعداد بردارهای مشخصه n بعدی است. تغییر مقیاس ویژگی‌ها به‌دلیل اختلافات موجود در واریانس، بزرگی نسبی و اهمیت آن‌ها ضروری است. مجموعه‌ای از N بردار مشخصه می‌تواند به‌صورت یک ماتریس $n \times N$ داده‌ها مانند X مطابق رابطه (۲) نمایش داده شود.

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nN} \end{bmatrix} \tag{۲}$$

افزون بر این، اگر $V = (v_1, \dots, v_c)$ نشان‌دهنده یک گروه cتایی از نمونه‌ها باشد، که هر یک از آن‌ها مرکز یکی از c خوشه را توصیف می‌کند. الگوریتم FCM ماتریس X را از طریق کمینه‌سازی تابع هدفی که در رابطه (۳) معرفی می‌شود، به c زیرمجموعه (یا خوشه) دارای هم‌پوشانی تقسیم می‌کند.

$$J(U, V: X) = \sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^{\mu} d^2(x_k, v_i) \tag{۳}$$

کمینه‌سازی تابع هدف، تابع روابط (۴) و (۵) است.

$$\sum_{i=1}^{c} u_{ik} = 1 \quad \forall k \in \{1, \dots, N\} \tag{۴}$$

$$0 < \sum_{k=1}^{N} u_{ik} < N \quad \forall i \in \{1, \dots, c\} \tag{۵}$$

که در آن‌ها، $u_{ik} \in [0,1]$ درجه عضویت kامین بردار مشخصه تغییر مقیاس یافته x_k در خوشه فازی iام را نشان می‌دهد، U ماتریس افراز فازی مطابق رابطه (۶) است که شامل عضویت هر یک از بردارهای مشخصه تغییر مقیاس یافته در هر خوشه فازی است.

$$U = \begin{bmatrix} u_{11} & \cdots & u_{1k} & \cdots & u_{1N} \\ \vdots & & \vdots & & \vdots \\ u_{i1} & \cdots & u_{ik} & \cdots & u_{iN} \\ \vdots & & \vdots & & \vdots \\ u_{c1} & \cdots & u_{ck} & \cdots & u_{cN} \end{bmatrix}_{c \times N} \tag{۶}$$

پارامتر $\mu \in [1, \infty]$ مربوط به توان وزنی برای هر عضویت فازی است، $d^2(x_k, v_i)$ فاصله بین kامین بردار مشخصه تغییر مقیاس یافته x_k تا مرکز خوشه iام است.

توان وزنی μ در رابطه (۳) میزان فازی بودن خوشه‌ها را مشخص می‌کند. این پارامتر مقدار عضویت مشترک بین خوشه‌های فازی را کنترل می‌کند. در FCM، $\mu = 1$ در تئوری به جواب K-means همگرا می‌شود. به‌عبارت دیگر، هنگامی که μ به یک میل کند، مقادیر عضویت به یک یا صفر میل می‌کنند. برای $\mu \to \infty$، بردارهای مشخصه به عضویت برابر در تمام c خوشه گرایش پیدا می‌کنند. بنابراین درجه عضویت kامین بردار مشخصه تغییر مقیاس یافته x_k در خوشه فازی iام، یعنی u_{ik} به $1/c$ میل می‌کند. مراحل الگوریتم عددی FCM به‌صورت خلاصه به این صورت هستند.

گام اول) ماتریس اولیه افراز فازی U (یا ماتریس مراکز خوشه‌های فازی V) با استفاده از یک تولید کننده اعداد تصادفی ایجاد می‌شود.

گام دوم) اگر الگوریتم FCM با ماتریس افراز فازی U آغاز شده است، عضویت‌های اولیه u_{ik}^{init} مربوط به x_k متعلق به خوشه i با استفاده از رابطه (۷) به‌منظور ارضای رابطه (۴) اصلاح می‌شود.

$$u_{ik} = \frac{u_{ik}^{init}}{\sum_{i=1}^{c} u_{ik}^{init}} \quad for\ 1 \le i \le c, 1 \le k \le N \tag{۷}$$

اگر الگوریتم FCM با ماتریس مراکز خوشه‌های فازی V (شامل c مرکز خوشه فازی $(v_1^{init}, \dots, v_c^{init})$ آغاز شده است، عضویت‌های u_{ik} مربوط به x_k متعلق به خوشه i با استفاده از رابطه (۸) با جایگزین کردن v_i^{init} با v_i تعیین می‌شود.

$$u_{ik} = \frac{\left(\frac{1}{d^2(x_k, v_i)}\right)^{1/(\mu-1)}}{\sum_{i=1}^{c}\left(\frac{1}{d^2(x_k, v_i)}\right)^{1/(\mu-1)}} \quad for \quad 1 \le i \le c, \quad 1 \le k \le N \tag{۸}$$

گام سوم) مرکز فازی v_i برای $i = 1, 2, \dots, c$ به‌وسیله رابطه (۹) محاسبه می‌شود.

$$v_i = \frac{\sum_{k=1}^{N}(u_{ik})^{\mu} x_k}{\sum_{i=1}^{c} u_{ik}^{init}} \tag{۹}$$

گام چهارم) عضویت فازی u_{ik} با استفاده از رابطه (۸) به‌هنگام می‌شود.

دو گام آخر تا زمانی که تغییر در مقادیر عضویت‌ها بین دو تکرار متوالی به قدر کافی کوچک شود، تکرار

می‌شوند (Bezdek، ۱۹۸۱). در این نقطه، روش سنتی تحلیل خوشه‌ای فازی، غیرفازی کردن ماتریس افراز فازی U را برای اختصاص نهایی بردارهای مشخصه به خوشه‌ها توصیه می‌کند. در این روش هر سایت تنها به خوشه‌ای اختصاص می‌یابد که بالاترین میزان عضویت سایت مربوط به آن خوشه است. خوشه‌های حاصل از این روش، در نهایت از نظر نحوه عضویت سایت‌ها در آن‌ها به‌صورت خوشه‌های سخت هستند و مشابهت بسیاری با خوشه‌های به‌دست آمده از الگوریتم K-means دارند. در مقابل، اختصاص نهایی فازی وجود دارد که در آن، یک خوشه فازی با در برگرفتن سایت‌هایی که عضویت آن‌ها در آن خوشه از مقدار آستانه معینی تجاوز می‌کند، شکل می‌گیرد. عموماً انتخاب یک مقدار آستانه برای تشکیل خوشه‌های فازی تا حدی اختیاری و ذهنی است. در فازی‌ترین افراز، عضویت‌های یک بردار مشخصه در تمام خوشه‌ها برابر $1/c$ خواهد بود. از این‌رو مقدار $1/c$ به‌عنوان یک انتخاب قابل قبول برای عضویت فازی آستانه شناخته می‌شود. (Rao و Srinivas، ۲۰۰۸)

شاخص‌های صحت خوشه‌بندی و تعیین تعداد بهینه مناطق: ارزیابی صحت روشی برای ارزیابی و مقایسه خوشه‌های به‌دست آمده از یک الگوریتم خوشه‌بندی برای انتخاب‌های متفاوت پارامترها یا مقایسه خوشه‌های حاصل از الگوریتم‌های خوشه‌بندی مختلف است. همچنین، خوشه‌های تشکیل شده با استفاده از شاخص‌های صحت خوشه‌بندی[1] برای تعیین تعداد بهینه مناطق تفسیر می‌شوند. در مطالعه حاضر از میان شاخص‌های سنجش صحت خوشه‌بندی سخت، شاخص عرض silhouette به‌دلیل عملکرد قابل قبول در مطالعات پیشین انتخاب شده است. این شاخص برای iامین بردار مشخصه در خوشه با $s(i)$ نشان داده شده و مطابق رابطه (۱۰) تعریف می‌شود.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \qquad (۱۰)$$

که در آن، $a(i)$ فاصله متوسط بردار مشخصه iام نسبت به تمام بردارهای مشخصه دیگر موجود در خوشه و $b(i)$ کمینه فاصله متوسط بردار مشخصه‌ی iام نسبت به تمام بردارهای مشخصه خوشه دیگر است. بر اساس رابطه (۱۱)، خواهیم داشت $-1 \leq s(i) \leq 1$.

