نشریه علمی تحقیقات جنگل و صنوبر ایران
جلد ۲۷ شماره ۲، صفحه ۲۱۸-۲۰۷، (۱۳۹۸)

شناسه دیجیتال (DOI): 10.22092/ijfpr.2019.120127
شناسه دیجیتال (DOR): 98.1000/1735-0883.1398.27.207.76.2.1605.41

مقایسه روش‌های مختلف تجزیه‌وتحلیل خوشه‌ای (مطالعه موردی: جنگل‌های بلوط کرمانشاه)

نغمه پاک‌گهر [۱]، جواد اسحاقی راد [۲*]، غلامحسین غلامی [۳]، احمد علیجانپور [۴] و دیوید رابرتز [۵]

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

۲*- نویسنده مسئول، دانشیار، گروه جنگل‌داری، دانشکده منابع طبیعی، دانشگاه ارومیه، ارومیه، ایران. پست الکترونیک: javad.eshaghi@yahoo.com

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، بوزمن، آمریکا

تاریخ دریافت: ۱۳۹۷/۱۲/۲۱ تاریخ پذیرش: ۱۳۹۸/۰۳/۲۰

چکیده

طبقه‌بندی، ابزاری کارآمد برای پژوهش‌های جوامع گیاهی و بررسی پدیده‌های بوم‌شناختی است. هدف از پژوهش پیش‌رو، مقایسه روش‌های مختلف خوشه‌بندی در تجزیه‌وتحلیل خوشه‌ای بود. سه قطعه جنگلی با جهت جنوبی از توده‌های بلوط در جنگل‌های چهارزبر استان کرمانشاه با شرایط مشابه ازنظر شیب و ارتفاع از سطح دریا انتخاب شدند. در هر قطعه در فاصله‌های صفر، ۲۵، ۵۰، ۱۰۰ و ۱۵۰ متری با استفاده از سه خط‌نمونه که در فاصله‌های ۲۰۰ متری از هم قرار گرفتند، نمونه‌برداری انجام شد. در این بررسی از روش تجزیه‌وتحلیل خوشه‌ای برای طبقه‌بندی پوشش گیاهی استفاده شد. برای محاسبه ماتریس فاصله‌ها از روش Gower و برای اتصال خوشه‌ها از چهار روش نزدیک‌ترین همسایه، دورترین همسایه، اتصال میانگین و اتصال وارد استفاده شد. برای یافتن تعداد بهینه خوشه‌ها و بررسی کیفیت خوشه‌بندی در روش‌های مختلف از معیار سیلوئت استفاده شد. همچنین، انطباق بین ماتریس فاصله محاسبه‌شده و دندروگرام حاصل از روش‌های مختلف با ضریب همبستگی کوفنتیک ارزیابی شد. نتایج نشان داد که تعداد بهینه خوشه‌ها در جوامع بلوط منطقه مورد مطالعه، دو خوشه بود. مقدار همبستگی کوفنتیک بین ماتریس فاصله و دندروگرام به‌دست‌آمده از روش‌های میانگین و نزدیک‌ترین همسایه بیشتر از دو روش وارد و دورترین همسایه به‌دست آمد. همچنین، کیفیت خوشه‌بندی روش‌های نزدیک‌ترین همسایه و میانگین بهتر از دو روش دیگر بود، اما میانگین شاخص سیلوئت در خوشه دوم روش نزدیک‌ترین همسایه، بسیار کم بود، بنابراین روش اتصال خوشه میانگین همراه با ضریب فاصله Gower برای داده‌های ترتیبی مطلوب‌تر است و تغییری در داده‌ها ایجاد نمی‌کند.

واژه‌های کلیدی: خوشه‌بندی، داده‌های ترتیبی، طبقه‌بندی، ضریب فاصله Gower.

مقدمه

مسیری که یک بوم‌شناس برای تجزیه‌وتحلیل داده‌های پوشش گیاهی طی می‌کند، شامل نمونه‌برداری، جمع‌آوری داده‌ها، محاسبه ماتریس تشابه یا عدم تشابه و طبقه‌بندی یا رسته‌بندی داده‌ها است (Podani, 2005). نمونه‌برداری از پوشش گیاهی نخستین مرحله از پژوهش‌های بوم‌شناسی است که داده‌های آن به دو صورت کمی و کیفی برداشت می‌شوند (Damgaard, 2014).

