نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد 28، شماره 3، مرداد – شهریور 1393، ص. 480-471

Journal of Water and Soil
Vol. 28, No. 3, Jul.-Agu. 2014, p. 471-480

افزایش کارایی فرایند خوشه‌بندی K- میانگین توسط روش‌های سلسله مراتبی

بیژن قهرمان[1*] - کامران داوری[2]

تاریخ دریافت:1392/1/20
تاریخ پذیرش:1392/11/28

چکیده

به علت کمبود آمار و اطلاعات همیشه امکان استفاده از تحلیل فراوانی مکانی جهت تخمین چندک‌های سیلاب وجود ندارد. از آن‌جاکه استفاده از یک روش واحد برای ناحیه‌ای کردن معمولاً نتایج قابل قبولی را به دست نمی‌دهد، لذا معمولاً چندین روش منطقه‌ای به‌طور توأم مورد استفاده قرار می‌گیرد. در این مطالعه سه الگوریتم خوشه‌ای هیبرید که هر یک به طور جداگانه فرایند خوشه‌ای کردن را برای تعیین نواحی مشابه به کار می‌برند، مورد بررسی قرار گرفت. از الگوریتم‌های خوشه‌ای سلسله مراتبی متراکمی از روش‌های پیوند تکی، پیوند کامل و وارد، و از الگوریتم خوشه‌ای تفکیکی، از الگوریتم K- میانگین استفاده شد. تأثیر تحلیل خوشه‌ای هیبرید در ناحیه‌ای کردن با استفاده از آمار روزآمد شده‌ی 68 حوضه‌ی آبریز استان‌های خراسان مورد بررسی قرار گرفت. از چهار شاخص آزمون خوشه‌ای شامل ضریب کوفنتیک، متوسط عرض سیلهوت، نمایه‌های دان و دیویس- بولدین جهت تعیین تعداد بهینه‌ی خوشه‌ها استفاده گردید. نتایج نشان داد که روش‌های پیوند تکی و کامل برپایه‌ی نمایه‌های اعتبارسنجی ضریب کوفنتیک و متوسط عرض سیلهوت بهتر بودند ولی منجر به تشکیل خوشه‌هایی نامتجانس (یک خوشه‌ی بزرگ و تعدادی خوشه‌ی بسیار کوچک) گردید که در تحلیل فراوانی سیلاب مناسب نمی‌باشد. گرچه تحلیل خوشه‌ای هیبرید در حداقل‌سازی تلاش لازم جهت نیل به نواحی همگن مؤثر بود ولی درجه‌ی موثر بودن به تعداد خوشه‌ها بستگی داشت. معیارهای ناهمگنی هاسکینگ منفی بود که بیانگر همبستگی سیلاب در ایستگاه ها بود. نهایتاً هیبرید الگوریتم وارد و K- میانگین برای استفاده در ناحیه‌ای کردن پیشنهاد گردید. چهار ناحیه همگن تشخیص داده شد.

واژه‌های کلیدی: تحلیل فراوانی منطقه‌ای سیلاب، خراسان، خوشه بندی، گشتاورهای خطی، ناحیه‌بندی، هیبرید

مقدمه

مشکلات فراوانی در اندازه‌گیری سیلاب وجود دارد. در نتیجه، طول دوره‌ی آماری مقادیر سیلاب حداکثر لحظه‌ای در بسیاری از ایستگاه‌های آب‌سنجی کوتاه است. پذیرفته شده است که برازش تابع احتمال توزیع مناسب بر داده‌هایی با طول دوره‌ی آماری کوتاه رضایت‌بخش نیست، نتایج اریب بوده و مقادیر سیلاب تخمین زده شده از خطای استاندارد بالایی برخوردار خواهند بود. یکی از راه‌های برون‌رفت از این دشواری، استفاده از مفهوم سیلاب منطقه‌ای است که در آن ایستگاه‌هایی که رفتاری یکسان دارند در ناحیه‌ای همگن، به نام خوشه، قرار می‌گیرند. تحلیل سیلاب منطقه‌ای کاربردی گسترده در تخمین مقادیر سیلاب در طراحی و عملکرد سامانه‌های منابع آب، طراحی و مدیریت کاربری اراضی، مسایل مربوط به بیمه‌ی سیلاب و حفاظت از نواحی مسکونی و سایر موارد مشابه دارد. از دیرباز پژوهشگران درصدد یافتن متغیرهایی بوده‌اند که بر روی سیلاب تاثیر دارند. نگرش‌هایی در راستای تشخیص و تفسیر الگوهای نهفته در داده‌ها منطقاً منجر به خوشه‌بندی می‌شود. خوشه‌بندی مقوله‌ای از علم آمار کاربردی چند متغیره به‌شمار می‌آید (4).

تا به امروز یک روش واحد برای خوشه‌بندی حوضه‌های آبریز بین هیدرولوژیست‌ها وجود ندارد. چنین چارچوبی برای طبقه‌بندی باید ویژگی‌های عمومی حوضه را همراه با شرایط هیدروکلیماتولوژی موثر در پاسخ حوضه و با نگاهی به عدم‌قطعیت‌های زمانی و مکانی در بر داشته باشد. واگنر و همکاران (29) نگرش‌های موجود برای تعریف شباهت هیدرولوژیکی و طبقه‌بندی حوضه‌های آبریز را مرور کردند؛ مولفه‌ها یا ویژگی‌های بارزی که باید در الگوی طبقه‌بندی درنظر گرفته شود را بحث کرده‌اند؛ و چارچوبی پایه‌ای را برای طبقه‌بندی حوضه‌ها به‌عنوان نقطه‌ی آغازین برای تحلیل‌های بعدی تهیه کردند. نام‌بردگان پرسش‌هایی همچون (الف) چگونه می‌توان به بهترین گونه ویژگی‌های فیزیکی حوضه و شرایط هیدروکلیماتولوژی را ارایه نمود؟

1 و 2- استاد و دانشیار گروه مهندسی آب، دانشکده کشاورزی، دانشگاه فردوسی مشهد
(*- نویسنده مسئول: (Email: bijangh@um.ac.ir

(ب) این ارایه چگونه در مقیاس‌های مکانی و زمانی تغییر می‌کند؟ (پ) چه ویژگی‌هایی از پاسخ حوضه در چه مقیاس مکانی و زمانی مناسب‌تر است؟ (ت) ساختار درونی و غیرهمگنی در چه مقیاسی مهم است و باید درنظر گرفته شود؟ را مطرح ساختند.