اگر مقدار $s(i)$ نزدیک به یک باشد، می‌توان این طور نتیجه‌گیری کرد که بردار مشخصه iام در خوشه‌ای مناسب جای گرفته است. از سوی دیگر چنان‌چه مقدار $s(i)$ به ۱- نزدیک باشد، می‌توان این گونه استنتاج کرد که iامین بردار مشخصه به خوشه مناسبی تعلق نیافته است. از میانگین تمام مقادیر $s(i)$ برای قضاوت کلی در مورد خوشه‌بندی انجام گرفته استفاده می‌شود (Rao و Srinivas، ۲۰۰۸).

در تحلیل خوشه‌ای فازی، ارزیابی صحت با استفاده از شاخص‌های صحت خوشه‌بندی فازی اجرا می‌شود که متفاوت از تابع هدفی که با استفاده از الگوریتم خوشه‌بندی فازی بهینه می‌شود، در نظر گرفته می‌شوند.

معیارهایی که در ارزیابی و انتخاب خوشه در نظر گرفته می‌شوند، عبارتند از فشردگی و جدایی خوشه‌ها. منظور از فشردگی این است که اعضای یک خوشه حتی‌الامکان به یکدیگر نزدیک باشند و مقصود از جدایی فاصله داشتن هر چه بیشتر خوشه‌ها از یکدیگر و گسترش آن‌ها در فضا است. در ادامه برخی شاخص‌های مختلف صحت خوشه‌بندی که در ادبیات پژوهشی رایج هستند، به‌طور خلاصه توصیف می‌شوند.

- ضریب افراز[2]: این شاخص به‌منظور سنجش مقدار هم‌پوشانی بین خوشه‌ها به‌صورت رابطه (۱۱) طرح شد (Bezdek، ۱۹۸۱).

$$V_{PC}(U) = \frac{1}{N}\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^2 \qquad (۱۱)$$

- انتروپی افراز[3]: انتروپی افراز برای یک افراز c خوشه‌ای فازی به شکل رابطه (۱۲) تعریف می‌شود.

$$V_{PE}(U) = -\frac{1}{N}\left[\sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}\log_a(u_{ik})\right] \qquad (۱۲)$$

که در آن، $a \in (1, \infty)$ است (Bezdek، ۱۹۸۱).

افراز بهینه، متناظر با مقدار بیشینه V_{PC} یا مقدار کمینه V_{PE} است که بر کمینه هم‌پوشانی بین خوشه‌ها دلالت می‌کند. دامنه تغییرات V_{PC} در محدوده $[1/c\,, 1]$ است، در حالی‌که V_{PE} در بازه $[0, \log_a(c)]$ جای می‌گیرد. برای یک افراز سخت، V_{PC} برابر یک است، در حالی‌که V_{PE} مساوی صفر است.

[1] Cluster Validity Measures

[2] Partition Coefficient (PC)

[3] Partition Entropy (PE)

هرگاه عضویت هر یک از بردارهای مشخصه در تمام خوشه‌ها برابر باشد ($u_{ik} = 1/c \quad \forall i,k$) که در فازی-ترین افراز c روی می‌دهد، مقدار V_{PC} برابر $1/c$ و V_{PE} مساوی $\log_a(c)$ خواهد شد. نقطه ضعف V_{PC} و V_{PE} این است که با هیچ یک از مشخصات داده‌ها ارتباط مستقیمی ندارند.

- شاخص صحت ژی-بنی[1]: شاخص پیشنهادی توسط Xie و Beni (۱۹۹۱) تابعی از مجموعه داده‌ها و مراکز خوشه‌ها است که به‌صورت رابطه (۱۳) تعریف می‌شود.

$$V_{XB}(U,V:X) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^2\|v_i - x_k\|^2}{N \min_{i\neq k}\|v_i - v_k\|^2} \qquad (۱۳)$$

که در آن، عبارت صورت کسر مجموع مجذورات انحراف هر یک از بردارهای مشخصه از مرکز هر خوشه فازی است. اندازه این عبارت با افزایش در فشردگی خوشه‌ها کاهش می‌یابد. عبارت مخرج کسر که کمینه جدایی بین مراکز خوشه‌ها را می‌سنجد، برای خوشه-هایی که کاملاً از یکدیگر جدا هستند، مقدار بزرگ‌تری خواهد داشت. کمینه مقدار V_{XB} مبین یک افراز خوب است که متناظر با خوشه‌هایی فشرده و کاملاً جدا از هم است (Xie و Beni، ۱۹۹۱). Xie و Beni (۱۹۹۱) جایگزینی $(u_{ik})^{\mu}$ با $(u_{ik})^2$ را در رابطه (۱۳) هنگامی که در رابطه (۳) $\mu \neq 2$، پیشنهاد کردند. Bezdek (۱۹۸۷) به این مورد به‌عنوان شاخص ژی-بنی توسعه یافته الگوریتم فازی c-means ($V_{XB,m}$) که به‌صورت رابطه (۱۴) حاصل می‌شود، اشاره کرد.

$$V_{XB,m}(U,V:X) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^{\mu}\|v_i - x_k\|^2}{N \min_{i\neq k}\|v_i - v_k\|^2} \qquad (۱۴)$$

مقدار V_{XB} به‌صورت یکنواخت با افزایش تعداد خوشه‌ها، کاهش می‌یابد. برای برطرف کردن این مشکل، Kwon (۱۹۹۸) یک شاخص جدید صحت خوشه‌بندی V_K مطابق رابطه (۱۵) ارائه کرد که در صورت کسر دارای یک جمله اضافه است.

$$V_K = \frac{\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^2\|v_i - x_k\|^2 + \frac{1}{c}\sum_{i=1}^{c}\|v_i - \bar{v}\|^2}{\min_{i\neq k}\|v_i - v_k\|^2} \qquad (۱۵)$$

برآورد سیلاب در تحلیل فراوانی منطقه‌ای تا زمانی قابل اعتماد است که تعداد سال‌های آمار موجود در ایستگاه‌های یک منطقه بزرگ‌تر یا مساوی پنج برابر دوره بازگشت مورد نظر برای برآورد سیلاب باشد (Rao و Srinivas، ۲۰۰۸). از این‌رو تعداد سایت‌های موجود در هر منطقه و سال‌های آماری موجود برای هر یک از آن‌ها عاملی مهم در انتخاب تعداد خوشه‌ها است.

همچنین، اغلب از شاخص‌های صحت خوشه‌بندی برای تعیین تعداد بهینه خوشه‌ها در یک مجموعه داده-ها استفاده می‌شود. در مورد شاخص‌های صحت خوشه-بندی فازی، این فرایند نیازمند آن است که کلیه پارامترهای الگوریتم خوشه‌بندی به‌جز تعداد خوشه‌ها c، ثابت نگاه داشته شوند. سپس پارامتر c از یک تا یک مقدار بیشینه در افزایش‌های یک واحدی تغییر می‌کند. در این پژوهش، با توجه به تعداد سایت‌ها و سال‌های آماری موجود در ناحیه مورد مطالعه، عملکرد الگوریتم-های خوشه‌بندی مورد مطالعه، در تشکیل تعداد دو تا پنج منطقه مورد بررسی قرار گرفته است.

تحلیل فراوانی منطقه‌ای سیلاب با استفاده از الگوریتم گشتاورهای خطی: پس از اجرای فرایند منطقه‌بندی، فرایند تحلیل فراوانی منطقه‌ای سیلاب با استفاده از روش معرفی شده توسط Hosking و Wallis (۱۹۹۷) که مبتنی بر استفاده از گشتاورهای خطی است، اجرا می‌شود.

در مرحله نخست این فرایند که غربال کردن داده‌ها است، هدف تشخیص سایت‌هایی است که به‌صورت فاحشی با گروهی از سایت‌ها به‌عنوان یک مجموعه، ناجور هستند. بدین منظور شاخص ناجوری D_i برای تمامی سایت‌های مورد مطالعه محاسبه می‌شود و چنان‌چه مقدار D_i یک سایت از مقدار بحرانی که برای بیش از ۱۴ سایت برابر سه است تجاوز کند، آن سایت ناجور در نظر گرفته می‌شود و از ادامه فرایند تحلیل فراوانی حذف می‌شود (Hosking و Wallis، ۱۹۹۷).