به‌طور کلی، برای تجزیه‌وتحلیل پوشش گیاهی از روش‌های کمی رسته‌بندی و طبقه‌بندی استفاده می‌شود که

به‌عنوان ابزاری بنیادی برای شناسایی الگوهای بوم‌شناسی به‌کار گرفته می‌شوند (Suh *et al*., 2009). رسته‌بندی، یک روش قوی برای جوامع زیستی است که می‌تواند شیب تغییرات محیطی را شناسایی کند. با این وجود، وارد کردن داده‌های پرت، اثرات چشمگیری در تمام فنون رسته‌بندی دارد (Belbin & McDonald, 1993)، بنابراین استفاده از روش‌های مؤثر طبقه‌بندی به‌منظور یافتن الگوهای بوم‌شناسی برای حفاظت از طبیعت، نقشه‌برداری از منابع طبیعی و آمایش سرزمین غیرقابل انکار است (De Cáceres *et al*., 2010; McGranahan *et al*., 2013). از معمول‌ترین روش‌ها برای تحلیل پوشش گیاهی، روش‌های مختلف طبقه‌بندی هستند (Lengyel & Podani, 2015). ایده اصلی طبقه‌بندی اطلاعات، جدا کردن نمونه‌ها از هم و قرار دادن آن‌ها در گروه‌های شبیه به‌هم است (Ken *et al*., 2008)، بنابراین طبقه‌بندی پوشش گیاهی برای تفکیک داده‌های ناهمگن به گروه‌های همگن‌تر انجام می‌شود تا بررسی تغییرات پوشش گیاهی آسان‌تر شود (lengyel & Podani, 2015)، اما با وجود مزیت‌های فراوان، این روش‌ها نیز محدودیت‌هایی دارند.

با توسعه علوم رایانه‌ای و به‌منظور بهره‌گیری از روش‌های عددی چندمتغیره در فرآیندهای طبقه‌بندی پوشش گیاهی، تلاش برای کاهش عامل ذهنیت در توصیف پوشش گیاهی و عوامل محیطی مطرح شده است (Grabherr *et al*., 2003). برای گروه‌بندی نمونه‌ها، روش‌های مختلف طبقه‌بندی وجود دارد که بیشتر آن‌ها نتایج متفاوتی ارائه می‌دهند. از این‌رو، تاکنون هیچ روش طبقه‌بندی ازنظر تئوری یا عملی پذیرفته نشده است (Vavrek, 2016). روش تجزیه‌وتحلیل خوشه‌ای در پژوهش‌های متعدد داخلی و خارجی استفاده شده است (Mahmoodi *et al*., 2015; Alamgir *et al*., 2016; Lechner *et al*., 2016). تجزیه‌وتحلیل خوشه‌ای یا طبقه‌بندی خوشه‌ای، یک روش آماری برای گروه‌بندی داده‌ها یا مشاهدات باتوجه‌به شباهت یا درجه نزدیکی آن‌ها است. تجزیه خوشه‌ای، ابزار میان‌بر تحلیل داده‌ها محسوب می‌شود. هدف آن نظم دادن به اشیاء مختلف به گروه‌هایی است که درجه ارتباط بین دو شیء، اگر آن‌ها به یک گروه تعلق داشته باشند، بیشترین و در غیر این صورت، کمترین است (Shakeri *et al*., 2011). در این روش ابتدا ماتریس داده‌های خام تنظیم می‌شود. سپس درجه جور کردن هر جفت از قطعه‌نمونه‌ها براساس ضریب تشابه یا عدم تشابه محاسبه می‌شو. هنگامی‌که تشابه یا عدم تشابه بین کلیه جفت‌های قطعه‌نمونه‌ها محاسبه شد، ماتریس تشابه یا عدم تشابه به‌وجود می‌آید. درنهایت، جفت‌هایی با بیشترین مقدار تشابه باهم در یک گروه ترکیب می‌شوند. سپس دو جفت بعدی با بیشترین تشابه پیدا می‌شود و به‌همین ترتیب جفت‌های دیگر با تشابه بیشتر باهم ترکیب خواهند شد (Everitt *et al*., 2011). در پژوهش‌های بوم‌شناختی، اغلب با هدف کاهش هزینه و زمان برداشت، داده‌های ترتیبی جمع‌آوری می‌شوند، اما در داده‌های ترتیبی از یک‌سو از کیفیت داده‌ها کاسته می‌شود و ازسوی دیگر، امکان تبدیل این داده‌ها وجود ندارد (Zuur *et al*., 2010). از آنجایی‌که تمام تجزیه‌وتحلیل‌ها به داده‌های ورودی بستگی دارند (Gill & Tipper, 1978)، استفاده از روش‌های متریک برای داده‌های ترتیبی سبب انحراف تجزیه‌وتحلیل می‌شود، بنابراین باید درنظر داشت که برای طبقه‌بندی نیز باید از روش‌های آماری متناسب با سری داده‌ها استفاده شود (Hall & Richardson, 2016). باتوجه‌به روش‌هایی که برای تجزیه‌وتحلیل داده‌ها انتخاب می‌شود، مقیاس داده‌ها ممکن است ثابت بماند یا تغییر کند. به‌طور مثال، داده‌هایی که با مقیاس ترتیبی جمع‌آوری شده‌اند، با استفاده از شاخص ضریب تشابه متریک، تبدیل به داده متریک شوند و برعکس. به‌عبارت دیگر، در تجزیه‌وتحلیل داده‌ها، ویژگی‌های متریک و ترتیبی در هم ادغام می‌شوند. از معایب این تغییرات می‌توان به از دست دادن اطلاعات در هنگام تغییر مقیاس متریک به ترتیبی و افزایش اطلاعات در هنگام تغییر از مقیاس ترتیبی به متریک اشاره کرد که این پدیده ازنظر ریاضی درست نیست (Podani, 2005). برای مثال، در روش طبقه‌بندی خوشه‌ای برای محاسبه ماتریس

