مقاله پژوهشی

نشریه علمی تحقیقات جنگل و صنوبر ایران
جلد ۲۹ شماره ۴، صفحه ۳۲۵-۳۱۵، (۱۴۰۰)

شناسه دیجیتال (DOI): 10.22092/ijfpr.2022.356241.2022
شناسه دیجیتال (DOR): 20.1001.1.17350883.1400.29.4.1.4

مقایسه عملکرد دو روش خوشه‌بندی غیرسلسله‌مراتبی در داده‌های پوشش گیاهی

نغمه پاک‌گهر [۱]، جواد اسحاقی‌راد [۲*]، غلامحسین غلامی [۳]، احمد علیجانپور [۴] و دیوید دابلیو. رابرتز [۵]

۱- دکتری جنگل‌داری، گروه جنگل‌داری، دانشکده منابع طبیعی، دانشگاه ارومیه، ارومیه، ایران

۲*- نویسنده مسئول، استاد، گروه جنگل‌داری، دانشکده منابع طبیعی، دانشگاه ارومیه، ارومیه، ایران. پست الکترونیک: j.eshaghi@urmia.ac.ir

۳- استادیار، گروه ریاضی، دانشکده علوم، دانشگاه ارومیه، ارومیه، ایران

۴- دانشیار، گروه جنگل‌داری، دانشکده منابع طبیعی، دانشگاه ارومیه، ارومیه، ایران

۵- استاد، گروه اکولوژی، دانشگاه ایالتی مونتانا، بووزمن، آمریکا

تاریخ دریافت: ۱۴۰۰/۰۷/۱۷ تاریخ پذیرش: ۱۴۰۰/۰۹/۲۶

چکیده

هدف طبقه‌بندی پوشش گیاهی، بهینه‌سازی و خلاصه کردن تغییرات آن به‌عنوان نماینده تغییرات محیطی است که سبب دسترسی به اطلاعات مفید و قابل‌تفسیر از بوم‌سازگان می‌شود. باتوجه‌به وجود تعداد زیادی از روش‌های طبقه‌بندی، انتخاب روش بهینه، چالشی بزرگ است. در پژوهش پیش‌رو، دو روش خوشه‌بندی غیرسلسه‌مراتبی شامل K-means و K-medoids برای بوم‌سازگان‌های جنگلی مقایسه شدند. داده‌های مورد استفاده در این راستا شامل دو مجموعه داده واقعی جمع‌آوری‌شده از نوشهر (جنگل‌های هیرکانی) و اسلام‌آباد غرب (جنگل‌های زاگرس) و شش مجموعه داده شبیه‌سازی‌شده بودند. برای آماده‌سازی داده‌ها از تبدیل داده هلینگر استفاده شد. سپس، سه روش اندازه‌گیری فاصله اقلیدسی، بری‌کورتیس و منهتن به‌کار گرفته شدند تا عملکرد دو روش غیرسلسله‌مراتبی مذکور بررسی شود. نتایج طبقه‌بندی به‌دست‌آمده از روش‌های مختلف با سه روش ارزیابی‌کننده سیلوئت، همبستگی فی و ISAMIC مقایسه شدند. نتایج نشان داد که ترکیب ماتریس تشابه بری‌کورتیس و روش‌های خوشه‌بندی K-means و K-medoids به‌ترتیب رتبه‌های اول و دوم را در بین خوشه‌بندی‌های مختلف داشتند. ضعیف‌ترین خوشه‌بندی مربوط به ترکیب ماتریس تشابه منهتن و روش K-medoids بود. روش K-means در داده‌های ناهمگن‌تر مانند داده‌های زاگرس و شبیه‌سازی‌شده، کارایی بیشتری داشت. همچنین، تبدیل داده هلینگر سبب بهبود عملکرد ضریب فاصله اقلیدسی شد. باتوجه‌به نتایج تحلیل‌های مربوطه، ترکیب روش خوشه‌بندی K-means و ماتریس تشابه بری‌کورتیس برای داده‌های جوامع گیاهی پیشنهاد می‌شود.

واژه‌های کلیدی: تبدیل داده‌ها، داده شبیه‌سازی‌شده، روش اندازه‌گیری فاصله، کیفیت خوشه‌بندی.

مقدمه

از دیرباز، خوشه‌بندی با داده‌های کمی و کیفی پوشش گیاهی از اهداف اصلی علوم زیستی بوده است و امروزه نیز بخش جداناپذیر علوم گیاهی محسوب می‌شود (Lengyel *et al.*, 2018). شناسایی، تفکیک، توصیف و تهیه نقشه جوامع گیاهی بدون خوشه‌بندی واحدهای گیاهی غیرممکن است (Lengyel *et al.*, 2018). همچنین، نحوه شناسایی ترکیب پوشش گیاهی به انتخاب روش خوشه‌بندی بستگی دارد، بنابراین وجود یک الگوریتم مؤثر برای خوشه‌بندی پوشش گیاهی ضروری است (Aho *et al.*, 2008). خوشه‌بندی عددی با هدف بهینه‌سازی و خلاصه کردن تغییرات پوشش گیاهی انجام می‌شود که نماینده تغییرات

محیطی هستند. خوشه‌بندی، اطلاعات مفیدی از داده‌های چندمتغیره در اختیار کاربران قرار می‌دهد تا با اتکا بر آن، تفسیر درست و منطقی از داده‌های خود داشته باشند (Schmidtlein *et al*., 2010).