در گام بعدی، همگنی مناطق حاصل از عملیات خوشه‌بندی، با استفاده از شاخص‌های ناهمگنی H مورد ارزیابی قرار می‌گیرد. سه شاخص ناهمگنی H_1، H_2 و H_3 بر اساس گشتاورهای خطی تعریف می‌شوند. در هر منطقه اگر $H < 1$ باشد، آن منطقه همگن، اگر $1 \leq H < 2$ باشد، منطقه نسبتاً ناهمگن و اگر $H \geq 2$ باشد، منطقه کاملاً ناهمگن است. از آن‌جا که در منطقه‌بندی با استفاده از روش‌های تحلیل خوشه‌ای،

[1] Xie-Beni Validity Measure

شاخص H_1 اهمیت و قابلیت اعتماد بیشتری دارد، لذا در بررسی همگنی مناطق حاصل، توجه و تأکید بیشتری به این شاخص معطوف می‌شود. مناطقی که بیشتر به حالت ناهمگنی نزدیک هستند، به‌منظور بهبود وضعیت همگنی‌شان می‌توانند با حذف یا جابه‌جایی محدود یک یا چند سایت که تأثیر بیشتری در افزایش ناهمگنی دارند و یا باز تعریف مناطق در صورت نیاز، اصلاح شوند (Hosking و Wallis، ۱۹۹۷).

در ادامه تحلیل فراوانی منطقه‌ای، باید یک توزیع فراوانی واحد که برآوردهای چندک دقیقی را برای هر سایت حاصل می‌کند، به‌عنوان توزیع منطقه‌ای بر داده-های به‌دست آمده از سایت‌های متعدد برازش داده شود. به‌منظور شناسایی چنین توزیعی می‌توان از یک شاخص نکویی برازش استفاده کرد. Z^{DIST} شاخصی است که چنین کارکردی دارد ($DIST$ معرف نوع توزیع است). در برازش یک توزیع بر یک منطقه، اگر $|Z^{DIST}| \leq 1.64$ باشد، می‌توانیم برازش را مناسب بدانیم. هر چه مقدار $|Z^{DIST}|$ به صفر نزدیک‌تر باشد، برازش بهتر خواهد بود. مجموعه‌ای از توزیع‌های سه پارامتری منطقی محتمل شامل لجستیک تعمیم یافته، مقادیر حدی تعمیم یافته، پارتوی تعمیم یافته، نرمال تعمیم یافته و پیرسون تیپ III بر نسبت‌های گشتاور خطی متوسط منطقه‌ای برازش داده می‌شود. توزیع منطقه‌ای با اعمال ضریبی مشخص که می‌تواند دبی متوسط سیلاب هر سایت باشد، تبدیل به توزیع ویژه آن سایت می‌شود (Hosking و Wallis، ۱۹۹۷).

معرفی حوزه آبخیز سفیدرود بزرگ: حوزه آبخیز سفیدرود بزرگ به‌خاطر وجود اقلیم‌های متفاوت و منابع غنی آب و خاک از اهمیت خاصی برخوردار است. مساحت این حوزه آبخیز ۶۳۹۴۵ کیلومتر مربع می‌باشد و در محل تلاقی رشته کوه‌های البرز، زاگرس و مرکزی واقع شده است. این حوزه آبخیز از دو شاخه رودخانه‌ای اصلی به‌نام قزل‌اوزن و شاهرود تشکیل شده است که در محل سد سفیدرود بهم می‌پیوندند و رودخانه سفیدرود را تشکیل می‌دهند. در این پژوهش تعداد ۳۹ ایستگاه هیدرومتری در این حوزه آبخیز که دارای اطلاعات مورد نیاز برای اجرای عملیات خوشه‌بندی و تحلیل فراوانی منطقه‌ای سیلاب بودند، انتخاب شدند.

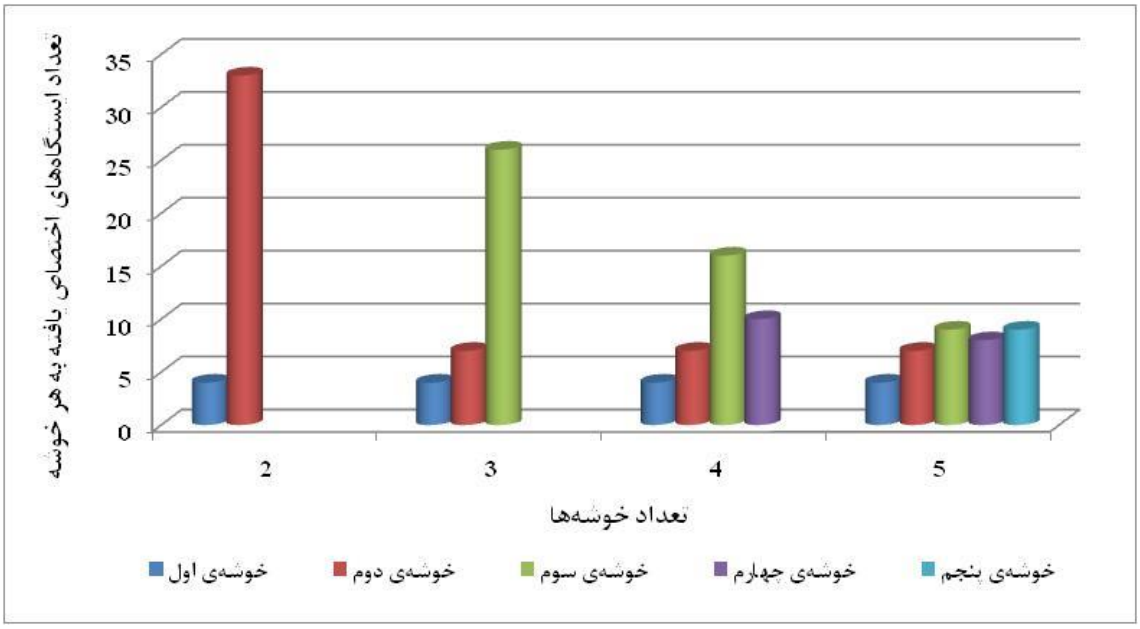
زبان برنامه نویسی *R*: در این مطالعه، برای اجرای محاسبات و عملیات مربوط به خوشه‌بندی و تحلیل فراوانی منطقه‌ای از محیط نرم‌افزاری زبان برنامه نویسی آماری R (نسخه ۳,۰,۲) و بسته‌های محاسباتی kohonen_2.0.11، cluster_1.14.4، e1071_1.6-1 و lmomRFA_2.5 استفاده شده است.

نتایج و بحث

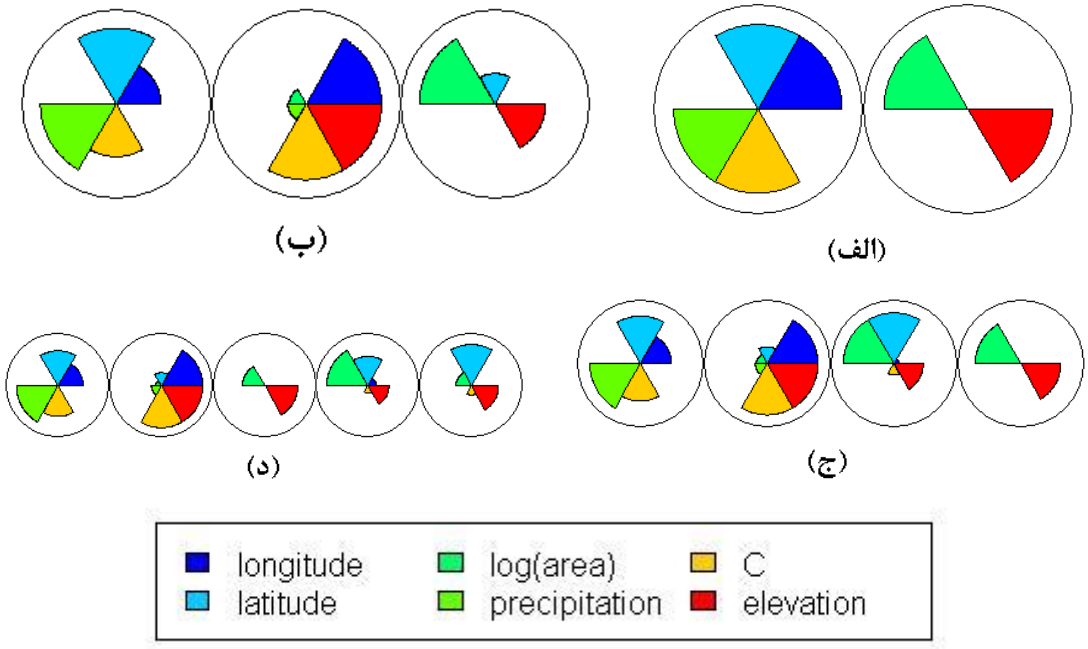
با محاسبه شاخص ناجوری برای ایستگاه‌های هیدرومتری مورد مطالعه، از میان ۳۹ سایت، دو سایت که بر اساس شاخص D، ناجور تشخیص داده شدند از ادامه فرایند تحلیل کنار گذاشته شدند.