فاصله بین متغیرها از شاخص‌های ضریب تشابه یا عدم تشابه متریک (مانند ضریب فاصله اقلیدوسی) استفاده می‌شود (Everitt *et al.*, 2011)، درصورتی‌که روش مناسب با ساختار داده‌های ترتیبی در محاسبه این ماتریس درنظر گرفته نشود، نتایج ازنظر آماری، اشکال خواهند داشت که سبب تغییر ماهیت داده‌ها از یک مرحله به مرحله دیگر می‌شود. پژوهش‌های بسیار کمی به این نکته اشاره کرده‌اند (Yang *et al.*, 2007)، برای نمونه، Podani (۲۰۰۵) روش ضریب فاصله Gower به‌همراه روش اتصال خوشه نزدیک‌ترین همسایه را برای داده‌های ترتیبی پیشنهاد کرد. همچنین، Yang و همکاران (۲۰۰۷) گزارش کردند که روش‌های OrdClAn و NMDS براساس شاخص Gower همبستگی معنی‌داری بین ماتریس فاصله و دندروگرام ایجاد می‌کنند. با این‌حال، پژوهش‌های بی‌شماری به این نکته بی‌توجه بوده‌اند (Lengyel & Podani, 2015; Hüllbusch *et al.*, 2016; Mokaram Kashtiban *et al.*, 2018)، بنابراین در روش خوشه‌بندی، نوع داده ورودی است که خروجی طبقه‌بندی را تحت تأثیر قرار می‌دهد. به‌این‌ترتیب، تشخیص درستی یا نادرستی نتایج به‌دست‌آمده، بسیار مشکل خواهد بود (Yang *et al.*, 2007). برای نمونه، در برداشت‌های جامعه‌شناسی که داده‌های پوشش علفی براساس مقیاس براون- بلانکه تخمین زده می‌شوند، نوع داده‌ها ترتیبی است (Podani, 2005). از آنجایی‌که اندازه‌گیری دقیق پوشش و یا فراوانی گونه‌های علفی، بسیار وقت‌گیر و مشکل است، پژوهشگران سعی می‌کنند با استفاده از مقیاس براون- بلانکه که بازه‌ای از سطح پوشش را در اختیار پژوهشگر قرار می‌دهد، سطح پوشش یک گونه در قطعه‌نمونه را برآورد کنند. باتوجه‌به ماهیت برآوردی و غیرخطی این اعداد در روش طبقه‌بندی خوشه‌ای، به‌جای میانگین درصد تاج‌پوشش هر طبقه که برای یک گونه اختصاص می‌یابد، از داده ترتیبی متناسب با آن طبقه استفاده می‌شود.