در منابع روش‌های متعددی برای تشخیص نواحی همگن وجود دارد. استفاده از سامانه‌ی خبره‌ی فازی با کمک الگوریتم ژنتیک برای تهیه‌ی معیاری از شباهت بین حوضه‌ها (27)، تشخیص هیدرولوژیکی نواحی همگن براساس شکل تابع توزیع تجمعی تجربی و شباهت‌های ویژگی‌های فیزیوگرافیک و اقلیمی (6)، استفاده از مفهوم ناحیه‌ی تاثیر که دربرگیرنده‌ی ایستگاه‌های دارای آمار هیدرولوژیکی بوده و با ناحیه‌ی مدنظر تشابه معقولی دارد (10) از آن جمله است. این ناحیه عموماً یا در فضای جغرافیایی یا این‌که در فضای متغیرهایی که برای تخمین جریان رودخانه‌ای استفاده می‌شود (متغیرهای پیشگو) قابل تعریف است. این نگرش‌ها مکمل هم بوده و ترکیبی از این دو بر هرکدام از آن‌ها به تنهایی برتری دارد. بنابراین انگ و همکاران (19) روش رگرسیون ناحیه‌ی تاثیر هیبرید را معرفی کرده که دو نگرش را با هم ترکیب می‌کند. برای مشخص کردن نواحی همگن، اتیم و هارمانک‌لوگلو (7) ناحیه را به زیرناحیه‌هایی تقسیم و ساختار آن را مجدداً تعریف کردند. این کار در دو مرحله انجام می‌شود: (الف) حذف برخی ایستگاه‌ها از ناحیه و استفاده از تخصیصی کاملاً متفاوت از ایستگاه‌ها به زیرناحیه‌های همگنِ مشخص شده که اضافه شده‌اند، و (ب) اضافه کردن یک ویژگی از خوشه به نواحی منطقه‌ای. روش‌های بیان شده یا بسیار به قضاوت شخصی متکی است (روش آخر) یا این که بسیار پیچیده‌اند. نگرش سنتی به مسئله به‌طور متداول هنوز هم با تحلیل خوشه‌ای انجام می‌شود.

هر خوشه از یک یا چند ایستگاه تشکیل شده و هر ایستگاه دربرگیرنده‌ی چندین «متغیر» است. در هیدرولوژی، متغیرها که از آن‌ها برای تحلیل فراوانی منطقه‌ای سیلاب (تف‌م‌س) استفاده می‌شود شامل: (الف) ویژگی‌های فیزیوگرافیکی حوضه‌های آبریز از قبیل مساحت، شیب متوسط حوضه، شیب متوسط آبراهه‌ی اصلی، طول آبراهه‌ی اصلی، تراکم آبراهه‌ای، نمایه‌ی ذخیره در حوضه، نمایه‌ی گونه‌ی خاک از قبیل پتانسیل نفوذپذیری، ویژگی‌های روان‌آب یا متوسط کمبود موثر رطوبت خاک، جزئی از حوضه که توسط دریاچه، مخزن آب یا باتلاق پوشیده شده است؛ (ب) متغیرهای موقعیت جغرافیایی از قبیل طول و عرض جغرافیایی و ارتفاع مرکز ثقل حوضه؛ (پ) معیاری از زمان پاسخ حوضه از قبیل زمان تاخیر یا زمان تا اوج؛ (ت) عوامل هواشناسی از قبیل جهت رگبار، میانگین باران سالانه، شدت بارندگی؛ و (ث) آمارهای سیلاب در محل ایستگاه می‌باشد. انتخاب مجموعه‌ای مناسب برای تف‌م‌س، کلید اصلی به‌شمار می‌آید زیرا امکان دارد نواحی همگن را کاملاً تغییر دهد. گرچه برای این کار نگرش قضاوت شخصی هنوز هم در هیدرولوژی متداول است (مثلاً 24)، نگرش‌های ساختارمندی نیز وجود دارد (مثلاً دین‌پژوه و همکاران (17) با استفاده از تحلیل عاملی و لین و همکاران (22) با استفاده از تحلیل مولفه‌های اصلی).

تشکیل خوشه‌ها فرآیندی است که با کمک آن مجموعه‌ای از ایستگاه‌ها به خوشه‌ها یا گروه‌هایی تقسیم می‌شوند به‌گونه‌ای که ایستگاه‌های درون هر خوشه تا حد ممکن مشابه بوده درحالی‌که ایستگاه‌ها در خوشه‌های متفاوت تا حد ممکن متمایز باشند. الگوریتم‌های خوشه‌سازی در حالت کلی به دو دسته‌ی خوشه‌سازی سلسله مراتبی و خوشه‌سازی تفکیکی تقسیم می‌شوند.

در خوشه‌سازی سلسله مراتبی برای تشکیل خوشه‌های بزرگ‌تر، خوشه‌های کوچک‌تر به‌طور پیوسته با هم ترکیب می‌شوند (تجمعی) یا این‌که خوشه‌های بزرگ‌تر به خوشه‌های کوچک‌تر شکسته می‌شوند (تقسیمی). در نقطه‌ی مقابل، در خوشه‌سازی تفکیکی گروه‌بندی طبیعی موجود در داده‌ها به‌طور پیوسته بهبود بخشیده می‌شود. مثال‌هایی از این دسته از الگوریتم‌ها شامل فرایند K- میانگین (ف‌کم) (23)، الگوریتم‌های K- میانه و K- نما است. مثالی از کاربرد الگوریتم خوشه‌سازی سلسله مراتبی تجمعی برای ناحیه‌بندی حوضه‌های آبریز در کانادا توسط برن و همکاران (13) و ف‌کم و شکل‌های دیگر آن برای تف‌م‌س در هیدرولوژی توسط ویلشایر (31)، برن (9)، بهاسکار و اُ-کانل (8) و برن و گوئل (12) ارایه شده است.

درحالی‌که روش‌های خوشه‌سازی سلسله مراتبی تحت تاثیر شرایط اولیه و کمینه‌های محلی قرار نمی‌گیرند، روش‌های خوشه‌سازی تفکیکی تحت تاثیر حدس‌های اولیه (تعداد خوشه‌ها، مراکز خوشه‌ها) قرار می‌گیرد. روش‌های خوشه‌سازی تفکیکی با این مفهوم پویا هستند که ایستگاه‌ها می‌توانند از خوشه‌ای به خوشه‌ی دیگر منتقل شوند تا تابع هدف را کمینه کند. در مقابل، ایستگاه‌های منسوب شده به یک خوشه در مراحل اولیه در روش‌های خوشه‌سازی سلسله مراتبی نمی‌توانند تغییر مکان دهند. بنابراین هر دو روش خوشه‌سازی با کمبودهایی مواجه هستند.

هدف این مقاله، تلفیق دو روش خوشه‌سازی است. به این‌صورت که خوشه‌سازی تفکیکی به‌دلیل ویژگی پویایی به‌عنوان هسته‌ی اصلی درنظر گرفته می‌شود. سپس از روش‌های مختلف سلسله مراتبی به‌عنوان تشکیل خوشه‌های اولیه استفاده و از آن‌ها برای خوشه‌سازی تفکیکی استفاده خواهد شد. با این شیوه انتظار می‌رود که همگرا شدن خوشه‌سازی نهایی، به حالت مقاوم میل کند.