در گام بعد، خوشه‌بندی ۳۷ ایستگاه هیدرومتری مورد نظر با استفاده از شبکه‌های کوهونن با دو تا پنج گره در لایه خروجی برای حالت‌های دو تا پنج خوشه‌ای اجرا شد. در شکل ۱، تعداد ایستگاه‌های اختصاص یافته به هر یک از گره‌های لایه خروجی در حالت‌های دو تا پنج خوشه‌ای مشاهده می‌شود.

همچنین، شکل ۲، نشان‌دهنده وزن ویژگی‌های تغییر مقیاس یافته ایستگاه‌های اختصاص یافته به هر گره یا خوشه در حالت‌های دو تا پنج خوشه‌ای تشکیل شده به‌وسیله نگاشت‌های خود سازمانده است. هر قطاع در هر دایره که با رنگی خاص مشخص شده است، معرف وزن مقدار تغییر مقیاس یافته یکی از ویژگی‌های مورد استفاده در خوشه‌بندی در آن خوشه است. برای نمونه، در شکل ۲-الف که مربوط به حالت دو خوشه‌ای است، خوشه‌ی اول (سمت چپ) دارای وزن مثبت مقادیر استاندارد شده مثبت و بزرگ برای طول و عرض جغرافیایی، متوسط بارندگی سالانه و ضریب رواناب است و در مقابل خوشه دوم (سمت راست) دارای وزن مثبت و بزرگ برای لگاریتم مساحت زهکشی و ارتفاع از سطح دریا است. وزن ویژگی‌ها در سایر حالت‌های خوشه‌بندی نیز به همین صورت قابل تفسیر است.



شکل ۱- تعداد ایستگاه‌های متعلق به هر یک از خوشه‌های تشکیل شده به‌وسیله نگاشت‌های خود سازمانده در حالت‌های دو تا پنج خوشه‌ای



شکل ۲- وزن ویژگی‌های تغییر مقیاس یافته گره‌های لایه‌های خروجی در حالت‌های دو تا پنج خوشه‌ای؛ شکل‌های (الف)، (ب)، (ج) و (د) به‌ترتیب معرف حالت‌های دو، سه، چهار و پنج خوشه‌ای هستند.

در ادامه با محاسبه شاخص صحت خوشه‌بندی عرض silhouette و مقدار میانگین آن برای حالت‌های دو تا پنج خوشه‌ای مشخص شد که مقدار این شاخص در حالت سه خوشه‌ای بیش از سایر حالت‌ها است و لذا حالت سه منطقه‌ای را می‌توان، حالت بهینه منطقه‌بندی با استفاده از نگاشت‌های خودسازمانده از نظر وضعیت خوشه‌بندی دانست. در جدول ۱ مقادیر میانگین محاسبه شده شاخص عرض silhouette برای تعداد خوشه‌های مختلف مشاهده می‌شود.

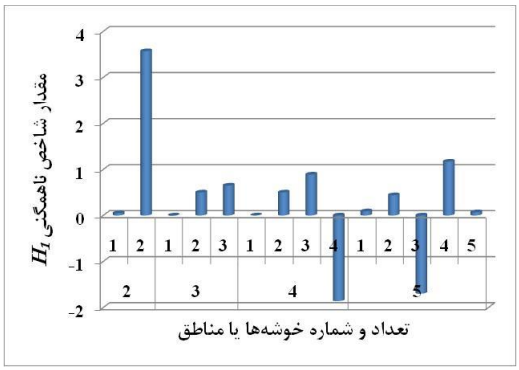
در گام بعد، مقدار شاخص ناهمگنی H_1 برای تمامی مناطق تشکیل شده در حالت‌های دو تا پنج منطقه‌ای حاصل از استفاده از نگاشت‌های خود-سازمانده محاسبه شد که نتایج به‌دست آمده، در شکل ۳ ارائه شده است. همان‌طور که در شکل مشاهده می-

شود، در حالت‌های سه و چهار منطقه‌ای، تمام مناطق تشکیل شده همگن هستند، اما در هر یک از حالت-های دو و پنج منطقه‌ای برای یک منطقه، شاخص ناهمگنی H_1 بیشتر از یک بوده و در نتیجه آن منطقه ناهمگن است.

پس از این مراحل، مراکز خوشه‌های تشکیل شده به‌وسیله نگاشت‌های خود سازمانده، به‌عنوان مراکز اولیه خوشه‌ها در خوشه‌بندی با استفاده از الگوریتم c-means تعیین شدند.

جدول ۱- مقادیر میانگین شاخص صحت خوشه‌بندی عرض silhouette برای حالت‌های دو تا پنج خوشه‌ای

تعداد خوشه‌ها	۲	۳	۴	۵
میانگین عرض silhouette	۰/۴۴	۰/۴۹	۰/۴۲	۰/۴۱



شکل ۳- مقادیر شاخص ناهمگنی H_1 برای مناطق تشکیل شده با استفاده از نگاشت‌های خود سازمانده برای ۳۷ ایستگاه هیدرومتری مورد مطالعه در حوزه آبخیز سفیدرود بزرگ

برای مقادیر مختلف عامل فازی‌کننده μ در فاصله [1.1,2.5] که در هر گام به اندازه ۰/۱ تغییر نمود، الگوریتم c-means با ۱۰۰۰ مرتبه تکرار برای هر یک از حالت‌های دو تا پنج خوشه‌ای اجرا شد.

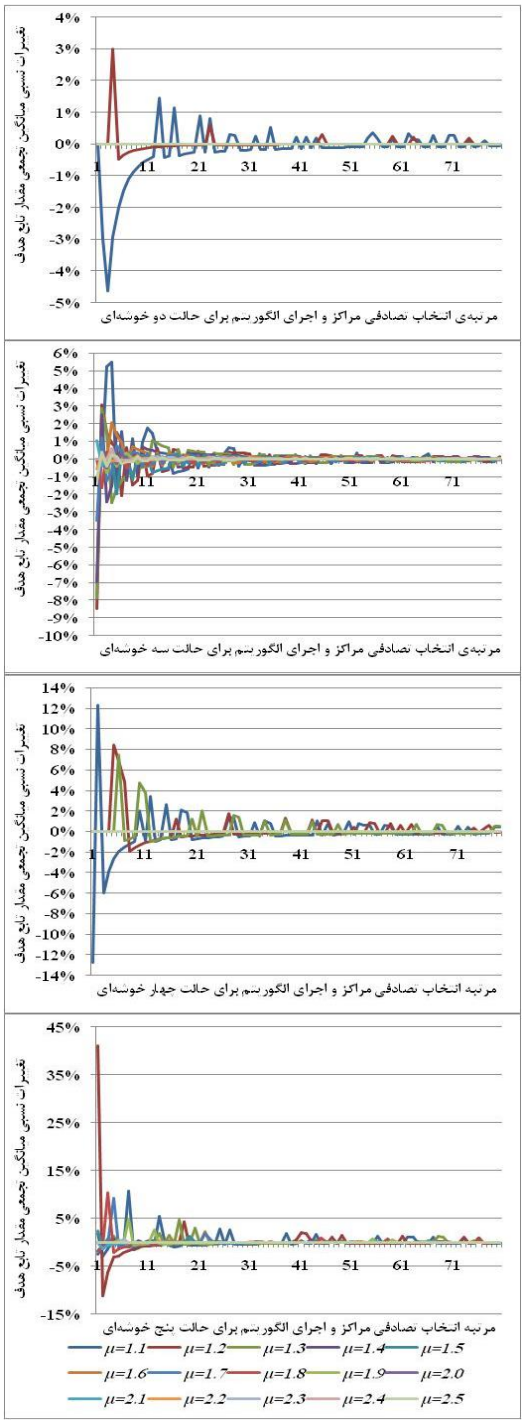
برای این‌که تأثیر انتخاب مراکز خوشه‌های حاصل از به‌کارگیری نگاشت‌های خود سازمانده به‌عنوان مراکز اولیه خوشه‌ها در الگوریتم فازی c-means مشخص شود، توانایی کمینه‌سازی تابع هدف در این شرایط با وضعیت متوسطی از اجرای الگوریتم c-means با مراکز خوشه‌های اولیه تصادفی مورد مقایسه قرار گرفت. بدین منظور با ۱۰۰ مرتبه انتخاب تصادفی مراکز اولیه خوشه‌ها، مقدار تابع هدف برای اجرای الگوریتم c-means با ۱۰۰۰ مرتبه تکرار و برای مقادیر مختلف عامل فازی‌کننده μ در فاصله [1.1,2.5] در هر یک از حالت‌های دو تا پنج خوشه‌ای محاسبه شد. سپس میزان تغییرات نسبی مقدار میانگین تجمعی تابع هدف در دفعات متوالی اجرای الگوریتم c-means با استفاده از مراکز اولیه تصادفی خوشه‌ها مشخص شد. همان‌طور که در شکل ۴ مشاهده می‌شود، بیشینه پس از اجرای الگوریتم c-means با استفاده از ۷۰ گزینه تصادفی برای انتخاب مراکز اولیه خوشه‌ها، در تمام حالت‌های دو تا پنج منطقه‌ای و برای تمام مقادیر مورد بررسی μ، میزان تغییرات نسبی میانگین تجمعی مقدار تابع هدف به کمتر از یک درصد می‌رسد. از این رو، مقایسه مقادیر به‌دست آمده برای تابع هدف با استفاده از مراکز خوشه‌های حاصل از به‌کارگیری نگاشت‌های خود سازمانده به‌عنوان مراکز اولیه خوشه-ها در الگوریتم فازی c-means، با میانگین مقادیر تابع هدف در اجرای الگوریتم c-means با استفاده از ۱۰۰ انتخاب اولیه مراکز خوشه‌ها به‌صورت تصادفی، می-تواند تأثیر روش مذکور در کارایی الگوریتم c-means را نشان دهد.