باید به این نکته توجه داشت که در طبقه‌بندی داده‌ها نمی‌توان با داده‌های ترتیبی همانند داده‌های کمی عمل کرد. بهترین رویکرد برای تجزیه‌وتحلیل داده‌های ترتیبی، استفاده از روشی است که مقیاس داده‌ها را تغییر ندهد (De Cáceres *et al.*, 2010). باتوجه‌به اینکه استفاده از روش‌های متریک ضرایب تشابه در داده‌های ترتیبی، مقیاس داده‌ها را تغییر می‌دهد، ضروری است که برای تجزیه‌وتحلیل ماتریس داده‌های ترتیبی از ضرایب تشابه مناسب داده‌های ترتیبی استفاده شود (Podani, 2005). باتوجه‌به استفاده گسترده از داده‌های ترتیبی در علوم زیستی، اهمیت تجزیه‌وتحلیل داده‌ها براساس مقیاس داده‌ها و باتوجه‌به مشکلات یادشده درمورد داده‌هایی که با مقیاس براون- بلانکه برداشت می‌شوند، انتخاب مقیاس فاصله و روش اتصال خوشه‌های مناسب در تجزیه‌وتحلیل خوشه‌ای پوشش گیاهی بسیار مهم هستند (Lewis, 2004)، بنابراین هدف از پژوهش پیش‌رو، مقایسه روش‌های مختلف تجزیه‌وتحلیل خوشه‌ای در پوشش گیاهی بخشی از جنگل‌های بلوط کرمانشاه بود.

مواد و روش‌ها

منطقه مورد مطالعه

این پژوهش در دامنه‌های جنوبی جنگل‌های چهارزبر (واقع در ۳۴ کیلومتری شهرستان کرمانشاه) با طول جغرافیایی ′۳۹ °۴۶ تا ′۴۹ °۴۶ شرقی و عرض جغرافیایی ′۹ °۳۴ تا ′۱۴ °۳۴ شمالی انجام شد. میانگین بارش سالانه این منطقه ۴۸۹ میلی‌متر است که ۸۰ درصد آن در فصل‌های پاییز و زمستان اتفاق می‌افتد. محاسبه نمایه خشکی دومارتن، اقلیم مدیترانه‌ای سرد را در این منطقه نشان می‌دهد. کمینه و بیشینه ارتفاع از سطح دریا در این منطقه به‌ترتیب برابر با ۱۴۰۰ و ۱۸۰۰ متر هستند. پوشش درختی غالب منطقه از بلوط (*Quercus brantii* Lindl.) تشکیل شده است (Eshaghi Rad *et al.*, 2017).

روش پژوهش

سه قطعه جنگلی به‌مساحت ۴/۵، پنج و ۵/۵ هکتار و

با جهت جنوبی از جنگل‌های بلوط منطقه با شرایط مشابه ازنظر شیب و ارتفاع از سطح دریا انتخاب شد. در هر قطعه با استفاده از سه خط‌نمونه با طول ۱۵۰ متر که در فاصله‌های ۲۰۰ متری از هم قرار گرفته بودند و در جهت شیب غالب پیاده شدند، نمونه‌برداری از پوشش گیاهی انجام شد (اولین خط‌نمونه به‌صورت تصادفی پیاده شد) (شکل ۱). نقاط برداشت پوشش گیاهی در هر خط‌نمونه در فاصله‌های صفر، ۲۵، ۵۰، ۱۰۰ و ۱۵۰ متری تعیین شد (Gehlhausen *et al.*, 2000). برای برداشت پوشش علفی نیز در هر نقطه برداشت، پنج قطعه‌نمونه ۰/۲۵ متر مربعی (۰/۵ × ۰/۵ متر مربع) به فاصله یک متر (Euskirchen *et al.*, 2001) عمود بر خط‌نمونه در سمت چپ و راست پیاده شد (Gehlhausen *et al.*, 2000). در هر قطعه‌نمونه، نوع گونه ثبت شد و فراوانی و درصد پوشش گونه‌های گیاهی براساس مقیاس براون- بلانکه تخمین زده شد (شکل ۱). گونه‌های جمع‌آوری‌شده به هرباریوم مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی استان کرمانشاه منتقل شدند و تاکسون‌های گیاهی شناسایی شدند.