در پژوهش Peet و Roberts (۲۰۱۳) با مقایسه روش‌های مختلف خوشه‌بندی عددی مرسوم در بوم‌شناسی گیاهان، سه روش اصلی خوشه‌بندی شامل سلسله‌مراتبی تجمعی، سلسله‌مراتبی مقسمی و غیرسلسله‌مراتبی به‌عنوان روش‌های وابسته به ماتریس تشابه (ماتریس فاصله) شناسایی شدند. اولین خوشه در الگوریتم خوشه‌بندی سلسله‌مراتبی فقط شامل یک نمونه بود. سپس، نمونه‌های مشابه در خوشه‌های یکسان قرار گرفتند و درنهایت، همه نمونه‌ها در یک خوشه جای گرفتند. در مقابل اولین خوشه، الگوریتم خوشه‌بندی سلسله‌مراتبی مقسمی شامل همه نمونه‌ها است. در مرحله‌های بعدی، تقسیم‌بندی نمونه‌ها و قرارگیری نمونه‌های مشابه در خوشه‌های یکسان شروع می‌شود. تقسیم شدن خوشه‌ها تازمانی‌که آن‌ها بسیار کوچک شوند و قابلیت تقسیم و تبدیل به خوشه جدید را نداشته باشند، ادامه می‌یابد (Legendre & Legendre, 2012). بااین‌حال از معایب روش‌های سلسله‌مراتبی می‌توان به تصمیم‌های اولیه خوشه‌بندی مانند ادغام و تقسیم خوشه‌ها اشاره کرد که نتیجه نهایی را تحت تأثیر قرار می‌دهند (Lengyel & Botta-Dukát, 2019). الگوریتم‌های خوشه‌بندی غیرسلسله‌مراتبی هیچ‌گونه ساختار سلسله‌مراتبی ندارند (Roberts, 2015) و به‌طور مستقیم نمونه‌ها را در بین مرکزهای خوشه تعیین‌شده پراکنش می‌دهند. در این روش‌ها، اطلاع از تعداد خوشه‌ها برای خوشه‌بندی ضروری است که می‌توان در هر تکرار با کمینه کردن ناهمگنی در خوشه‌ها، نتایج بهتری به‌دست آورد.

در جوامع گیاهی، استفاده از روش‌های خوشه‌بندی سلسله‌مراتبی مرسوم‌تر است و روش‌های غیرسلسله‌مراتبی کمتر استفاده می‌شوند. Aho و همکاران (۲۰۰۸) با مقایسه روش‌های سلسله‌مراتبی و غیرسلسله‌مراتبی با استفاده از ارزیابی‌کننده‌های مختلف، عملکرد بهتری را در روش‌های خوشه‌بندی غیرسلسله‌مراتبی مشاهده کردند. این الگوریتم‌ها به‌عنوان یکی از روش‌های مؤثر با عملکرد مطلوب برای داده‌های حجیم معرفی شده‌اند (Rodriguez *et al*., 2019). یکی از معروف‌ترین آن‌ها، روش K-means است که به‌دلیل الگوریتم ساده در زمینه‌های مختلف استفاده شده است. این الگوریتم خوشه‌بندی، اغلب در رقابت با روش‌های سلسله‌مراتبی، عملکرد بهتری نشان داده است (Aho *et al.*, 2008). روش K-medoids یکی دیگر از روش‌های خوشه‌بندی غیرسلسله‌مراتبی است. Tichý و همکاران (۲۰۱۴) K-means و K-medoids را کاربردی‌ترین روش‌های غیرسلسله‌مراتبی خوشه‌بندی پوشش گیاهی معرفی کردند. براساس نتایج Roberts (۲۰۱۵) نیز روش‌های مذکور بهتر از روش‌های خوشه‌بندی سلسله‌مراتبی مانند وارد (Ward's) و بتای انعطاف‌پذیر (Flexible-β) هستند. بااین‌حال، هنوز انتخاب روش خوشه‌بندی مناسب براساس ویژگی‌های داده‌ها و یافتن تعداد مناسب خوشه‌ها، یک موضوع چالش‌برانگیز است (Lengyel & Botta-Dukát, 2019). انتخاب روش بهینه خوشه‌بندی، یکی از مسائل مهم در علم جامعه‌شناسی گیاهی محسوب می‌شود (Lengyel *et al*., 2018)، اما به‌دلیل پیچیدگی داده‌های پوشش گیاهی تاکنون پژوهشگران در این مورد اتفاق‌نظر نداشته‌اند. انتخاب روش مناسب خوشه‌بندی در جامعه‌شناسی گیاهی با آشکارسازی الگوهای نهفته در جوامع گیاهی و تولید بالقوه رویشگاه (Janatbabaei *et al.*, 2020) سبب مدیریت بهینه، حفظ منابع طبیعی و تهیه نقشه‌های آمایش سرزمین می‌شود (Lengyel *et al.*, 2018)، بنابراین پژوهش پیش‌رو با هدف ارزیابی عملکرد دو روش مرسوم خوشه‌بندی غیرسلسله‌مراتبی شامل K-means و K-medoids در بوم‌سازگان جنگلی انجام شد. در پژوهش‌های پیشین داده‌های شبیه‌سازی برای ارزیابی روش‌های مختلف طبقه‌بندی کمتر مورد توجه قرار گرفته است در حالی که، شبیه‌سازی داده یک روش مؤثر برای ارزیابی روش‌های جدید و مقایسه روش‌های مختلف محسوب می‌شود (Morris *et al*., 2019). شبیه‌سازی داده‌ها، امکان آگاهی از ویژگی‌های آن‌ها را فراهم می‌کند که

در تصمیم‌گیری بهتر درمورد انتخاب روش خوشه‌بندی بهینه بسیار مؤثر است (Morris *et al*., 2019). در این راستا، Lengyel و همکاران (۲۰۱۸) از داده‌های شبیه‌سازی و واقعی به‌طور هم‌زمان استفاده کردند. باتوجه‌به اهمیت استفاده از داده‌های شبیه‌سازی‌شده، پژوهش پیش‌رو با استفاده از داده‌های واقعی و شبیه‌سازی‌شده انجام شد تا روش‌های K-means و K-medoids به‌طور دقیق‌تری مقایسه شوند.