در شکل‌های ۵ و ۶ مقادیر محاسبه شده تابع هدف در روش ترکیبی نگاشت‌های خود سازمانده و الگوریتم c-means با میانگین مقادیر تابع هدف در اجرای الگوریتم c-means با استفاده از ۱۰۰ انتخاب تصادفی مراکز اولیه خوشه‌ها مقایسه شده است.

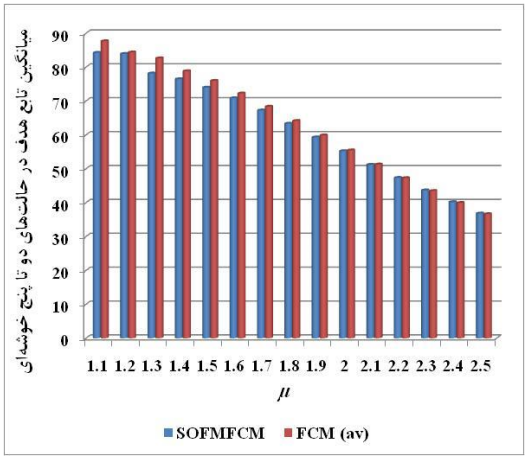
شکل ۵، نشان‌دهنده حالتی است که برای هر یک از مقادیر μ در فاصله [1.1,2.5]، میانگین مقدار تابع هدف در چهار حالت دو تا پنج خوشه‌ای، هم برای نتایج حاصل از ترکیب نگاشت‌های خود سازمانده و الگوریتم c-means و هم برای نتایج حاصل از ۱۰۰ انتخاب تصادفی مراکز اولیه الگوریتم c-means محاسبه شده است. نتایج مندرج در شکل نشان می-دهد که استفاده از مراکز خوشه‌های تشکیل شده به-وسیله نگاشت‌های خود سازمانده به‌عنوان مراکز اولیه خوشه‌ها در الگوریتم c-means در فاصله [1.1,2.1]، موجب بهبود توانایی الگوریتم c-means در کمینه-سازی تابع هدف می‌شود، اگر چه در فاصله‌ای که $\mu\epsilon$[2.2,2.5] باشد، عملکرد این ترکیب در حد بسیار اندکی، ضعیف‌تر از وضعیت متوسط است.

شکل ۶ مربوط به وضعیتی است که برای هر یک از حالات دو، سه، چهار و پنج خوشه‌ای، میانگین

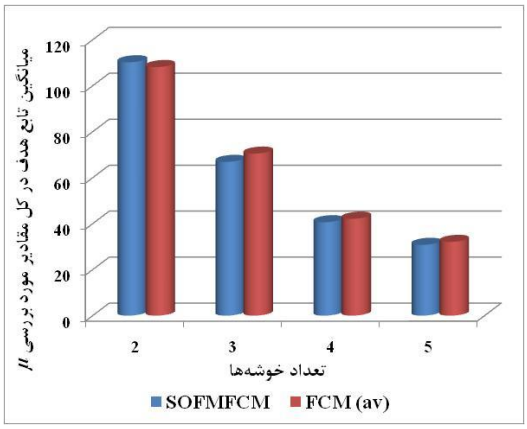
مقدار تابع هدف در تمام مقادیر مورد بررسی μ در فاصله [1.1,2.5]، محاسبه شده و بین آن‌ها میانگین گرفته شده است. این فرایند هم برای نتایج حاصل از ترکیب نگاشت‌های خود سازمانده و الگوریتم c-means و هم برای نتایج حاصل از ۱۰۰ انتخاب تصادفی مراکز اولیه الگوریتم c-means اجرا شده است.



شکل ۴- تغییرات نسبی میانگین تجمعی مقدار تابع هدف در اجرای الگوریتم c-means با ۱۰۰ انتخاب تصادفی مراکز اولیه برای حالت‌های دو تا پنج خوشه‌ای



شکل ۵- میانگین مقدار تابع هدف چهار حالت دو، سه، چهار و پنج خوشه‌ای در مقادیر مختلف μ؛ SOFMFCM معرف روش ترکیبی نگاشت‌های خود سازمانده و الگوریتم فازی c-means و FCM (av) معرف میانگین تجمعی تابع هدف الگوریتم فازی c-means برای ۱۰۰ انتخاب تصادفی مراکز اولیه است.



شکل ۶- میانگین مقدار تابع هدف در مقادیر مورد بررسی μ در فاصله [1.1,2.5] برای تعداد مختلف خوشه‌ها

آن‌چنان‌که در شکل ۶ مشاهده می‌شود، ترکیب نگاشت‌های خود سازمانده با الگوریتم فازی c-means، در حالت‌های سه، چهار و پنج خوشه‌ای عملکرد بهتر این الگوریتم در کمینه‌سازی تابع هدف را نسبت به وضعیت متوسط اجرای الگوریتم c-means به‌دنبال دارد، با این حال در حالت دو خوشه‌ای، عملکرد این ترکیب تا حدی ضعیف‌تر از وضعیت متوسط است.

در مرحله بعدی، مقادیر شاخص‌های صحت خوشه‌بندی فازی ضریب افراز، انتروپی افراز، شاخص ژی-بنی توسعه یافته و شاخص کوان برای حالت‌های دو تا پنج منطقه‌ای حاصل از ترکیب نگاشت‌های خود-سازمانده با الگوریتم c-means محاسبه شدند که نتایج به‌دست آمده در شکل ۷ منعکس شده است.