مواد و روش‌ها

الگوریتم هیبرید: هر یک از N ایستگاه آب‌سنجی در منطقه

دارای m متغیر مشخصه است، به‌طوری‌که x_{ij} مقدار متغیر j- ام از ایستگاه i- ام می‌باشد. ازآن‌جا که محدوده‌ی تغییرات این متغیرها عموماً با هم بسیار متفاوت است، غالباً آن‌ها را مقیاس (مثلاً نرمال یا استاندارد بین صفر و 1 یا 1- تا 1+) می‌کنند. متغیر مقیاس شده‌ی متناظر با متغیر اصلی x_{ij}، y_{ij} نامیده می‌شود. فرض می‌کنیم که در مرحله‌ی مشخصی از فرایند خوشه‌سازی سلسله مراتبی تجمعی، K خوشه تشکیل شده باشد. این K خوشه برای مرحله‌ی آغازین خوشه-سازی در فکم استفاده می‌شود. فکم روش تکراری است به-طوری‌که ایستگاه‌ها از خوشه‌ای به خوشه‌ی دیگر منتقل می‌شوند تا تابع هدف (رابطه‌ی 1) را کمینه کند:

$$OF = \sum_{k=1}^{K}\sum_{j=1}^{m}\sum_{i=1}^{N_k} d^2\left(y_{ij}^{(k)} - y_{\bullet j}^{(k)}\right) \qquad (1)$$

که در آن K تعداد خوشه‌ها، N_k تعداد ایستگاه‌ها در خوشه‌ی k، $y_{ij}^{(k)}$ مقدار استاندارد شده‌ی متغیر j در ایستگاه i مربوط به خوشه‌ی k، $d(\times)$ معیاری مناسب از فاصله در فضای m- بُعدی (مثلاً فاصله‌ی اقلیدوسی) و $y_{\bullet j}^{(k)}$ مقدار میانگین متغیر j برای خوشه‌ی k می‌باشد. با کمینه کردن OF در رابطه‌ی 1، فاصله‌ی هر ایستگاه تا مرکز خوشه‌ای که به آن تعلق دارد کمینه می‌شود.

مقدار بهینه‌ای که توسط تابع هدف OF به‌دست می‌آید به موقعیت مراکز اولیه‌ی خوشه‌ها در فکم بستگی دارد. هیچ روش یکتایی برای تعریف شرایط اولیه‌ی مراکز خوشه‌ها که منجر به کمینه‌ی سراسری شود وجود ندارد. بنابراین چندین روش برای ایجاد شرایط اولیه استفاده می‌شود. ویلشایر (31) شرایط اولیه برای خوشه-سازی را با تقسیم داده‌ها به‌طور تصادفی درنظر گرفت. بهاسکار و اُ-کانر (8) ایستگاه‌هایی را برای شرایط اولیه‌ی خوشه‌سازی درنظر گرفتند که فاصله‌ی آن‌ها با هم از یک حداقلی بیش‌تر باشند. برن (9) K ایستگاه از N ایستگاه را به‌عنوان مراکز خوشه‌ها به‌گونه‌ای درنظر گرفت که هر خوشه دست‌کم یک عضو داشته باشد. در پژوهش حاضر نتایج به‌دست آمده از سه الگوریتم پیوند تکی، پیوند کامل و وارد (از الگوریتم‌های خوشه‌سازی سلسله مراتبی) برای شرایط اولیه‌ی خوشه-سازی در فکم درنظر گرفته شد.

هر ایستگاه در بین K خوشه به خوشه‌ای منتسب می‌شود که فاصله‌ی آن تا مرکز آن خوشه از فاصله‌اش تا سایر مراکز خوشه‌ها کمتر باشد. پس از به پایان رسیدن فرایند انتساب تمامی ایستگاه‌ها، مراکز تمامی خوشه‌ها روزآمد شده و مقدار تابع هدف OF مجدداً محاسبه می‌شود. در گام بعد مجدداً انتساب تمامی ایستگاه‌ها روزآمد می‌شود. این فرایند تا جایی تکرار می‌شود که مقدار تابع هدف در دو گام متوالی تغییر نکند.

الگوریتم‌های پیوند تکی و پیوند کامل: این الگوریتم‌ها با N خوشه که هرکدام یک عضو دارند آغاز می‌شود. سپس از بین تمامی ترکیبات بین هر دو خوشه، دو خوشه‌ای که فاصله‌ی آن‌ها کم‌تر از سایر فاصله‌ها باشد در هم ادغام می‌شوند. در الگوریتم پیوند تکی، فاصله‌ی بین خوشه‌ی ترکیب شده و یکی از خوشه‌های منفرد دیگر به‌صورت کوچک‌ترین فاصله بین هرکدام از اعضای خوشه‌ی ترکیب شده و آن خوشه‌ی منفرد درنظر گرفته می‌شود. از طرف دیگر در الگوریتم پیوند کامل، فاصله‌ی بین خوشه‌ی ترکیب شده و هر خوشه‌های منفرد دیگر به‌صورت بیش‌ترین فاصله بین هرکدام از اعضای خوشه‌ی ترکیب شده و خوشه‌ی منفرد درنظر گرفته می‌شود. در هر گام هر دو خوشه‌ای که حداقل فاصله را با هم داشته باشند در هم ادغام می‌شوند. در نتیجه، در هر گامِ زمانی از تعداد خوشه‌ها یکی کم می‌شود. پایان الگوریتم زمانی است که تعداد خوشه‌ها برابر با تعداد خوشه‌های از پیش فرض شده برسد.

الگوریتم وارد (W): فرایند وارد (30) نیز با خوشه‌های تک عضوی آغاز می‌کند. برای شروع، هر مرکز خوشه منطبق بر یک ایستگاه است. بنابراین مقدار تابع هدف (رابطه‌ی 2) نیز صفر خواهد بود. در هر گام، تمامی ترکیب‌های دوتایی درنظر گرفته شده و آن ترکیبی انتخاب می‌شود که منجر به کم‌ترین افزایش در تابع هدف (مجموع مربعات انحراف ایستگاه‌ها از مراکز خوشه‌های مربوطه‌شان) شود. این دو خوشه درهم ادغام می‌شوند. الگوریتم وارد منجر به خوشه‌هایی تقریباً کروی و با تعداد اعضای تقریباً برابر می‌شود. هاسکینگ و والیس (3) این ویژگی را برای تشکیل نواحی همگن در فرایند تحلیل ناحیه‌ای مناسب می‌داند.

$$OF_W = \sum_{k=1}^{K}\sum_{j=1}^{m}\sum_{i=1}^{N_k}\left(y_{ij}^{k} - y_{\bullet j}^{k}\right)^2 \qquad (2)$$

اعتبارسنجی خوشه‌ها: از نمایه‌های اعتبارسنجی به‌طور گسترده به منظور تشخیص تعداد بهینه‌ی خوشه‌ها (K) در مجموعه‌ای از داده‌ها استفاده می‌شود (20). در این پژوهش کارایی روش هیبرید برای الگوریتم‌های خوشه‌سازی برای تفمس، از چهار نمایه‌ی ارزیابی خوشه‌ها یعنی ضریب همبستگی کوفنتیک (28)، متوسط عرض سیلهوت (25)، نمایه‌ی دان (18) و نمایه‌ی دیویس- بولدین (16) استفاده گردید.