مواد و روش‌ها

داده‌های واقعی

در این پژوهش از دو مجموعه داده واقعی پوشش گیاهی استفاده شد. مجموعه اول از شش سری متفاوت در حوزه استحفاظی اداره کل منابع طبیعی و آبخیزداری استان مازندران- نوشهر جمع‌آوری شدند (Khanalizadeh *et al*., 2020). از هر سری، یک پارسل مدیریت‌شده و یک پارسل شاهد انتخاب شد (جدول ۱). برای تعیین مرکز قطعه‌نمونه‌ها از روش منظم تصادفی با ابعاد شبکه ۲۰۰×۱۰۰ متر مربع استفاده شد. در مرکز هر قطعه‌نمونه، یک زیرقطعه‌نمونه ۱۰۰ متر مربعی (۱۰×۱۰ متر مربع) برای بررسی پوشش علفی و یک زیرقطعه‌نمونه ۴۰۰ متر مربعی (۲۰×۲۰ متر مربع) برای بررسی پوشش درختی و درختچه‌ای برداشت شد (Eshaghi Rad *et al*., 2009).

مجموعه دوم داده از سه قطعه‌نمونه در جنگل‌های اسلام‌آباد غرب (استان کرمانشاه) جمع‌آوری شدند (جدول ۲). در هر قطعه‌نمونه با استفاده از سه ترانسکت که در فاصله‌های ۲۰۰ متری از هم و در جهت شیب قرار داشتند، پوشش گیاهی نمونه‌برداری شد. اولین ترانسکت به‌صورت تصادفی و ترانسکت‌های بعدی به‌صورت منظم- تصادفی پیاده شدند. نقاط برداشت پوشش گیاهی در هر ترانسکت در فاصله‌های صفر، ۲۵، ۵۰، ۱۰۰ و ۱۵۰ متری قرار داشتند (Eshaghi Rad *et al*., 2014).

میانگین شاخص‌های تنوع شانون‌وینر و تنوع بتا در داده‌های هیرکانی به‌ترتیب ۱/۲۹ و ۰/۵۴ و برای داده‌های زاگرس ۲/۴۵ و ۰/۶۴ به‌دست آمد، بنابراین داده‌های زاگرس، تغییرات گونه‌ای و ناهمگنی بیشتری داشتند.

جدول ۱- ویژگی‌های عمومی منطقه‌های تعیین‌شده برای نمونه‌برداری (مجموعه داده‌های واقعی اول در استان مازندران)

سری	شماره پارسل	مساحت (هکتار)	تعداد قطعه‌نمونه	جهت عمومی	محدوده ارتفاعی (متر از سطح دریا)
سری چهار دهگا	۴۱۴ (مدیریت‌شده)	۷۱	۲۰	شرقی	۳۰۰-۴۰۰
	۴۱۲ (شاهد)	۵۹		شمالی	
سری ۱۰ لالیس	۲۹ (مدیریت‌شده)	۶۷	۲۰	شمال شرقی	۱۲۵۰-۱۴۰۰
	۲۶ (شاهد)	۵۹		شمالی	
سری یک شیراکنس	۱۳۴ (مدیریت‌شده)	۱۱۳	۲۰	جنوب شرقی	۱۶۰۰-۱۸۵۰
	۱۳۵ (شاهد)	۱۲۶		جنوب شرقی	
سری پنج لاکوبن	۵۳۰ (مدیریت‌شده)	۲۴/۷	۲۰	شمال غربی	۸۰۰-۹۰۰
	۵۴۱ (شاهد)	۹۲		شمال غربی	
سری هفت واشمرد	۷۲۵ (مدیریت‌شده)	۳۴	۲۰	جنوب شرقی	۷۰۰-۸۵۰
	۷۲۴ (شاهد)	۵۸		شرقی	
سری ده جمند	۳۱۷ (مدیریت‌شده)	۶۴	۲۰	شمال شرقی	۱۱۳۰-۱۳۰۰
	۳۱۸ (شاهد)	۴۷		شمال شرقی	

جدول ۲- ویژگی‌های عمومی منطقه‌های تعیین‌شده برای نمونه‌برداری (مجموعه داده‌های واقعی دوم در استان کرمانشاه)

جنگل چهارزبر	مساحت (هکتار)	تعداد قطعه‌نمونه	جهت عمومی	ارتفاع (متر از سطح دریا)
قطعه‌نمونه یک	۴/۵	۱۵	شمال شرقی	۱۷۰۰
قطعه‌نمونه دو	۵	۱۵	شمال شرقی	۱۶۸۰
قطعه‌نمونه سه	۵/۵	۱۵	شمال شرقی	۱۶۵۰

داده‌های شبیه‌سازی‌شده

به‌منظور ارزیابی روش‌های خوشه‌بندی، شش گروه‌داده شبیه‌سازی‌شده با استفاده از بسته آماری Cluster Generation (Qiu & Joe, 2015) و Coenoflex (Roberts, 2016) به‌کار برده شد. برای شبیه‌سازی، دو گرادیان محیطی با ۲۵۰ گونه در نظر گرفته شد. فراوانی گونه‌ها برای قطعه‌نمونه‌ها نیز صددرصد انتخاب شد. هر گروه‌داده که در زیر به آن‌ها اشاره شده است، پنج خوشه ازپیش‌تعیین‌شده داشت و برای تعیین خوشه‌ها از معیار جدایش استفاده شد.