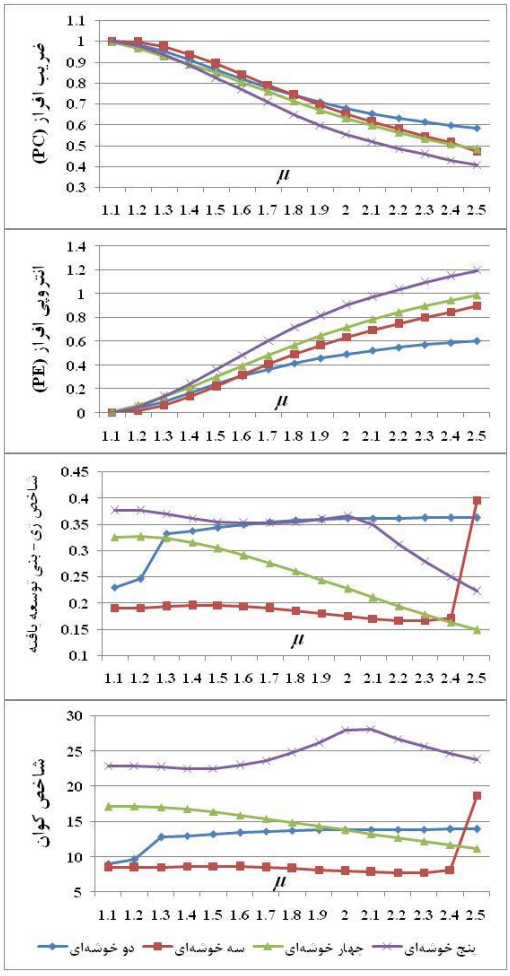
همان‌طور که در نمودارهای موجود در شکل ۷ دیده می‌شود، مقدار شاخص ضریب افراز (V_{PC}) برای تمامی حالات دو تا پنج خوشه‌ای با افزایش مقدار μ در محدوده مورد بررسی، کاهش می‌یابد. این قاعده در مورد شاخص انتروپی افراز (V_{PE})، دقیقاً برعکس بوده و با افزایش مقدار μ در محدوده مورد بررسی، مقدار این شاخص افزایش می‌یابد. در مقابل با بررسی نمودارهای مربوط به شاخص ژی-بنی توسعه یافته ($V_{XB,m}$) و شاخص کوان (V_K) به‌نظر می‌رسد که تغییرات این دو شاخص روند خاصی را با تغییرات مقدار μ دنبال نمی‌کند.

بررسی نمودار مربوط به تغییرات شاخص ضریب افراز (V_{PC}) نشان می‌دهد که بزرگ‌ترین (بهترین) مقادیر این شاخص در فاصله $1.1 \leq \mu \leq 1.8$ مربوط به حالت سه خوشه‌ای و در فاصله $1.9 \leq \mu \leq 2.5$ متعلق به حالت دو خوشه‌ای است. این روند در مورد شاخص انتروپی افراز (V_{PE}) هم تا حدودی مشاهده می‌شود، به‌طوری‌که کوچک‌ترین مقادیر این شاخص در محدوده $1.1 \leq \mu \leq 1.5$ در حالت سه خوشه‌ای و در محدوده $1.6 \leq \mu \leq 2.5$ در حالت دو خوشه‌ای مشاهده می‌شوند.

رفتار شاخص‌های صحت خوشه‌بندی در محدوده $2 < \mu \leq 2.5$ نیز مورد بررسی قرار گرفت که در این محدوده تغییر خاصی در رفتار این شاخص‌ها مشاهده نشد. به‌طور خلاصه بر اساس نتایج محاسبه این دو شاخص به ازای تمام مقادیر مورد بررسی μ، حالت‌های سه خوشه‌ای و دو خوشه‌ای همواره گزینه‌های مناسب‌تری برای اجرای خوشه‌بندی هستند.

در بررسی نمودار مربوط به شاخص ژی-بنی توسعه یافته ($V_{XB,m}$) مشاهده می‌شود که حالت سه خوشه‌ای در سراسر گستره $1.1 \leq \mu \leq 2.3$ در مقایسه با سایر حالت‌ها کوچک‌ترین (بهترین) مقادیر را اختیار می‌کند و تنها در بازه $2.4 \leq \mu \leq 2.5$، کمترین مقادیر به حالت چهار خوشه‌ای تعلق می‌گیرد.

مشابه این وضعیت در بررسی نمودار مربوط به شاخص کوان (V_K) نیز دیده می‌شود، با این تفاوت که محدوده‌ای که حالت سه خوشه‌ای در آن حالت بهینه خوشه‌بندی است به بازه $1.1 \leq \mu \leq 2.4$ گسترش می‌یابد و تنها در نقطه $\mu = 2.5$ است که حالت چهار خوشه‌ای نسبت به حالت سه خوشه‌ای وضعیت بهتری پیدا می‌کند. همچنین، بررسی نتایج به‌دست آمده برای محاسبه این دو شاخص نشان می‌دهد که حالت دو خوشه‌ای اگرچه بهترین مقادیر را اختیار نمی‌کند، اما در موارد متعددی پس از حالت سه خوشه‌ای می‌تواند بهترین گزینه برای اجرای خوشه‌بندی باشد.



شکل ۷- نمودار تغییرات شاخص‌های صحت خوشه‌بندی فازی V_{PC}، V_{PE}، $V_{XB,m}$ و V_K با تغییر مقدار μ و تعداد خوشه‌ها

نهایی‌سازی اختصاص سایت‌ها به خوشه‌های حاصل از به‌کارگیری الگوریتم c-means به‌صورت افراز فازی انجام گرفت و هر سایت به تمام خوشه‌هایی که درجه عضویتش در آن‌ها بیش از آستانه $1/c$ بود، اختصاص یافت. در این حالت، با افزایش مقدار μ و خاصیت فازی بودن خوشه‌ها، حضور برخی سایت‌ها در چند خوشه مختلف، موجب افزایش تعداد سایت‌ها و داده‌های آماری موجود در آن خوشه‌ها می‌شود. اما باید توجه داشت که افزودن سایت‌های بیشتر، ممکن

است با افزایش میزان ناهمگنی همراه باشد.

در شکل ۸ مجموع تعداد ایستگاه‌های اختصاص یافته به هر یک از مناطق در حالت‌های دو تا پنج خوشه‌ای در هر یک از مقادیر مورد بررسی μ مشاهده می‌شود. در این شکل تعداد سایت‌های عضو در هر منطقه نیز مشخص شده است. همان‌طور که در این شکل دیده می‌شود، در حالت دو منطقه‌ای به‌دلیل آن که مقدار درجه عضویت یک ایستگاه همواره تنها در یک منطقه می‌تواند بالاتر از میزان آستانه عضویت (۰،۵) باشد، لذا با تغییر مقدار عامل فازی‌کننده μ هیچ تغییری در مجموع تعداد سایت‌های عضو در کل دو منطقه پدید نمی‌آید و تنها در دو گام انتقال از ۱،۲ به ۱،۳ و از ۲،۳ به ۲،۴ تغییراتی در نحوه اختصاص سایت‌ها به مناطق و تعداد سایت‌های عضو در هر منطقه اتفاق می‌افتد. بر خلاف حالت دو خوشه‌ای، در حالت‌های سه، چهار و پنج خوشه‌ای، کاهش مقدار آستانه عضویت در مناطق موجب می‌شود که گاه یک سایت به‌صورت هم‌زمان در بیش از یک منطقه عضویت داشته باشد که این امر سبب افزایش مجموع تعداد سایت‌های عضو در مناطق تشکیل شده در هر حالت می‌شود.

بر اساس نتایج مندرج در شکل ۸، افزایش مجموع تعداد سایت‌ها در حالت سه منطقه‌ای از مقدار $\mu = 1.8$ آغاز شده و تا انتهای فاصله مورد بررسی ادامه می‌یابد به‌طوری‌که در مقدار $\mu = 2.5$ حاصل جمع تعداد سایت‌های عضو در کل سه منطقه به عدد ۴۳ می‌رسد. همچنین، در حالت‌های چهار و پنج منطقه‌ای با کاهش بیشتر آستانه عضویت، مجموع تعداد سایت‌های عضو در مناطق در $\mu = 2.5$، به‌ترتیب به ۴۵ و ۵۷ می‌رسد.

افزایش تعداد ایستگاه‌های هیدرومتری یا سایت‌های عضو در مناطق، یکی از مزایای استفاده از منطقه‌بندی فازی در مقایسه با منطقه‌بندی سخت است، زیرا این افزایش موجب افزایش تعداد داده‌های آماری موجود در هر منطقه شده و امکان برآورد سیلاب برای دوره‌های بازگشت طولانی‌تر را فراهم می‌کند. البته در کنار این ویژگی مثبت این نکته نیز باید مدنظر قرار گیرد که افزایش مجموع تعداد سایت‌ها، بسته به ناحیه و سایت‌های مورد مطالعه گاهی ممکن است با افزایش میزان ناهمگنی همراه باشد.