ضریب همبستگی کوفنتیک (ضک) معیار اعتبارسنجی برای الگوریتم‌های خوشه‌سازی سلسله مراتبی است. مراحل انجام این الگوریتم در ساختار درخت- مانندی به نام دندوگرام نمایان می‌شود. ضک برای کمّی کردن نمایش دندوگرام در فضای دوبُعدی به‌کار رفته و مقدار آن بین صفر و یک تغییر می‌کند. هرچه مقدار آن به یک نزدیک‌تر باشد، بیانگر این است که خوشه‌سازی موفقیت‌آمیزتر است. برای محاسبه‌ی ضک از نرم‌افزار MATLAB استفاده شد. عرض سیلهوت (25) برای هر ایستگاه معیاری مقایسه‌ای است و نشان می‌دهد که آیا بهتر است که این ایستگاه در خوشه‌ای که در آن قرار

دارد باقی بماند یا این‌که به خوشه‌ی دیگری منتقل شود. عرض سیلهوت برای ایستگاه i- ام در خوشه‌ی k- ام برابر است با:

$$SW_i = \frac{O_i - I_i}{\max\{I_i, O_i\}} \quad (3)$$

که در آن I_i متوسط فاصله‌ی ایستگاه i- ام تا تمامی ایستگاه‌های خوشه‌ی k-ام و O_i حداقل فاصله بین ایستگاه i- ام تا سایر خوشه‌ها (فاصله‌ی یک ایستگاه تا خوشه‌ای که به آن تعلق ندارد متوسط فاصله‌ی آن ایستگاه تا تمامی ایستگاه‌های آن خوشه است) می‌باشد. از این‌رو مقدار s(i) بین 1+ و 1- خواهد بود. هرچه مقدار SW_i به 1+ نزدیک‌تر باشد نشان‌دهنده‌ی درستی تعلق ایستگاه i- ام به خوشه‌ای که در آن قرار دارد است. در مقابل، نزدیک‌تر بودن به 1- نشان از انتساب نادرست این ایستگاه به خوشه می‌دهد. بر این اساس، متوسط عرض سیلهوت (مع‌س) میانگین تمامی عرض‌های سیلهوت خواهد بود. بنابراین خوشه‌سازی بهینه است که از بیشینه‌ی متوسط عرض سیلهوت برخوردار باشد. نمایه‌ی دیویس- بولدین (16) با رابطه‌ی 4 داده شده (K تعداد خوشه‌ها، S میانگین فاصله‌ی اقلیدوسی بین مرکز ثقل خوشه‌ی مدنظر تا تمامی ایستگاه‌های آن و d_{jk} فاصله‌ی اقلیدوسی بین مراکز خوشه‌های j و k) و در حقیقت پراکنش درون- خوشه‌ای را با پراکنش بین- خوشه‌ای مقایسه می‌کند. هرچه مقدار این نمایه کوچک‌تر باشد (صورت کسر کوچک و مخرج آن بزرگ) بیانگر مجموعه خوشه‌های متراکمی است که به‌خوبی از هم مجزا شده باشند. نمایه‌ی دان (18) با رابطه‌ی 5 داده می‌شود [$D(C_i, C_j)$ فاصله‌ی بین دو خوشه‌ی C_i و C_j (حداکثر فاصله بین ایستگاهی در خوشه‌ی C_i تا ایستگاهی در خوشه‌ی C_j) و $D(C_k)$ فاصله‌ی درون- خوشه‌ای برای خوشه‌ی C_k (حداکثر فاصله بین دو ایستگاه در خوشه‌ی C_k)]، تعداد بهینه‌ی خوشه‌ها متناظر با بیشینه‌ی نمایه‌ی دان می‌باشد.

$$DB = \frac{1}{K}\sum_{k=1}^{K}\max_{j,\, j\neq k}\left\{\frac{S_k + S_j}{d_{jk}}\right\} \quad (4)$$

$$D_n = \min_{i=\{1,\ldots,K\}}\left\{\min_{j=\{1,\ldots,K\}, j\neq i}\left\{\frac{D(C_i, C_j)}{\max_{k=\{1,\ldots,K\}} D(C_k)}\right\}\right\} \quad (5)$$

آزمون همگنی ناحیه‌ای: هاسکینگ و والیس (3 و 21) روش‌هایی را بر پایه‌ی ویژگی‌های گشتاورهای خطی برای آزمون‌های همگنی و ناهمگنی مجموعه‌ای از نواحی محتمل که از تحلیل خوشه‌ای منتج شده است ارایه داده‌اند. در این پژوهش از این روش‌ها استفاده شد. عمده‌ترین معیار ناهمگنی بر پایه‌ی گشتاورهای خطی، ضریب تغییرات (L-CV) بنانهاده شده است:

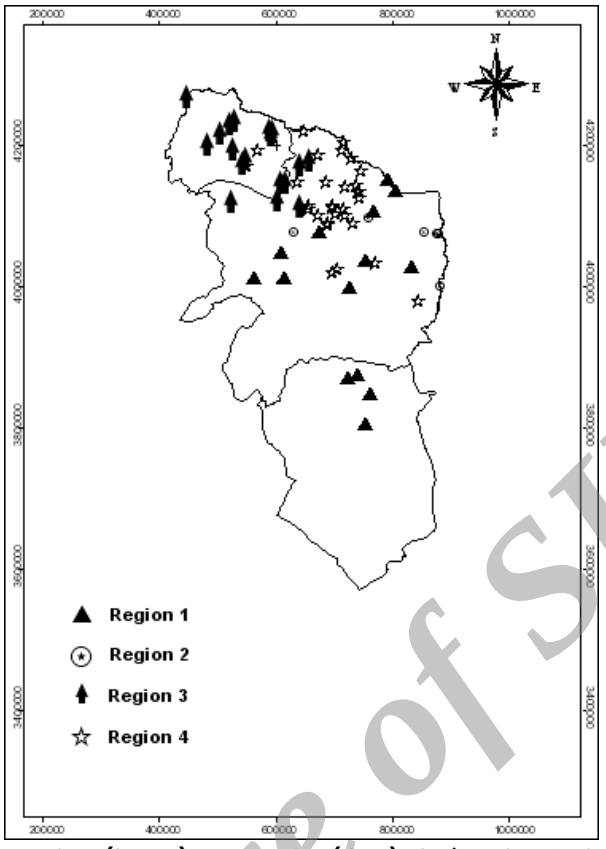
$$H = \frac{(V - \mu_v)}{\sigma_v} \quad (6)$$

که در آن V انحراف استاندارد وزن‌دار شده‌ی L-CVها بر مبنای داده‌های مقیاس شده، μ_v میانگین داده‌های شبیه‌سازی شده به حجم N_{sim} (غالباً 500) و σ_v انحراف استاندارد مقادیر شبیه‌سازی شده‌ی آن‌ها می‌باشند. جزییات کار به‌خوبی در صفحات 80-76 از هاسکینگ و والیس (3) توضیح داده شده است. نواحی «کاملاً همگن»، «نسبتاً همگن» و «کاملاً غیرهمگن» به‌ترتیب با H<1، 1<H<2 و H>2 مشخص می‌شود.