۱- گروه‌داده‌های شبیه‌سازی‌شده با فاصله (درجه جدایش صفر) شامل ۱۴۰ گونه و ۲۵۰ قطعه‌نمونه (Sep 0)

۲- گروه‌داده‌های شبیه‌سازی‌شده با فاصله (درجه جدایش ۰/۱-) شامل ۱۰۵ گونه و ۲۵۰ قطعه‌نمونه (Sep 1)

۳- گروه‌داده‌های شبیه‌سازی‌شده با فاصله (درجه جدایش ۰/۲-) شامل ۱۷۸ گونه و ۲۵۰ قطعه‌نمونه (Sep 2)

۴- گروه‌داده‌های شبیه‌سازی‌شده با فاصله (درجه جدایش ۰/۳-) شامل ۱۶۷ گونه و ۲۵۰ قطعه‌نمونه (Sep 3)

۵- گروه‌داده‌های شبیه‌سازی‌شده با فاصله (درجه جدایش ۰/۴-) شامل ۹۸ گونه و ۲۵۰ قطعه‌نمونه (Sep 4)

۶- گروه‌داده‌های شبیه‌سازی‌شده با فاصله (درجه جدایش ۰/۵-) شامل ۱۱۸ گونه و ۲۵۰ قطعه‌نمونه (Sep 5)

تجزیه‌وتحلیل داده‌ها

ازآنجایی‌که تبدیل داده بر نتایج خوشه‌بندی تأثیر می‌گذارد، تبدیل داده هلینگر (Hellinger) به‌منظور آماده‌سازی داده‌ها به‌کار برده شد (Legendre & Gallagher, 2001). برای بررسی عملکرد دو روش K-means و K-medoids از سه روش اندازه‌گیری تشابه شامل فاصله اقلیدسی (Euclidean distance)، فاصله بری‌کورتیس (Bray-Curtis distance) و فاصله منهتن (Manhattan distance) استفاده شد که به‌دلیل محبوبیت و پرکاربرد بودن در علوم بوم‌شناسی انتخاب شدند. برحسب همگنی داده‌ها، دو تا پنج خوشه برای داده‌های هیرکانی و دو تا شش خوشه برای داده‌های زاگرس انتخاب شدند. سپس، خوشه‌بندی براساس تعداد خوشه‌های انتخاب‌شده انجام شد و کیفیت خوشه‌ها با استفاده از معیارهای ارزیابی بررسی شد. نتایج به‌دست‌آمده برای هر معیار از بهترین به بدترین رتبه‌بندی شدند. همچنین، میانگین معیارهای رتبه‌بندی‌شده محاسبه و به‌این‌ترتیب، خوشه‌بندی مناسب معرفی شد. برای بررسی بهتر نتایج، یافته‌های هر ارزیابی‌کننده به‌شکل نمودار جعبه‌ای رسم شد (Schmidtlein *et al*., 2010; Roberts, 2015). همه محاسبات در نرم‌افزار R ver. 3.6.1 انجام گرفت. روش‌های خوشه‌بندی و نیز روش‌های ارزیابی خوشه‌ها در ادامه تشریح شده است.

تبدیل داده هلینگر

تبدیل داده هلینگر برای داده‌های فراوانی، بسیار مناسب است. در این روش، وزن کمی به گونه‌های نادر داده می‌شود. تبدیل داده شامل تقسیم هر مقدار به مجموع هر ردیف در ماتریس است. سپس، مجذور هر مقدار محاسبه می‌شود (Legendre & Gallagher, 2001).

روش‌های خوشه‌بندی غیرسلسله‌مراتبی

خوشه‌بندی K-means یک روش تکرارشونده است (MacQueen, 1967) که در آن k خوشه به‌عنوان ورودی در نظر گرفته می‌شود و مجموعه n نمونه در k خوشه قرار می‌گیرد. اعضای خوشه با محاسبه مرکز هر گروه و قرارگیری نمونه‌ها به نزدیک‌ترین مرکز تعیین می‌شوند. به‌طوری‌که شباهت داخلی خوشه‌ها، زیاد و شباهت بین خوشه‌ها کم باشد. شباهت هر خوشه نسبت به متوسط نمونه‌های آن خوشه سنجیده می‌شود که این متوسط، مرکز خوشه نیز نامیده می‌شود. قرارگیری نمونه‌ها و تغییرات مرکز آن‌ها به‌طور مرتب تغییر می‌کنند تازمانی‌که همگنی خوشه‌ها بیشینه شود (Liu & Graham, 2019).

روش خوشه‌بندی K-medoids یا PAM (Kaufman & Rousseeuw, 1990) مانند روش K-means است، بااین‌تفاوت که به‌جای استفاده از میانگین، خود نمونه‌ها برای مرکز ثقل و نمایندگی خوشه‌ها به‌کار برده می‌شوند. این روش، K نمونه معرف را برای کمینه‌سازی عدم تشابه بین نمونه‌ها استفاده می‌کند و می‌تواند از هر ماتریس تشابه/عدم تشابه استفاده کند. PAM به‌طور پیش‌فرض، الگوریتمی برای انتخاب اولیه مرکز خوشه‌ها از داده‌ها دارد که نیاز به شروع تصادفی را از بین می‌برد.

معیارهای ارزیابی کیفیت خوشه‌بندی

شاخص میانگین سیلوئت (Silhouette) (Rousseeuw, 1987): این معیار برای یافتن تعداد بهینه خوشه‌ها و ارزیابی کیفیت آن‌ها به‌کار برده می‌شود. شاخص مذکور از فشردگی (میانگین فاصله داخل خوشه‌ها) و جدایش (میانگین فاصله بین خوشه‌ها) استفاده می‌کند. محدوده این شاخص بین ۱- تا یک تغییر می‌کند. مقدار نزدیک به یک نشان‌دهنده طبقه‌بندی خوب است، درحالی‌که اگر مقدار این شاخص صفر باشد، طبقه‌بندی منطقی به‌نظر نمی‌رسد و اگر برابر ۱- باشد، نشان‌دهنده طبقه‌بندی اشتباه است.