شکل ۱- موقعیت خط‌نمونه‌ها، نقاط برداشت و ریزنمونه‌ها در هر قطعه جنگلی

تجزیه‌وتحلیل داده‌ها

در این پژوهش، داده‌های پوشش علفی برای هر نقطه از نمونه‌برداری خط‌نمونه درنظر گرفته شد. ابتدا ماتریس گونه‌ها براساس مقادیر کیفی (کدهای مقیاس براون- بلانکه) برای ۱۳۰ گونه شناسایی‌شده و ۴۳ قطعه‌نمونه تشکیل شد. دو قطعه‌نمونه به‌دلیل شرایط نامناسب محیطی و عدم یکنواختی پوشش گیاهی، ترکیب فلورستیک مناسبی برای ورود به تجزیه‌وتحلیل داده‌ها نداشتند، بنابراین از ابتدا از ماتریس داده‌ها حذف شدند. سپس از تجزیه‌وتحلیل خوشه‌ای برای طبقه‌بندی پوشش گیاهی جوامع بلوط منطقه استفاده شد. در این روش ابتدا ماتریس فاصله بین قطعه‌نمونه‌ها با فرمول Gower محاسبه شد (Podani, 1999) (رابطه ۱). این روش، توانایی محاسبه فاصله بین قطعه‌نمونه‌ها با انواع مختلف داده‌ها همچون داده‌های ترتیبی و داده‌های کمی را دارد، اما روش‌های دیگر محاسبه فاصله بین قطعه‌نمونه‌ها همچون روش‌های اقلیدوسی و ضریب

همبستگی پیرسون این توانایی را ندارند (Podani, 1999).

$$S_{ij} = \frac{\sum_k^n w_{ijk} s_{ijk}}{\sum_k^n w_{ijk}} \qquad \text{رابطه (۱)}$$

که در آن: Sijk فاصله بین i و j در متغیر k و wijk وزن متغیر k بین مشاهدات i و j است.

برای ادغام خوشه‌ها از روش‌های مختلفی به‌شرح زیر استفاده شد:

– روش نزدیک‌ترین همسایه (Single linkage): در این روش، اتصال بین دو خوشه براساس کمترین فاصله بین یک عضو از یک گروه با یک عضو از گروه دیگر است (Everitt *et al*., 2011).

– روش دورترین همسایه (Complete linkage): در این روش، اتصال بین دو خوشه براساس بیشترین فاصله بین یک عضو از یک گروه با یک عضو از گروه دیگر است (Everitt *et al*., 2011).

– روش اتصال میانگین (Average linkage): در این روش برای تعیین اتصال بین دو خوشه، متوسط فاصله بین خوشه‌ها مورد توجه قرار می‌گیرد (Everitt *et al*., 2011).

– روش وارد (Ward's method): در این روش از مجموع مربعات تفاضل هر داده یک خوشه نسبت به بردار میانگین آن خوشه، معیاری برای سنجش یک خوشه استفاده می‌شود. الگوریتم زیر را می‌توان برای روش وارد درنظر گرفت (Everitt *et al*., 2011):

۱– ابتدا هر داده به‌عنوان یک خوشه درنظر گرفته می‌شود.

۲– به‌ازای تمام جفت خوشه‌های ممکن از مجموعه خوشه‌ها، آن دو خوشه‌ای که مجموع مربعات تفاضل داده‌های خوشه حاصل از اجتماع آن‌ها با بردار میانگین خوشه حاصل، کمینه باشد، انتخاب می‌شوند.