ایستگاه‌ها ویژگی‌های آن‌ها: از 68 ایستگاه آب‌سنجی متناظر با 68 حوضه‌ی آبریز استان‌های خراسان شمالی، رضوی و جنوبی در شرق و شمال شرق ایران (شکل 1) استفاده شد. این سه استان مساحتی درحدود 313000 km^2 را دربر داشته و اقلیم آن خشک و نیمه خشک است. بیش‌ترین ارتفاع برابر با 3300 m در قله‌ی بینالود و کم‌ترین ارتفاع آن 250 m در دشت سرخس واقع شده است. تنوع اقلیمی و تغییرپذیری مکانی و زمانی بارندگی در منطقه ناشی از وجود بیابان‌ها و قلل مرتفع بوده و بارندگی به‌طور کلی از شمال- غرب به جنوب- شرق کاهش می‌یابد. داده‌ها از شرکت‌های سهامی آب منطقه‌ای سه استان خراسان تهیه شدند. برای خوشه‌بندی برای هر حوضه‌ی آبریز متغیرهایی همچون ارتفاع و طول و عرض جغرافیایی مرکز ثقل، مساحت، ضریب شکل، شیب و طول آبراهه‌ی اصلی تهیه شد. شامکوئیان و همکاران (1) ت‌ف‌م‌س را برای سه استان خراسان شمالی، رضوی و جنوبی انجام دادند. پس از آن قهرمان و داوری (2) ت‌ف‌م‌س را با روزآمد کردن آمار سیلاب خراسان رضوی تکرار کردند. در این‌جا آمار سیلاب دو استان خراسان شمالی و جنوبی نیز روزآمد گردید.

نتایج و بحث

تحلیل خوشه‌ای: به ازاء هر تعدادی از خوشه‌ها (2 تا 10)، شرایط اولیه برای ف‌ک‌م به کمک روش تحلیل خوشه‌ای تجمعی سلسله مراتبی به‌دست آمد. به‌طور کلی با افزایش تعداد خوشه‌ها مقدار تابع هدف (رابطه‌ی 1) کاهش یافت. بیشینه‌ی مقدار زمانی است که تمامی ایستگاه‌ها در خوشه‌ای واحد قرار گرفته و کمینه‌ی آن (صفر) زمانی است که K برابر با تعداد ایستگاه‌ها باشد. شکل 2 مقدار تابع هدف (رابطه‌ی 1) را برای ترکیب‌های مختلفی از هیبرید کردن ف‌ک‌م با روش‌های سلسله مراتبی پیوند تکی (SL)، پیوند کامل (CL) و وارد (W) نشان می‌دهد. هیچ‌کدام از روش‌های سلسله مراتبی به تنهایی به کمینه‌ی تابع هدف منجر نشدند. با این حال روش W نسبت به دو روش دیگر تفاوتی محسوس داشت (31/3 درصد کم‌تر از پیوند تکی و 4/8 درصد کم‌تر از پیوند کامل- دربرگیرنده‌ی K بین 3 تا 10)، برتری روش وارد (W) در توزیع ایستگاه‌ها در خوشه‌ها نیز به چشم می‌خورد.



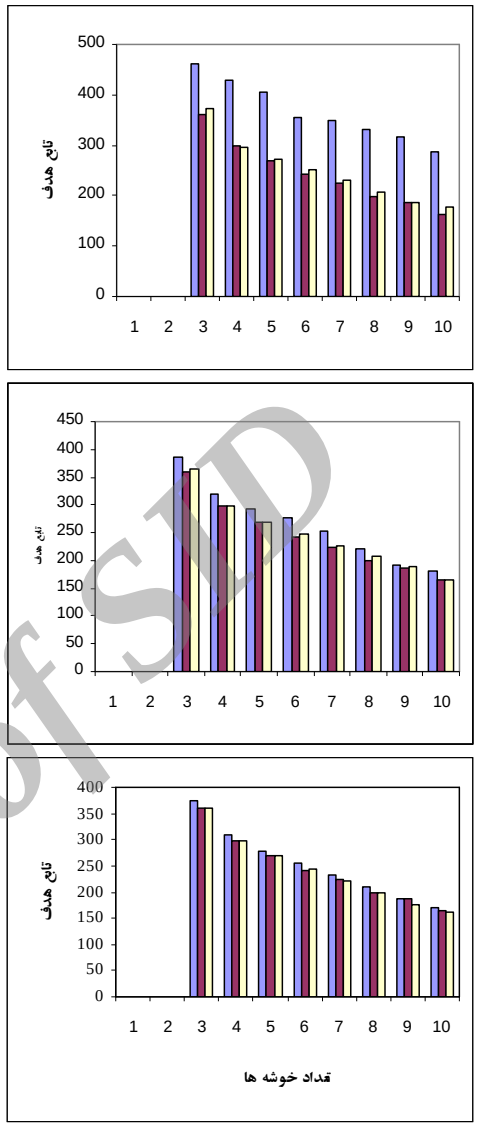
شکل 1- ایستگاه‌های آب‌سنجی در سه استان خراسان شمالی (در بالا)، خراسان رضوی (در وسط) و خراسان جنوبی (در پایین) و ناحیه‌بندی آن‌ها به 4 ناحیه

به‌طور کلی خوشه‌هایی که با پیوند تکی (و تا حدی پیوند کامل) به‌دست آمدند نامتجانس بود. یک خوشه‌ی بزرگ و تعدادی خوشه‌ی کوچک که نشان می‌دهد برای ناحیه‌بندی تناسب ندارد. در حالی‌که ایستگاه‌ها در روش W به‌خوبی توزیع شده بودند.

برتر بودن روش وارد از دیگر روش‌های سلسله مراتبی در هیبرید کردن با ف‌کم نیز به چشم می‌خورد؛ گرچه تفاوت چشم‌گیر نبود (3/2 درصد کم‌تر از پیوند تکی و 1/8 درصد کم‌تر از پیوند کامل). کارایی هرکدام از سه روش هیبریدی در کمینه کردن تابع هدف (رابطه‌ی 1) بهتر از عملکرد آن‌ها به تنهایی است (32 درصد بهبود برای پیوند تکی، 7/2 درصد بهبود برای پیوند کامل و 4/3 درصد برای وارد). هیبرید کردن نه تنها منجر به افزایش کارایی روش‌های سلسله مراتبی شد بلکه کارایی ف‌کم را نیز بهبود بخشید. با این حال تنها هیبرید کردن با روش W کارایی ف‌کم را بهبود بخشید (0/6 درصد) در حالی‌که انتخاب تصادفی برای مراکز خوشه‌ها در ف‌کم از هیبرید کردن آن با دو روش دیگر سلسله مراتبی (پیوند تکی و کامل) بهتر بود (2/6 درصد برای پیوند تکی و 1/2 درصد برای پیوند کامل).