ضریب همبستگی فی (Phi) (Tichý & Chytrý, 2006): این ضریب، مقدار وفاداری گونه‌ها را براساس داده‌های حضوروغیاب و با تعداد رویشگاه‌های مختلف محاسبه می‌کند. مقدار آن بین ۱- تا یک متغیر است. مقدار مثبت به‌منظور شناسایی گونه‌های شاخص اهمیت دارد، درحالی‌که مقدار منفی برای یافتن اختلاف بین جوامع مهم است.

آنالیز گونه‌های معرف خوشه‌ها برای کمینه کردن ثبات میانی (ISAMIC): تجزیه‌وتحلیلی است که ثبات گونه‌ها (گونه‌هایی که همیشه در خوشه‌ها حضور دارند یا همیشه بین گونه‌ها غایب هستند) را بین خوشه‌ها محاسبه می‌کند (Roberts, 2015). مقدار این شاخص بین ۱- تا یک است که مقدار مثبت، خوشه‌بندی بهتر را نشان می‌دهد.

نتایج

در پژوهش پیش‌رو، شش روش ترکیبی به‌منظور تعیین مناسب‌ترین آن‌ها برای خوشه‌بندی داده‌های پوشش گیاهی استفاده شد. به‌این‌منظور، دو روش خوشه‌بندی غیرسلسله‌مراتبی با استفاده از سه روش اندازه‌گیری تشابه در هشت مجموعه داده ارزیابی شدند. جدول ۳، نتایج رتبه‌بندی ارزیابی‌کننده‌های مختلف برای داده‌های مربوط به جنگل‌های هیرکانی را نشان می‌دهد. در این مجموعه داده، ترکیب ماتریس تشابه بری‌کورتیس و خوشه‌بندی K-mediods رتبه اول را در بین روش‌های دیگر کسب کرد. براساس ارزیابی‌کننده‌های سیلوئت و ISAMIC، ترکیب مذکور نیز بهترین عملکرد را در بین خوشه‌های مختلف مجموعه داده‌های هیرکانی به‌خود اختصاص داد، درحالی‌که بهترین روش خوشه‌بندی در همبستگی فی به ترکیب ماتریس تشابه منهتن و K-mediods تعلق گرفت.

جدول ۳- نتایج رتبه‌بندی کلی الگوریتم‌های مختلف خوشه‌بندی برای داده‌های جنگل‌های هیرکانی

ماتریس تشابه	خوشه‌بندی	Silhouette	Phi	ISMAIC	میانگین	رتبه
اقلیدسی	k-means	۴/۲۵	۴/۸۸	۲/۷۵	۳/۹۶	۵
	k-mediods	۴/۲۵	۱/۷۵	۳/۵	۳/۱۷	۲/۵
بری‌کورتیس	k-means	۱/۶۲	۴/۸۸	۳	۳/۱۷	۲/۵
	k-mediods	۱/۳۷	۳/۱۳	۲	۲/۱۷	۱
منهتن	k-means	۴/۷۵	۴/۸۸	۵	۴/۸۹	۶
	k-mediods	۴/۷۵	۱/۵	۴/۷۵	۳/۶۷	۴

جدول ۴، نتایج رتبه‌بندی ارزیابی‌کننده‌های مختلف برای خوشه‌بندی‌های آزمون‌شده در داده‌های جنگل‌های زاگرس را نشان می‌دهد. در این مجموعه داده، ترکیب بهتر به ماتریس تشابه بری‌کورتیس و روش خوشه‌بندی K-means تعلق گرفت. ارزیابی‌کننده‌های سیلوئت نیز همین ترکیب را بهتر از روش‌های دیگر معرفی می‌کند، درحالی‌که عملکرد خوشه‌بندی ماتریس تشابه منهتن و خوشه‌بندی K-means توسط ارزیابی‌کننده ISAMIC و ترکیب ماتریس تشابه منهتن و خوشه‌بندی K-mediods توسط ارزیابی‌کننده فی به‌عنوان روش‌های برتر تشخیص داده شده‌اند.

جدول ۴- نتایج رتبه‌بندی کلی الگوریتم‌های مختلف خوشه‌بندی برای داده‌های جنگل‌های زاگرس

ماتریس تشابه	خوشه‌بندی	Silhouette	Phi	ISMAIC	میانگین	رتبه
اقلیدسی	k-means	۲/۲	۴/۸	۴/۴	۳/۸	۵
	k-medoids	۲/۸	۳	۵/۲	۳/۶۷	۴
بری‌کورتیس	k-means	۱	۳/۹	۳/۸	۲/۹	۱
	k-medoids	۴/۲	۲/۸	۲	۳	۲
منهتن	k-means	۴/۸	۳/۹	۱/۸	۳/۵	۳
	k-medoids	۶	۲/۶	۳/۸	۴/۱۳	۶

نتایج رتبه‌بندی شش مجموعه داده شبیه‌سازی‌شده در جدول ۵ ارائه شده است. براین‌اساس، ترکیب روش اقلیدسی و خوشه‌بندی k-means نسبت به روش‌های دیگر، برتری دارد. ارزیابی‌کننده همبستگی فی و ISAMIC نیز بر ارجحیت عملکرد ترکیب مذکور تأکید دارد، اما براساس ارزیابی‌کننده سیلوئت، خوشه‌بندی ماتریس تشابه منهتن و k-means برتر از روش‌های دیگر بودند.