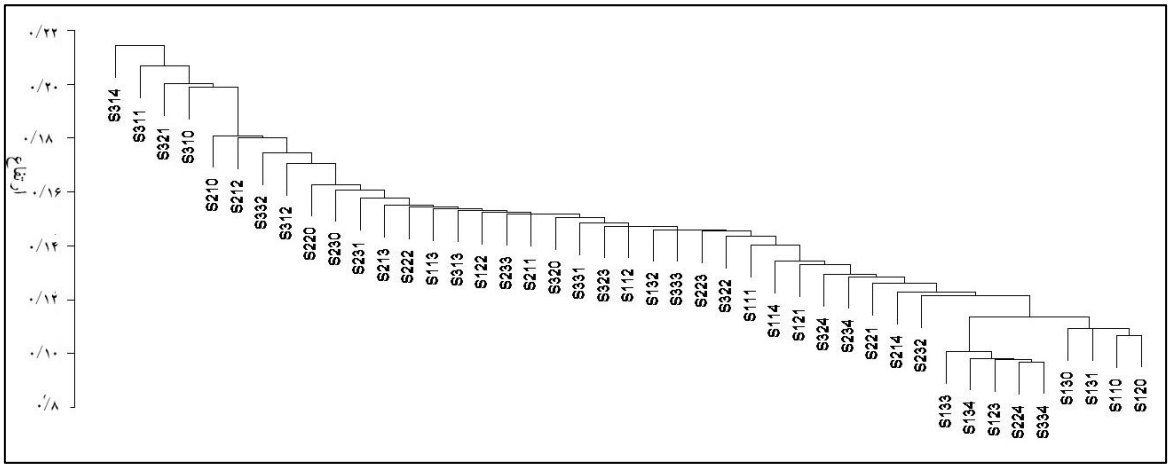
۳– دو خوشه انتخاب‌شده باهم ترکیب می‌شوند.

برای ارزیابی مقدار انطباق بین دندروگرام و ماتریس فاصله از ضریب همبستگی کوفنتیک (Cophenetic correlation coefficient) استفاده شد. هرچه اندازه ضریب همبستگی بیشتر باشد، انطباق بین ماتریس فاصله محاسبه‌شده و دندروگرام حاصل از روش‌های مختلف ادغام گونه‌ها بیشتر است. درنتیجه، ماهیت داده، کمتر تغییر می کند. ضریب همبستگی بیشتر از ۰/۹ نشان‌دهنده انطباق زیاد بین ماتریس فاصله و دندروگرام است. اگر این ضریب کمتر از ۰/۷۴ باشد، انطباقی بین ماتریس فاصله و دندروگرام وجود ندارد (Yang *et al*., 2007). برای تعیین تعداد بهینه خوشه‌ها و ارزیابی کیفیت خوشه‌ها از روش سیلوئت استفاده شد. این معیار، فاصله‌های درون‌خوشه‌ای و برون‌خوشه‌ای را هم زمان درنظر می‌گیرد. شاخص سیلوئت برای سنجش اعتبار خوشه‌ها، نسبت بین مجموع مربعات درون‌خوشه‌ای و مجموع مربعات بین‌خوشه‌ای را می‌سنجد. این شاخص همواره مقدار بین یک تا ۱– را کسب می‌کند. اگر شاخص سیلوئت نزدیک به یک باشد، نشان‌دهنده ساختار قوی خوشه‌بندی است، اما اگر این شاخص صفر باشد، خوشه‌بندی، ساختاری ضعیف دارد. اگر به شاخص مذکور، عدد منفی تعلق گیرد، به‌این معنی است که خوشه‌بندی هیچ ساختار منطقی ندارد (El-Serag, 2012).