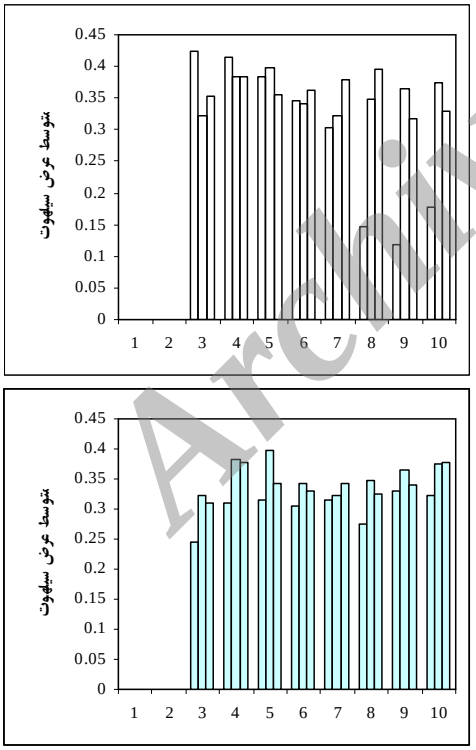
گرچه هیبرید کردن با روش W کارایی ف‌کم را بهبود بخشید ولی این بهبود به تعداد خوشه‌ها بستگی دارد: هیبرید کردن برای خوشه‌های 5 و 6 منجر به افزایش تابع هدف شد گرچه افزایش چشم‌گیر نبود (0/3 درصد برای K=5 و 1/2% برای K=6). بنابراین به‌نظر می‌رسد که هیبرید کردن همواره منجر به کاهش تابع هدف نشود و احتمالاً باید آن‌را تنها *پتانسیلی* برای بهبود کارایی دانست.

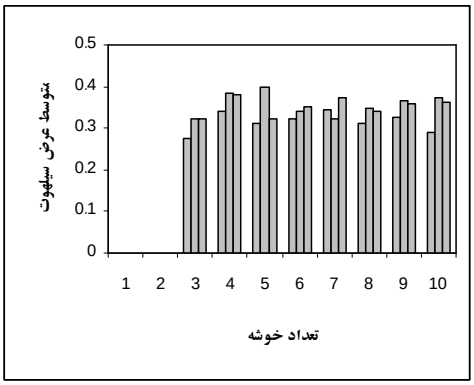
اعتبارسنجی خوشه‌ها. برای تعیین تعداد بهینه‌ی خوشه‌ها از نمایه‌های اعتبارسنجی استفاده شد. ضریب همبستگی کوفنتیک (ض‌ه‌ک) برای روش پیوند تکی (0/803) به‌طور مشخص بیش‌تر از دو روش پیوند کامل (0/655) و وارد (0/674) بود. با توجه به تعریف ض‌ه‌ک، می‌توان نتیجه‌گیری کرد که در مقایسه با روش‌های پیوند کامل و وارد، روابط چند بُعدی موجود در داده‌ها به‌گونه‌ای بهتر در ساختار درختی (دندوگرام) که روش پیوند تکی ارایه می‌دهد نمایان شده است. این نتیجه در راستای نتیجه‌ی پیش مبنی بر عملکرد ضعیف روش پیوند تکی در مقدار تابع هدف نمی‌باشد.



شکل 2- مقدار تابع هدف به‌عنوان تابعی از تعداد خوشه‌ها برای روش‌های خوشه‌بندی پیوند تکی (بالا)، پیوند کامل (وسط) و وارد (پایین). ستون‌ها در هر مقدار از تعداد خوشه به‌ترتیب از چپ به راست مربوط است به روش مشخص شده‌ای از سلسله مراتبی، K- میانگین و هیبرید این دو روش.

گرچه شاخصی برای مقایسه‌ی این دو نتیجه‌ی متفاوت وجود ندارد، ولی ما به این نکته اشاره می‌کنیم که روش پیوند تکی منجر به توزیع نامناسب (چوله) ایستگاه‌ها شده و بنابراین منطقاً برای ت‌ف‌م‌س فاقد شایستگی به شمار می‌آید. تغییرات متوسط عرض سیلهوت (مع-س) که دامنه‌ی مجاز آن بین 1- و 1+ است برای ترکیب‌های متفاوتی از روش‌های خوشه‌بندی محاسبه گردید (شکل 3). تمامی متوسط عرض‌های سیلهوت مثبت هستند که بدان مفهوم است که تمامی ایستگاه‌ها به خوشه‌های مناسب منتسب شده‌اند. مع‌س برای روش خوشه‌بندی پیوند تکی برای تعداد خوشه‌های 2 تا 6، برای پیوند کامل برای تعداد خوشه‌های بین 9 تا 10 و برای روش وارد برای تعداد خوشه‌های بین 7 تا 8 بیشینه بود. ولی همان‌طور که پیش از این نیز بیان شد دو روش پیوند تکی و پیوند کامل به‌دلیل ایجاد خوشه‌های نامتجانس مناسب نیستند. تقریباً برای کلیه‌ی حالات (بجز K=2) روش‌های هیبریدی افزایش مع‌س نسبت به روش سلسله مراتبی وجود داشت که نشان‌دهنده‌ی بهبود در کارایی به‌دلیل هیبرید است. در بین روش‌های هیبریدی، ترکیب SL+KM بیش‌ترین مع‌س را داشت (برای K≤8) که به‌دلیل ایجاد خوشه‌های نامتجانس برای تحلیل ت‌ف‌م‌س مناسب نمی‌باشد. مابین نتایج، هیبرید W+KM برای K=4 بالاترین مع‌س را دارا بود. بر اساس جدول 1 با افزایش K کارایی هر دو نمایه‌ی D و DB افزایش می‌یابد. این امر به‌ویژه تا K=6 نمایان‌تر است. با افزایش K به بیش‌تر از 6، هر دو نمایه نوسان می‌کنند. بنابراین در محدوده‌ی K≤6 به نظر می‌رسد که K=4 برای هیبرید W+KM بهتر باشد: D بیشینه و DB کمینه می‌شود. نمایه‌ی دان و نمایه‌ی دیوید- بولدین نیز نشان می‌دهند که هیبرید وارد و روش K- میانگین بهتراند (جدول 1).





شکل 3- اعتبارسنجی خوشه‌بندی با استفاده از متوسط عرض سیلهوت برای روش‌های خوشه‌بندی پیوند تکی (بالا)، پیوند کامل (وسط) و وارد (پایین). ستون‌ها در هر مقدار از تعداد خوشه به‌ترتیب از چپ به راست مربوط است به روش مشخص شده‌ای از سلسله مراتبی، K- میانگین و هیبرید این دو روش.