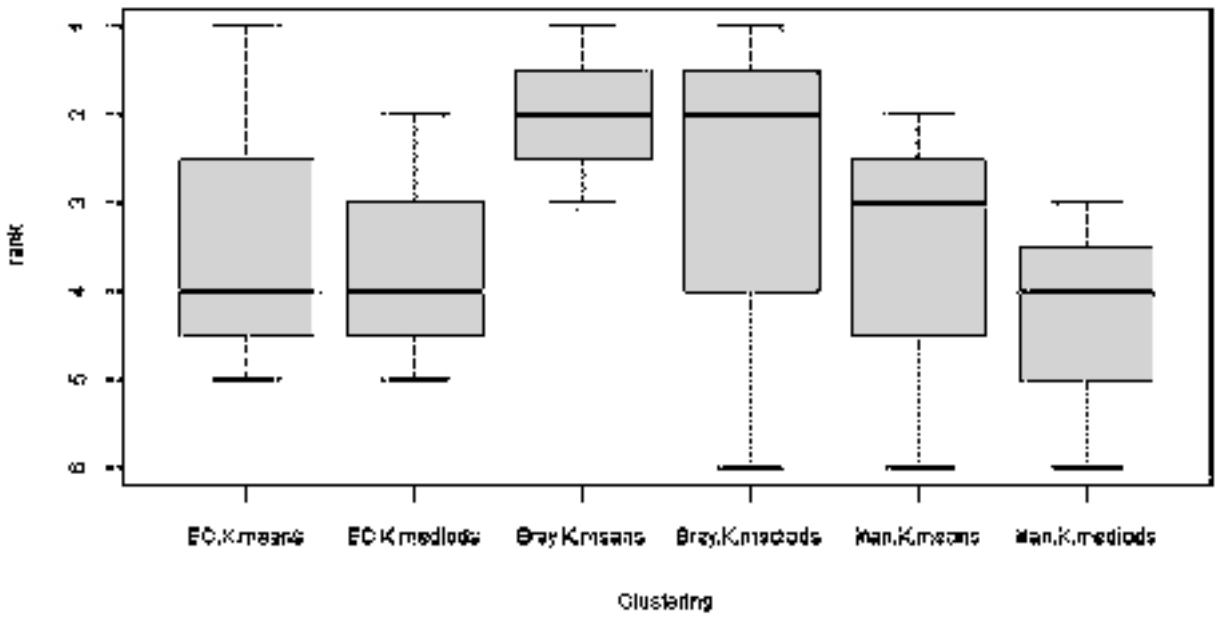
نتایج جدول ۶ و شکل ۱ نشان می‌دهند که ترکیب ماتریس تشابه بری‌کورتیس و روش‌های خوشه‌بندی K-means و K-medoids به‌ترتیب رتبه‌های اول و دوم را در میان خوشه‌بندی‌های مختلف دارند. ضعیف‌ترین خوشه‌بندی نیز در ترکیب ماتریس تشابه منهتن و روش K-medoids مشاهده شد.

جدول ۵- نتایج رتبه‌بندی کلی الگوریتم‌های مختلف خوشه‌بندی برای داده‌های شبیه‌سازی‌شده

ماتریس تشابه	خوشه‌بندی	Silhouette	Phi	ISMAIC	میانگین	رتبه
اقلیدسی	k-means	۲/۸۳	۲/۶۷	۱/۱۷	۲/۲۲	۱
	k-medoids	۴/۵	۴/۵	۴/۱۷	۴/۳۹	۵
بری‌کورتیس	k-means	۲/۸۳	۳/۳۳	۳/۵	۳/۲۲	۳
	k-medoids	۵/۳۳	۳/۸۳	۵	۴/۷۲	۶
منهتن	k-means	۲/۱۷	۳/۵	۲/۳۳	۲/۶۷	۲
	k-medoids	۳/۳۳	۳/۱۷	۴/۸۳	۳/۷۸	۴

جدول ۶- نتایج رتبه‌بندی کلی الگوریتم‌های مختلف خوشه‌بندی برای داده‌های مختلف

ماتریس تشابه	خوشه‌بندی	داده‌های جنگل‌های هیرکانی	داده‌های جنگل‌های زاگرس	داده‌های شبیه‌سازی‌شده	میانگین	رتبه
اقلیدسی	K-means	۳/۹۶	۳/۸	۲/۲۲	۳/۳۳	۳
	K-mediods	۳/۱۷	۳/۶۷	۴/۳۹	۳/۷۴	۵
بری‌کورتیس	K-means	۳/۱۷	۲/۹	۳/۲۲	۳/۱	۱
	K-mediods	۲/۱۷	۳	۴/۷۲	۳/۳	۲
منهتن	K-means	۴/۸۹	۳/۵	۲/۶۷	۳/۶۸	۴
	K-mediods	۳/۶۷	۴/۱۳	۳/۷۸	۳/۸۶	۶



شکل ۱- نمودار جعبه‌ای رتبه‌بندی‌شده برای خوشه‌بندی‌های مختلف مجموعه داده‌ها

در محور عمودی، رتبه یک بیانگر بهترین و رتبه شش نشان‌دهنده نامناسب‌ترین خوشه‌بندی است.