در شکل ۹، مقادیر محاسبه شده شاخص ناهمگنی H_1 برای حالت‌های دو تا پنج منطقه‌ای مشاهده می‌شوند. مطابق این شکل، در حالت دو منطقه‌ای اگر چه تعداد مجموع سایت‌های عضو در مناطق با تغییر مقدار μ افزایش نمی‌یابد، اما در حالی‌که در مقادیر $\mu = 1.1$ و $\mu = 1.2$ بر اساس شاخص H_1 یکی از مناطق در وضعیت ناهمگن قرار دارد، با تغییر در نحوه اختصاص سایت‌ها به مناطق از مقدار $\mu = 1.3$ تا انتهای محدوده مورد بررسی، هر دو منطقه وضعیت همگن پیدا می‌کنند. لذا به رغم عدم افزایش در مجموع تعداد سایت‌های عضو در مناطق، حالت دو منطقه‌ای حاصل از منطقه‌بندی با استفاده از ترکیب نگاشت‌های خود سازمانده و الگوریتم c-means از حیث همگنی مناطق نسبت به حالت دو منطقه‌ای حاصل از استفاده صرف از نگاشت‌های خود سازمانده برتری دارد.

در حالت سه منطقه‌ای و در فاصله $1.1 \le \mu \le 1.8$، بر اساس شاخص H_1 هر سه منطقه در وضعیت همگن قرار دارند، اما در بازه $1.9 \le \mu \le 2.5$، یکی از مناطق در وضعیت ناهمگن قرار می‌گیرد. در حالت چهار منطقه‌ای با وجود تغییراتی که بعضاً در مقادیر شاخص H_1 در برخی مناطق مشاهده می‌شود، در فاصله $1.1 \le \mu \le 2.0$، همواره هر چهار منطقه تشکیل شده دارای وضعیت همگن هستند. اما در بازه $2.0 \le \mu \le 2.5$، یکی از مناطق دچار ناهمگنی می‌شود. در حالت پنج منطقه‌ای در سراسر فاصله $1.1 \le \mu \le 2.5$، همواره یکی از مناطق تشکیل شده دارای وضعیت ناهمگن است.

با توجه به توضیحات پیشین و در نظر گرفتن سه معیار کیفیت خوشه‌بندی، بزرگی مناطق و همگنی مناطق به‌ترتیب بر اساس شاخص‌های صحت خوشه‌بندی، تعداد سایت‌های عضو در مناطق و شاخص‌های ناهمگنی، سه گزینه برای ادامه فرایند تحلیل فراوانی منطقه‌ای سیلاب انتخاب شدند که عبارتند از: حالت دو منطقه‌ای با $\mu = 2.5$، حالت سه منطقه‌ای با $\mu = 1.8$ و حالت چهار منطقه‌ای با $\mu = 2.0$.

در ادامه با استفاده از شاخص نکویی برازش Z، توزیع منطقه‌ای مناسب برای هر یک از مناطق تشکیل شده در سه حالت برگزیده نهایی مشخص شد که

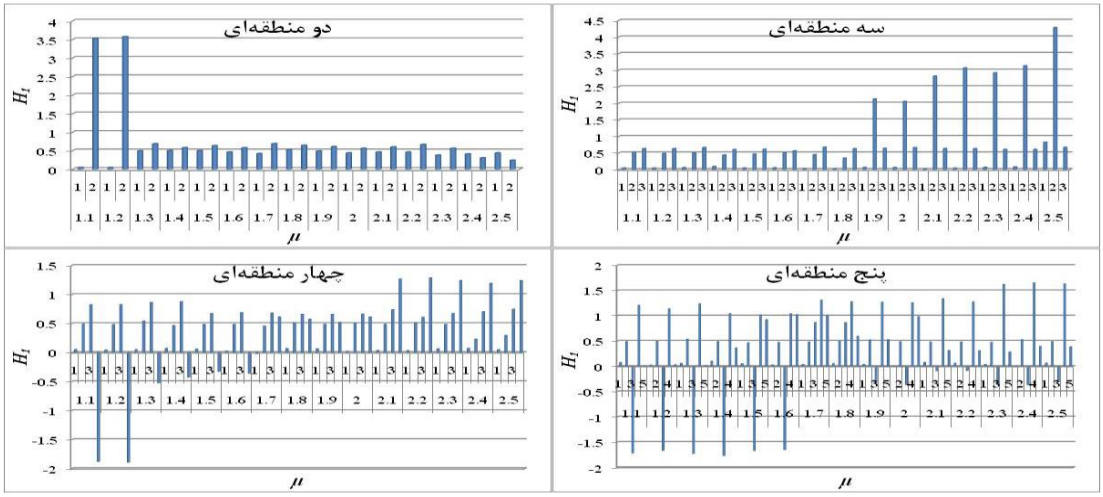
توزیع منتخب برای هر منطقه در جدول ۲ مشخص شده است. پس از آن با استفاده از تئوری گشتاورهای خطی، پارامترهای هر یک از توزیع‌های منطقه‌ای برآورد شده و سپس توزیع فراوانی ویژه هر سایت با استفاده از ضرب عامل شاخص سیلاب که برابر میانگین دبی لحظه‌ای بیشینه سیلاب سالانه آن سایت در نظر گرفته شد، در توزیع فراوانی منطقه‌ای مربوطه به‌دست آمد.

اگر چه با توجه به تعداد داده‌های آماری موجود در مناطق تشکیل شده، در بعضی مناطق بزرگ مانند منطقه دوم حالت دو منطقه‌ای و منطقه سوم حالت سه منطقه‌ای، امکان برآورد سیلاب حتی با دوره‌های بازگشت بزرگ‌تر از ۱۰۰ سال نیز وجود دارد، اما به‌منظور مقایسه برآوردهای سیلاب حاصل از سه گزینه نهایی باید بزرگ‌ترین دوره بازگشتی را که برآورد سیلاب نظیر آن برای تمام مناطق تشکیل شده قابل اعتماد است، برگزیده می‌شد. این دوره بازگشت تقریباً برابر ۲۵ سال است.

جدول ۲- توزیع‌های منتخب منطقه‌ای برای مناطق مربوط به سه حالت برگزیده نهایی

حالت برگزیده	شماره منطقه	توزیع منطقه‌ای
دو منطقه‌ای با $\mu = 2.5$	۱	نرمال تعمیم یافته
	۲	پیرسون تیپ III
سه منطقه‌ای با $\mu = 1.8$	۱	مقادیر حدی تعمیم یافته
	۲	نرمال تعمیم یافته
	۳	پیرسون تیپ III
چهار منطقه‌ای با $\mu = 2.0$	۱	مقادیر حدی تعمیم یافته
	۲	نرمال تعمیم یافته
	۳	پیرسون تیپ III
	۴	پیرسون تیپ III

شکل ۸- مجموع تعداد ایستگاه‌های عضو در کل مناطق برای هر یک از حالات دو تا پنج خوشه‌ای و در فاصله $1.1 \leq \mu \leq 2.5$

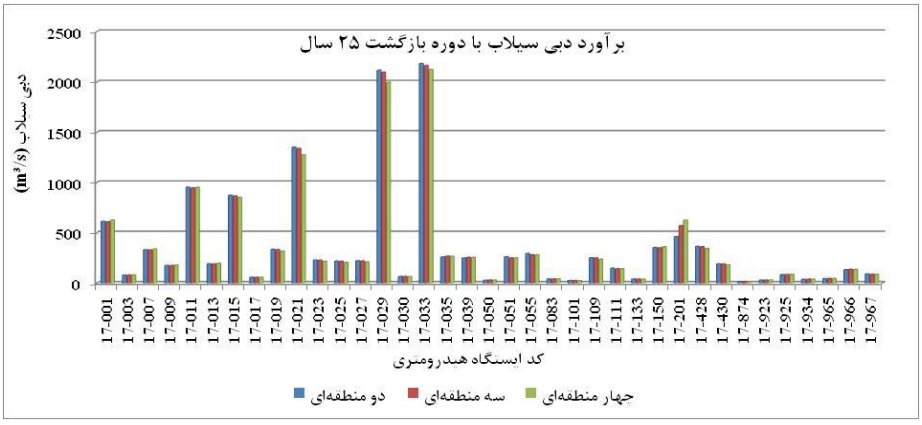


شکل ۹- مقادیر شاخص ناهمگنی H_1 در هر یک از حالات دو تا پنج خوشه‌ای و در فاصله $1.1 \leq \mu \leq 2.5$

در شکل ۱۰، مقادیر برآورد شده دبی سیلاب با دوره بازگشت ۲۵ سال برای هر یک از سایت‌ها در سه حالت منطقه‌بندی نهایی نشان داده شده است که بر اساس آن در میان سایت‌های مورد بررسی، بزرگ‌ترین دبی سیلاب برآورد شده با دوره بازگشت ۲۵ سال مربوط به ایستگاه ۱۷-۰۳۳ و در محدوده تقریبی ۲۱۲۳ تا ۲۱۸۳ مترمکعب بر ثانیه است. در مقابل کوچک‌ترین دبی سیلاب برآورد شده متعلق به ایستگاه ۱۷-۸۷۴ و حدود ۱۵ مترمکعب بر ثانیه است. بررسی این نتایج همچنین نشان می‌دهد که به‌جز ایستگاه ۱۷-۲۰۱ که در آن بیشینه اختلاف نسبی بین برآورد-های حاصل از سه گزینه منطقه‌بندی به حدود ۲۵ درصد می‌رسد، در هیچ یک از ایستگاه‌ها، میزان این اختلاف از شش درصد تجاوز نمی‌کند.