معیارهای ناهمگنی هاسکینگ و والیس (21) (رابطه‌ی 6) برای آزمون این‌که خوشه‌های منتج از هیبرید وارد و فک‌م برای K=4 از نظر آماری همگن هستند به‌کار برده شد. نتایج آزمون همگنی در جدول 2 ارایه شده است. تمامی مقادیر منفی هستند که ممکن است دالّ بر این باشد که آمار سیلاب ایستگاه‌ها با هم همبسته باشند (3). کاسترلارین و همکاران (14) به چندین آزمون همگنی اشاره کرده‌اند که در آن‌ها همبستگی سیلاب در ایستگاه‌های ناحیه ناچیز نمی‌باشد ولی به‌طور مشخص راه‌کاری برای درنظر گرفتن این همبستگی را ارایه نکرده‌اند. نامبردگان موثر بودن روش شناخته شده‌ی آزمون همگنی را در شرایطی که ایستگاه‌ها با هم همبستگی داخلی داشته باشند را در یک‌سری آزمایشات مونت کارلو تحلیل کرده‌اند. آن‌ها فرمول تصحیحی تجربی زیر را برای آزمون همگنی برپایه‌ی H_1 به منظور لحاظ کردن تقریبی همبستگی داخلی ارایه دادند:

$$H_{1,adj} = H_1 + C \times \rho^2 (R - 1) \qquad (7)$$

که در آن $H_{1,adj}$ مقدار تصحیح شده‌ی معیار ناهمگنی، H_1 مقدار به‌دست آمده از آزمون همگنی، C ضریب تصحیح تجربی که ثابت فرض می‌شود، ρ^2 میانگین ضرایب تعیین سیلاب در ناحیه و R تعداد ایستگاه‌های ناحیه است. مقادیر ρ^2 برای 4 ناحیه به‌ترتیب 0/11، 0/05، 0/14 و 0/08 به‌دست آمد. با اتخاذ C=0.122 (14)، مقادیر $H_{1,\ adj}$ محاسبه و در جدول 2 آورده شد. مقادیر تصحیح شده بیانگر این است که تمامی خوشه‌ها (نواحی) تقریباً همگن هستند.

جدول 1- نمایه‌ی دان و نمایه‌ی دیویس- بولدین (اعداد در پرانتز) برای اعتبارسنجی خوشه‌بندی در سه روش متفاوت هیبرید K- میانگین

تعداد خوشه	پیوند تکی	پیوند کامل	وارد
2	0/4000 (1/6700)	0/5000 (1/6420)	0/6000 (1/6000)
3	0/4817 (1/5348)	0/5564 (1/5392)	0/7498 (1/5000)
4	0/6667 (1/3380)	0/8800 (1/3830)	1/4000 (1/1140)
5	0/5872 (1/3732)	0/8502 (1/3165)	0/8591 (1/3568)
6	0/5872 (1/3223)	0/6319 (1/3370)	0/9083 (1/3530)
7	1/1591 (1/2292)	0/6319 (1/6054)	1/0900 (1/2096)
8	2/0228 (1/1556)	1/0534 (3/1018)	0/7799 (1/1868)
9	0/3619 (2/2051)	0/9362 (3/1092)	0/9021 (1/1751)
10	0/5603 (1/2681)	0/5206 (2/9741)	1/0800 (1/1156)

جدول 2- معیار ناهمگنی برای خوشه‌بندی ترکیبی هیبرید وارد و K- میانگین برای K=4.

شماره‌ی ناحیه	تعداد ایستگاه‌ها	تعداد ایستگاه-سال	آماره‌ی H1
1	14	293	-0/96 (-0/78)
2	5	142	-0/09 (-0/96)
3	19	381	-0/89 (-0/61)
4	30	725	-0/98 (-0/80)

اعداد داخل پرانتز مربوط به تصحیح ناشی از همبسته بودن آمار سیلاب است

موقعیت ایستگاه‌های قابل قبول و نواحی چهارگانه در شکل 1 ارایه شده است. این شکل نشان می‌دهد که امکان تداخل جغرافیایی بین ایستگاه‌های نواحی مختلف وجود دارد. با این حال این رفتار در هیدرولوژی متداول است. رائو و سری‌نیواس (24) چنین گسترشی را برای ت‌ف‌م‌س در ایندیانا واقع در آمریکا نشان دادند. نتیجه‌ی مشابهی توسط لین و همکاران (22) برای ناحیه‌بندی ایستگاه‌های باران‌سنج ثبات در بررسی الگوی توزیع زمانی باران در قسمت میانی تایلند گزارش شده است.

نتیجه‌گیری

الگوریتم‌های خوشه‌بندی هیبریدی که ترکیبی از روش‌های خوشه‌بندی سلسله مراتبی تجمعی و نیز تفکیکی هستند، مورد استفاده قرار گرفت و برای ناحیه‌بندی حوضه‌های آبریز به منظور تحلیل فراوانی سیلاب به‌کار رفت. الگوریتم‌های خوشه‌بندی سلسله مراتبی به‌کار رفته برای هیبرید کردن، الگوریتم‌های پیوند تکی، پیوند کامل و وارد بوده در حالی‌که الگوریتم خوشه‌بندی تفکیکی، K- میانگین درنظر گرفته شد. از سه مدل هیبرید ارایه شده، ترکیب روش وارد و ف‌ک‌م به‌طور منطقی تخمین‌های اولیه‌ی خوبی را برای گروه‌بندی حوضه‌های آبریز ارایه داد. در فرایند خوشه‌بندی هیبرید این توقع وجود دارد که مدل خوشه‌بندی سلسله مراتبی مقادیر اولیه‌ی معنی‌دارتری برای ف‌ک‌م بدهد به‌طوری‌که نتیجه‌ی آن معنی‌دارتر و بهتر باشد. با این حال نمی‌توان تضمین کرد که خوشه‌بندی هیبرید منجر به نتیجه‌ی بهتری از ف‌ک‌م شود. به بیان دیگر به‌نظر می‌رسد که باید به خوشه‌بندی هیبرید به‌عنوان گزینه‌ای «پتانسیل» برای شرایط اولیه در ف‌ک‌م نگریست و همواره منجر به بهترین نتیجه در بین تمامی گزینه‌ها برای شرایط اولیه در آن نمی‌شود. چهار معیار اعتبارسنجی خوشه‌ای یعنی ضریب همبستگی کوفنتیک، متوسط عرض سیلهوت، نمایه‌ی دان و نمایه‌ی دیویس- بولدین برای تعیین موثر بودنشان در تشخیص تفکیک بهینه ناشی از استفاده از الگوریتم‌های خوشه‌بندی آزمون شد. ض‌ه‌ک ناکارامد بود در حالی‌که کارایی م‌ع‌س تقریباً خوب بود. نشان داده شد که نمایه‌ی دان و نمایه‌ی دیویس- بولدین در تشخیص تفکیک بهینه کارا بوده و خوشه‌های تقریباً همگنی را تشکیل دادند. با استفاده از این نمایه‌های اعتبارسنجی، 4 خوشه از هیبرید وارد و خوشه‌بندی ف‌ک‌م به‌عنوان خوشه‌بندی بهینه درنظر گرفته شد. از معیارهای ناهمگنی هاسکینگ و والیس (21) برای آزمون این‌که آیا این خوشه‌ها از نظر آماری همگن هستند استفاده شد. نتایج نشان داد که تمامی نواحی تقریباً همگن هستند. در نهایت این خوشه‌ها را می‌توان برای ت‌ف‌م‌س به‌کار برد.

وقتی بخواهیم مقادیر سیلاب طراحی را برای افق‌های زمانی آینده محاسبه کنیم وجود ناایستایی معنی‌دار در سری زمانی هیدرولوژیکی، و البته تغییر اقلیم، را نمی‌توان نادیده گرفت. پژوهش‌های متعددی در منابع به چنین مفاهیمی در هیدرولوژی پرداخته‌اند (5، 11، 14 و 26). با این حال هیچ‌کدام از پژوهش‌های منتشر شده به مقوله‌ی ناحیه‌بندی نپرداخته‌اند. بنابراین یک خلاء در منابع علمی وجود دارد که نیاز به بررسی دقیق دارد.