بحث

خوشه‌بندی، روشی جدید در علوم پوشش گیاهی نیست، اما یکی از پرکاربردترین روش‌ها برای طبقه‌بندی جوامع گیاهی به‌شمار می‌آید (Lengyel *et al*., 2018). نتایج مقایسه روش‌های خوشه‌بندی غیرسلسله‌مراتبی در پژوهش پیش‌رو نشان داد که استفاده از چند ارزیابی‌کننده برای مقایسه روش‌های مذکور به‌منظور انتخاب روش برتر، مؤثرتر است. همچنین، به‌دلیل تأثیر پیچیدگی داده‌ها بر نتایج خوشه‌بندی پیشنهاد می‌شود که از چند مجموعه داده برای انتخاب روش خوشه‌بندی بهینه استفاده شود (Lengyel *et al*., 2018). براساس نتایج پژوهش پیش‌رو، ماتریس تشابه بری‌کورتیس بهتر از روش‌های منهتن و اقلیدسی عمل می‌کند. Lötter و همکاران (۲۰۱۳) نیز ماتریس تشابه بری‌کورتیس را برتر از روش‌های دیگر محاسبه ماتریس تشابه گزارش کردند. مطابق با نتایج دیگر پژوهش پیش‌رو، ماتریس فاصله اقلیدسی در داده شبیه‌سازی‌شده، روش کارآمدتری بود. همچنین، با استفاده از تبدیل داده هلینگر پیش از انجام ماتریس فاصله اقلیدسی، این ماتریس به‌روش مؤثری تبدیل می‌شود (Legendre & Gallagher, 2001). برخلاف روش فاصله‌های اقلیدسی، استفاده از ماتریس فاصله هلینگر به دو دلیل در جوامع گیاهی توصیه می‌شود. اول اینکه روش مذکور، محدودیت‌های روش فاصله‌های اقلیدسی را ندارد. زیرا تعداد صفر در داده‌های جوامع گیاهی زیاد است و اغلب یکی از متغیرها در مجموعه داده، چولگی دارد. به همین دلیل، روش اقلیدسی برای داده‌های جوامع گیاهی پیشنهاد نمی‌شود. دلیل دوم اینکه ماتریس فاصله هلینگر برای آنالیزهایی که به فضای متریک نیاز دارند، مناسب‌تر است (Legendre & De Cáceres, 2013). ماتریس تشابه منهتن در پژوهش پیش‌رو، ضعیف‌ترین عملکرد را در بین روش‌های ماتریس فاصله مورد استفاده داشت. Legendre و De Cáceres (۲۰۱۳) نیز این روش را برای داده‌های جوامع گیاهی، نامناسب گزارش کردند. ویژگی‌های مانند غیرمنفی بودن، تقارن و نابرابری مثلثی که برای فاصله‌های هندسی مانند روش ماتریس اقلیدسی و منهتن صدق می‌کند، استفاده از این روش‌ها را برای جوامع گیاهی، نامناسب و پیچیده کرده است (Roberts, 2017). در پژوهش پیش‌رو با استفاده از تبدیل داده هلینگر، اثر ویژگی‌های مذکور کمتر شد که بهبود عملکرد ماتریس فاصله اقلیدوسی را در پی داشت.

دومین عامل تصمیم‌گیری در خوشه‌بندی جوامع گیاهی، انتخاب روش خوشه‌بندی است. نتایج این پژوهش نشان داد که در داده‌های همگن‌تر (داده‌های جنگل‌های هیرکانی)، روش خوشه‌بندی K-medoids عملکرد بهتری دارد، اما K-means در داده‌های ناهمگن‌تر (داده‌های جنگل‌های زاگرس) و شبیه‌سازی‌شده، روش بهتری بود. همچنین، با تغییر ماتریس فاصله اقلیدسی به ماتریس فاصله هلینگر و بری‌کورتیس، ثبات روش خوشه‌بندی K-means افزایش یافت (Peterson *et al*., 2010). Pakgohar و همکاران (۲۰۲۱) در مقایسه روش‌های مختلف برای داده‌های پوشش گیاهی، روش خوشه‌بندی K-meansرا برتر از روش K-mediods معرفی کردند. روش K-means نسبت به نویز در داده‌ها حساس است (Hämäläinen *et al*., 2017). تبدیل داده هلینگر سبب بهبود چشم‌گیر عملکرد این روش شد، درحالی‌که روش K-medoidsبیشتر تحت تأثیر ابعاد داده‌ها قرار دارد (Hämäläinen *et al*., 2017) که این موضوع با افزایش تعداد داده در مجموعه داده‌های زاگرس و شبیه‌سازی‌شده مشهود بود. Roberts (۲۰۱۵) نیز به‌منظور محاسبه روش‌های خوشه‌بندی با ارزیابی‌کننده‌های هندسی، عملکرد بهتری را برای روش K-means گزارش کرد.

خوشه‌بندی باثبات به عوامل مختلفی بستگی دارد. سازگاری بین روش خوشه‌بندی و ماتریس اندازه‌گیری فاصله از عوامل مؤثر بر نتیجه خوشه‌بندی است. ترکیب مناسب روش خوشه‌بندی با ماتریس اندازه‌گیری فاصله سبب ایجاد نتایج خوشه‌بندی منطقی و قابل‌تفسیر می‌شود. براساس نتایج پژوهش پیش‌رو، ترکیب روش خوشه‌بندی K-means و ماتریس تشابه بری‌کورتیس برای داده‌های جوامع گیاهی پیشنهاد می‌شوند.