نتایج به‌دست آمده در این پژوهش نشان می‌دهد که ترکیب نگاشت‌های خود سازمانده با الگوریتم خوشه‌بندی فازی c-means، از نظر تشکیل مناطق همگن و برآوردهای قابل‌قبول سیلاب، می‌تواند روشی مناسب برای منطقه‌بندی و تحلیل فراوانی منطقه‌ای سیلاب در حوزه آبخیز سفیدرود باشد. این نتیجه با نتایج ارائه شده توسط Rao و Srinivas (۲۰۰۶b)، Srinivas و همکاران (۲۰۰۸) و Sadri و Burn (۲۰۱۱) مطابقت دارد. در این مطالعه مشخص شد که انتخاب مراکز اولیه خوشه‌ها برای اجرای الگوریتم فازی c-means با استفاده از مراکز خوشه‌های تشکیل شده به‌وسیله نگاشت‌های خود سازمانده، می‌تواند در کمینه‌سازی تابع هدف الگوریتم c-means اثری مثبت داشته باشد. همچنین، استفاده از ترکیب نگاشت‌های خود سازمانده و الگوریتم فازی c-means و اجرای منطقه‌بندی فازی می‌تواند با افزایش تعداد داده‌های آماری موجود در مناطق تشکیل شده، امکان برآورد سیلاب برای دوره‌های بازگشت طولانی‌تر را فراهم کند. این نکته نیز قابل توجه است که در مواردی مانند حالت دو منطقه‌ای در این پژوهش، استفاده ترکیبی از نگاشت‌های خود سازمانده و الگوریتم فازی c-means می‌تواند با تغییر نحوه توزیع سایت‌ها در مناطق، موجب تغییر و بهبود احتمالی وضعیت همگنی مناطق تشکیل شده شود.

پیشنهاد می‌شود در مطالعات آینده منطقه‌بندی حوزه‌های آبخیز با استفاده از ویژگی‌های دیگر فیزیوگرافی، هواشناسی، زمین‌شناسی، کاربری اراضی، خاک‌شناسی و دیگر عوامل احتمالی مؤثر بر فرایند تولید سیلاب مورد بررسی قرار گیرد. همچنین جستجو برای یافتن روش‌هایی کارآمدتر به‌منظور تعیین مراکز اولیه خوشه‌ها در الگوریتم‌های خوشه‌بندی K-means و c-means می‌تواند عملکرد این الگوریتم‌ها را در منطقه‌بندی حوزه‌های آبخیز بهبود بخشد.



شکل ۱۰- برآورد دبی سیلاب با دوره بازگشت ۲۵ سال برای ایستگاه‌های هیدرومتری مورد مطالعه در سه حالت نهایی منطقه‌بندی

منابع مورد استفاده

1. Bezdek, J.C. 1981. Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York, 256 pages.
2. Bezdek, J.C. 1987. Partition structures: a tutorial. In: “The analysis of fuzzy information”. CRC Press, Boca Raton, 312 pages.

3. Burn, D.H. 1989. Cluster analysis as applied to regional flood frequency. Journal of Water Resources Planning and Management, 115: 567–582.
4. Burn, D.H. and N.K. Goel. 2000. The formation of groups for regional flood frequency analysis. Hydrological Sciences Journal, 45: 97–112.
5. Fausett, L.V. 1994. Fundamentals of neural networks: architectures, algorithms, and applications. Englewood Cliffs, Prentice Hall, 461 pages.
6. Hall, M.J. and A.W. Minns. 1999. The classification of hydrologically homogeneous regions. Hydrological Sciences Journal, 44: 693–704.
7. Hall, M.J., A.W. Minns and A.K.M. Ashrafuzzaman. 2002. The application of data mining techniques for the regionalization of hydrological variables. Hydrology and Earth System Sciences, 6: 685–694.
8. Hosking, J.R.M. and J.R. Wallis. 1997. Regional frequency analysis: an approach based on l-moments. Cambridge University Press, New York, 224 pages.
9. Jingyi, Z. and M.J. Hall. 2004. Regional flood frequency analysis for the Gan-Ming river basin in China. Journal of Hydrology, 296: 98–117.
10. Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43: 59–69.
11. Kwon, S.H. 1998. Cluster validity index for fuzzy clustering. Electronics Letters, 34: 2176–2177.
12. Nathan, R.J. and T.A. McMahon. 1990. Identification of homogeneous regions for the purposes of regionalization. Journal of Hydrology, 121: 217–238.
13. Rao, A.R. and V.V. Srinivas. 2006a. Regionalization of watersheds by hybrid cluster analysis. Journal of Hydrology, 318: 37–56.
14. Rao, A.R. and V.V. Srinivas. 2006b. Regionalization of watersheds by fuzzy cluster analysis. Journal of Hydrology, 318: 57–79.
15. Rao, A.R. and V.V. Srinivas. 2008. Regionalization of watersheds-an approach based on cluster analysis, series: water science and technology library. Springer Publishers, 248 pages.
16. Tasker, G.D. 1982. Comparing methods of hydrologic regionalization. Water Resources Bulletin, 18: 965–970.
17. Sadri, S. and D.H. Burn. 2011. A Fuzzy C-Means approach for regionalization using a bivariate homogeneity and discordancy approach. Journal of Hydrology, 401: 231-239.
18. Srinivas, V.V., S. Tripathi, A.R. Rao and R.S. Govindaraju. 2008. Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. Journal of Hydrology, 348: 148–166.
19. Xie, X.L. and G. Beni. 1991. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13: 841–847.

Regionalization of watersheds by combining of self-organizing feature maps and fuzzy C-Means algorithm

Ali Ahani[*1] and Seyed Saeed Mousavi Nadoshani[2]
[1] MSc, Faculty of Agriculture, University of Shahrood, Iran and [2] Assistant Professor, Faculty of Water and Environmental Engineering, Shahid Beheshti University, Iran

Received: 21 April 2014 Accepted: 20 December 2014

Abstract

Cluster analysis methods are one of the most efficient approaches of regionalization of watersheds for Regional Flood Frequency Analysis (RFFA). Fuzzy regionalization is a kind of regionalization in which each site of interest may be assigned to more than one region simultaneously. In order to perform regionalization, a variety of cluster analysis algorithms named fuzzy clustering are used that fuzzy c-means algorithm is most well-known of them. Also Self-Organizing Feature Maps (SOFM) are a special class of Artificial Neural Networks (ANN) that has found several applications in the areas of pattern recognition. Capability of this class of networks in areas of pattern recognition and data clustering using their attributes has make some hydrologist interested in testing ability of these maps for regionalization of watersheds in order to perform regional flood frequency of analysis. In this study self-organizing feature maps have been used to determine initial centroids of clusters in fuzzy c-means algorithm for regionalization of Sefidrood watershed. Results of this study showed that this approach has an acceptable performance in formation of homogeneous regions and providing suitable estimates in regional flood frequency analysis using L-moment algorithm in interested watershed. Furthermore it is observed that fuzzy clustering may provide longest reliable flood estimates. Also based on fuzzy clustering validity measures it's seemed that two or three regions is appropriate number of regions for regional flood frequency analysis in this watershed.

Key words: Artificial neural networks, Cluster analysis, L-moments, Regional flood frequency analysis, Regionalization

[*] Corresponding author: ali.ahani66@yahoo.com