منابع

1- شامکوئیان ح.، قهرمان ب.، داوری ک. و سرمد م. 1388. تحلیل فراوانی سیلاب منطقه‌ای با استفاده از تئوری گشتاورهای خطی و سیلاب نمایه در حوضه‌های آبریز استان‌های خراسان. مجله آب و خاک (علوم و صنایع کشاورزی)، 23(1)، 31-43.

2- قهرمان ب. و داوری ک. 1388. استفاده از گشتاورهای خطی در تحلیل منطقه‌ای سیلاب در خراسان رضوی. شرکت سهامی آب منطقه‌ای خراسان رضوی. 88 صفحه.

3- هاسکینگ جی.آر.ام. و والیس جی.آر. 1392. تحلیل فراوانی ناحیه ای (نگرشی بر پایه گشتاورهای خطی). (مترجم: بیژن قهرمان) انتشارات طنین قلم، مشهد. 276 صفحه.

4- نیرومند ح.ع. 1378. تحلیل آماری چندمتغیره کاربردی. دانشگاه فردوسی مشهد.

5- Abdul Aziz O.I. and Burn D.H. 2006. Trends and variability in the hydrological regime of the Mackenzie River Basain, Journal of Hydrology, 319: 282-294.

6- Abida H. and Ellouze M. 2006. Hydrological delineation of homogeneous regions in Tunisia, Water Resources Management, 20: 961-977.

7- Atiem I. and Harmancloglu N.B. 2006. Assessment of regional floods using L-moments approach: the case of the River Nile, Water Resources Management, 20: 723-747.

8- Bhaskar N.R. and O'Connor C.A. 1989. Comparison of method of residuals and cluster analysis for flood regionalization, Journal of Water Resources Planning and Management, ASCE, 115(6): 793–808.

9- Burn D.H. 1989. Cluster analysis as applied to regional flood frequency, Journal of Water Resources Planning and Management, 115(5): 567–582.

10- Burn D.H. 1990. Evaluation of regional flood frequency analysis with a region of influence approach, Warer

Resources Research, 26(10): 2257-2265.
11- Burn D.H. and Elnur A.H. 2002. Detection of hydrologic trends and variability, Journal of Hydrology, 255: 107-122.
12- Burn D.H. and Goel N.K. 2000. The formation of groups for regional flood frequency analysis, Hydrological Sciences Journal, 45(1): 97–112.
13- Burn D.H., Zinji Z. and Kowalchuk M. 1997. Regionalization of catchments for regional flood frequency analysis, Journal of Hydrologic Engineering, ASCE, 2(2): 76-82.
14- Casterllarin A., Burn D.H., and Brath A. 2008. Homogeniety testing: how homogeneous do heterogeneous cross-correlated regions seem?, Journal of Hydrology, 360: 67-76.
15- Cunderlik J.M. and Burn D.H. 2003. Non-stationary pooled flood frequency analysis, Journal of Hydrology, 276: 210-223.
16- Davies, D.L. and D.W. Bouldin. 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1: 224–227.
17- Dinpashoh, Y., A., Fakheri-Fard, M., Moghaddam, S.Jahanbakhsh and M. Mirnia. 2004. Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods, Journal of Hydrology, 297: 109-123.
18- Dunn J.C. 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics, 3: 32–57.
19- Eng K., Milly P.C.D. and Tasker G.D. 2007. Flood regionalization: a hybrid geographic and predictor-variable region-of-influence regression method, Journal of Hydrologic Engineering, ASCE, 12(6): 585-591.
20- Halkidi M., Batistakis Y. and Vazirgiannis M. 2001. On clustering validation techniques, Journal of Intelligent Information systems, 17 (2/3): 107–145.
21- Hosking J.R.M. and Wallis J.R. 1993. Some statistics useful in regional frequency analysis, Water Resources Research, 29 (2): 271–281 (Correction: Water Resources Research 31(1): 251, 1995).
22- Lin G.F., Chen L.H., and Kao S.C. 2005. Development of regional design hyetographs, Hydrological Processes, 19: 937-946.
23- MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. In: Le Cam, L.M., Neyman, J. (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. University of California Press, Berkeley, CA, pp. 281–297.
24- Rao A.R. and Srinivas V.V. 2006. Regionalization of watersheds by hybrid-cluster analysis, Journal of Hydrology, 318: 37-56.
25- Rousseeuw P.J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, 20: 53–65.
26- Sharif M. and Burn D.H. 2006. Simulating climate change scenarios using an improved K-nearest neighbor model, Journal of Hydrology, 325: 179-196.
27- Shu C. and Burn D.H. 2004. Homogenous pooling delineation for flood frequency analysis using a fuzzy expert system with genetic enhancement, Journal of Hydrology, 291: 132.-149.
28- Sokal R.R. and Rohlf F.J. 1962. The comparison of dendrograms by objective methods, Taxonomy, 11: 33–40.
29- Wagner T., Sivapalan M., Troch P., and Woods R. 2007. Catchment classification and hydrologic similarity, Geography Compass, 1(4): 901-931, doi:_10.1111/j.1749-8198.2007.00039.x.
30- Ward Jr., J.H. 1963. Hierarchical grouping to optimize an objective function, Journal of American Statistical Association, 58: 236-244.
31- Wilshire S.E. 1986. Regional flood frequency analysis. II. Multivariate classification of drainage basins in Britain, Hydrological Sciences Journal, 31(3): 335-346.

Journal of Water and Soil
Vol. 28, No. 3, Jul.-Agu. 2014, p. 471-480

نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد 28، شماره 3، مرداد – شهریور 1393، ص. 480-471

Adopting Hierarchial Cluster Analysis to Improve The Performance of K-mean Algorithm

B. Ghahraman[1*] **- K. Davary**[2]
Received:09-04-2013
Accepted:17-02-2014

Abstract

Due to inadequate flood data it is not always possible to fit a frequency analysis to at-site stations. Reliable results are not always guaranteed by a single clustering algorithm, so a combination of methods may be used. In this research, we considered three clustering algorithms: single linkge, complete linkage and Ward (as hierarchial clustering methods), and K-mean (as partitional clustering analysis). Hybrid cluster analysis was tested for up-to-dated of floods data in 68 hydrometric stations in East and NE of Iran. Four cluster validity indices were used to find the optimum number of clusters. Based on the Cophenetic coefficient and average Silhouette width, single linkge, and complete linkage methods were performed well, yet they produced non-consistent clusters (one large and numerous small clusters) which are not amenable for flood frequency analysis. It was shown that hybridization was efficient to form homogeneous regions, however, the usefulness was dependent to the number of classes. Heterogeneity measure of Hosking was negative, due to inter-correlation of floods in the clusters. The hybrid of Ward and K-mean was shown to be the best combination for the region under study. Four homogeneous regions were delineated.

Keywords: Cluster analysis, Hyrid, Khorasan, Linear moments, Regional flood frequency analysis, Regionalyzation

1, 2- Professor and Associate Professor of Water Engineering Department, College of Agriculture, Ferdowsi University of Mashhad
(*- Corresponding Author Email: bijangh@um.ac.ir)