منابع مورد استفاده

- Aho, K., Roberts, D.W. and Weaver, T., 2008. Using geometric and non-geometric internal evaluators to compare eight vegetation classification methods. Journal of Vegetation Science, 19(4): 549-562.
- Eshaghi Rad, J., Soleimani, F. and Khodakarami, Y., 2014. Influence of edge effect on plant composition and distribution in oak forests (case study: Cheharzebar forests-Kermanshah). Iranian Journal of Forest and Poplar Research, 22(3): 527-539 (In Persian).
- Eshaghi Rad, J., Zahedi Amiri, Gh., Marvi Mohajer, M.R. and Mataji, A., 2009. Relationship between vegetation and physical and chemical properties of soil in *Fagetum* communities (Case study: Kheiroudkenar forest). Iranian Journal of Forest and Poplar Research, 17(2): 174-187 (In Persian).
- Hämäläinen, J., Jauhiainen, S. and Kärkkäinen, T., 2017. Comparison of internal clustering validation indices for prototype-based clustering. Algorithms, 10(3): 105.
- Janatbabaei, M., Moradi, Gh. and Feghhi, J., 2020. Effect of soil and topography characteristics on distribution of plant types in the Arasbaran forests, Iran. Journal of Forest Research and Development, 5(4): 583-597 (In Persian).
- Kaufman, L. and Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc., Hoboken, New Jersey, 342p.
- Khanalizadeh, A., Eshaghi Rad, J., Zahedi Amiri, Gh., Zare, H., Rammer, W. and Lexer, M.J., 2020. Assessing selected microhabitat types on living trees in Oriental beech (*Fagus orientalis* L.) dominated forests in Iran. Annals of Forest Science, 77(3): 91.
- Legendre, P. and De Cáceres, M., 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. Ecology Letters, 16(8): 951-963.
- Legendre, P. and Gallagher, E.D., 2001. Ecologically meaningful transformations for ordination of species data. Oecologia, 129(2): 271-80.
- Legendre, P. and Legendre, L., 2012. Numerical Ecology, 3rd Edition. Elsevier, Amsterdam, 1006p.
- Lengyel, A. and Botta-Dukát, Z., 2019. Silhouette width using generalized mean—A flexible method for assessing clustering efficiency. Ecology and Evolution, 9(23): 13231-13243.
- Lengyel, A., Landucci, F., Mucina, L., Tsakalos, J.L. and Botta-Dukát, Z., 2018. Joint optimization of cluster number and abundance transformation for obtaining effective vegetation classifications. Journal of Vegetation Science, 29(2): 336-347.
- Liu, D. and Graham, J., 2019. Simple measures of individual cluster-membership certainty for hard partitional clustering. The American Statistician, 73(1): 70-79.
- Lötter, M.C., Mucina, L. and Witkowski, E.T.F., 2013. The classification conundrum: species fidelity as leading criterion in search of a rigorous method to classify a complex forest data set. Community Ecology, 14(1): 121-132
- MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Volume 1: Statistics. Berkeley, California, 21 June-18 July 1965, 27 Dec. 1965 and 7 Jan. 1966: 281-297.
- Morris, T.P., White, I.R. and Crowther, M.J., 2019. Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38(11): 2074-2102.
- Pakgohar, N., Eshaghi Rad, J., Gholami, G.H., Alijanpour, A. and Roberts, D.W., 2021. A comparative study of hard clustering algorithms for vegetation data. Journal of Vegetation Science, 32(3): e13042.
- Peet, R.K. and Roberts, D.W., 2013. Classification of natural and semi-natural vegetation: 28-70. In: van der Maarel, E. and Franklin, J. (Eds.). Vegetation Ecology, Second Edition. Wiley-Blackwell, Oxford, 584p.
- Peterson, A.D., Ghosh, A.P. and Maitra, R., 2010. A systematic evaluation of different methods for initializing the k-means clustering algorithm. Technical Report 07, Department of Statistics, Iowa State University, Ames, Iowa, 105p.
- Qiu, W. and Joe, H., 2015. The clusterGeneration package. Available at: https://cran.r-project.org/web/packages/clusterGeneration/index.html
- Roberts, D.W., 2015. Vegetation classification by two new iterative reallocation optimization algorithms. Plant Ecology, 216(5): 741-758.
- Roberts, D.W., 2016. Package 'coenoflex: Gradient-Based Coenospace Vegetation Simulator, Version 2.2-0. Available at: https://cran.r-project.org/web/packages/coenoflex/index.html
- Roberts, D.W., 2017. Distance, dissimilarity, and mean–variance ratios in ordination. Methods in Ecology and Evolution, 8(11): 1398-1407.
- Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L.D.F. and Rodrigues, F.A., 2019. Clustering algorithms: A comparative approach. PLoS One, 14(1): e0210236.
- 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.
- Schmidtlein, S., Tichý, L., Feilhauer, H. and Faude, U., 2010. A brute-force approach to vegetation classification. Journal of Vegetation Science, 21(6): 1162-1171.
- Tichý, L. and Chytrý, M., 2006. Statistical determination of diagnostic species for site groups of unequal size. Journal of Vegetation Science, 17(6): 809-818.
- Tichý, L., Chytrý, M. and Botta-Dukát, Z., 2014. Semi-supervised classification of vegetation: preserving the good old units and searching for new ones. Journal of Vegetation Science, 25(6): 1504-1512.

Comparison of two non-hierarchal clustering performance in vegetation community datasets

N. Pakgohar [1], J. Eshaghi Rad [2*], Gh. Gholami [3], A. Alijanpour [4] and D.W. Roberts [5]

1- Ph.D. of Forestry, Department of Forestry, Faculty of Natural Resources, Urmia University, Urmia, Iran
2*- Corresponding author, Prof., Department of Forestry, Faculty of Natural Resources, Urmia University, Urmia, Iran
E-mail: j.eshaghi@urmia.ac.ir
3- Assistant Prof., Department of Mathematics, Faculty of Science, Urmia University, Urmia, Iran
4- Associate Prof., Department of Forestry, Faculty of Natural Resources, Urmia University, Urmia, Iran
5- Prof., Department of Ecology, Montana State University, Bozeman, USA

Received: 09.10.2021 Accepted: 17.12.2021

Abstract

Clustering task is optimized and summarized high dimensional vegetation datasets that indicator of environmental change and gathering to interpreting pattern form ecosystem. Variety clustering methods is available and the issue is chosen proper methods. The aim of the research was compared two non-hierarchical clustering as K-means and K-medoids in forest ecosystems. For this purpose, two real datasets from Hyrcanian and Zagros forests of Iran and six simulated datasets were applied. The Hellinger transformation was employed before calculating dissimilarity matrices. Euclidean distance, Manhattan distance and Bray-Curtis dissimilarity indices were then calculated on the transformed data sets. And three evaluators including silhouette width, phi coefficient and ISAMIC were chosen. The results show that combination of Bray-Curtis dissimilarity matrices and K-means and K-medoids have first and second ranks among other clustering methods. K-means clustering is more effective in heterogenous dataset as Zagros and simulated datasets. The weakest clustering algorithm was combination between Manhattan distance and K-medoids. Also results show that Hellinger data transformation cause to improve Euclidean distance matrix. Our results indicated that combination of Bray-Curtis dissimilarity with K-means is more significant and recommended.

Keywords: Clustering accuracy, data transformation, distance measure, simulated dataset.