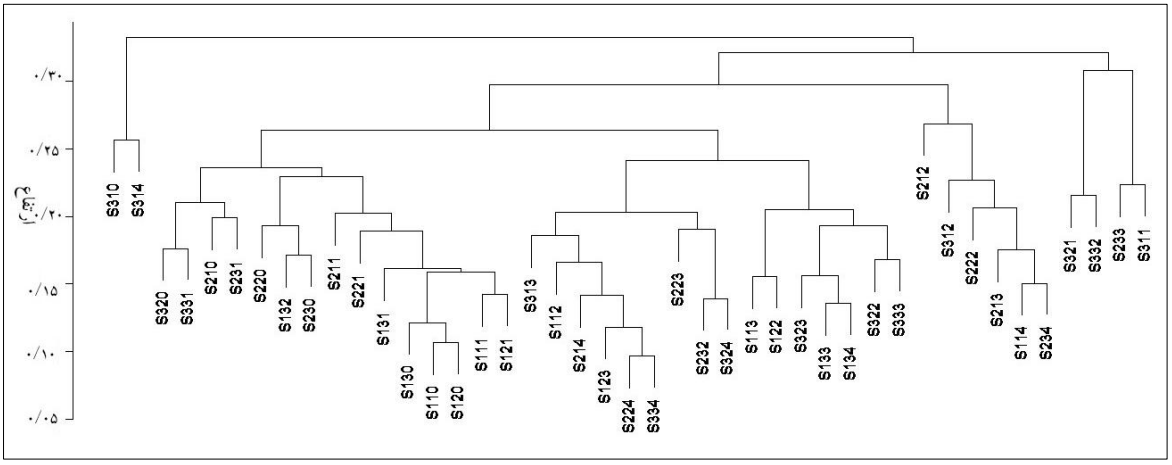
تمام تجزیه‌وتحلیل‌ها در نرم‌افزار R 3.5.1 و با استفاده از بسته‌های Cluster، Gower، Rtsne و Ggplot2 انجام شد.

نتایج

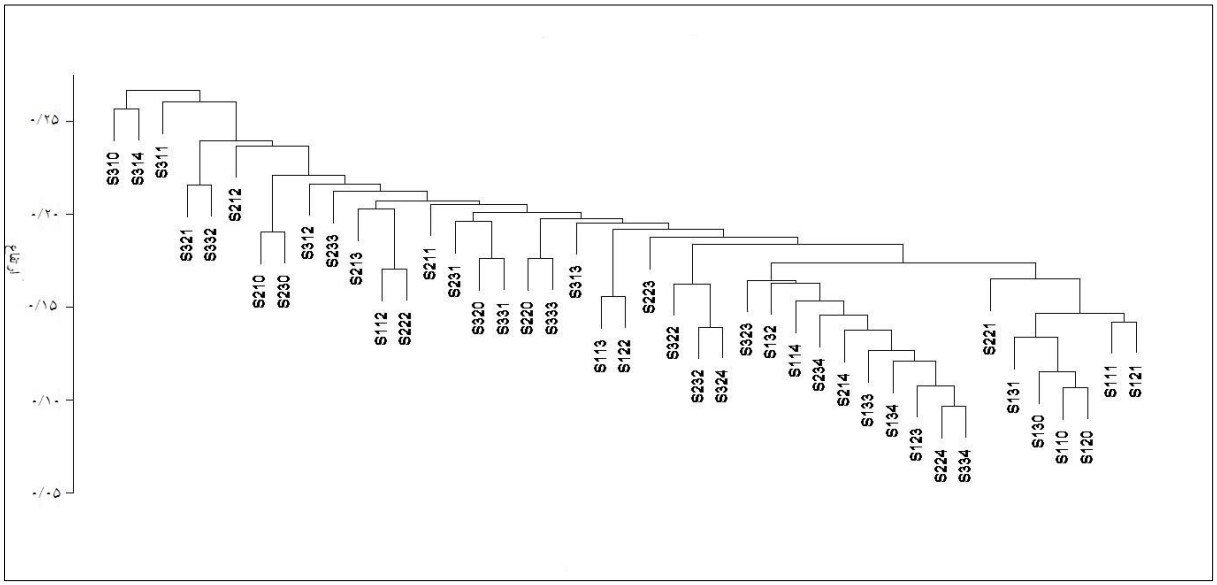
در ابتدا خوشه‌بندی براساس ۱۳۰ گونه و ۴۳ قطعه‌نمونه انجام شد. نتایج مربوط به روش تجزیه‌وتحلیل خوشه‌ای حاصل از روش ضریب فاصله Gower و روش ادغام نزدیک‌ترین همسایه در شکل ۲، دورترین همسایه در شکل ۳، روش میانگین در شکل ۴ و روش اتصال وارد در شکل ۵ نشان داده شده است. در شکل‌های مذکور، عدد صدگان: شماره قطعه، عدد دهگان: شماره خط‌نمونه، عدد یکان: فاصله نمونه‌برداری (صفر: حاشیه، یک: فاصله ۲۵ متر، دو: فاصله ۵۰ متر، سه: فاصله ۱۰۰ متر و چهار: فاصله ۱۵۰ متر) هستند.



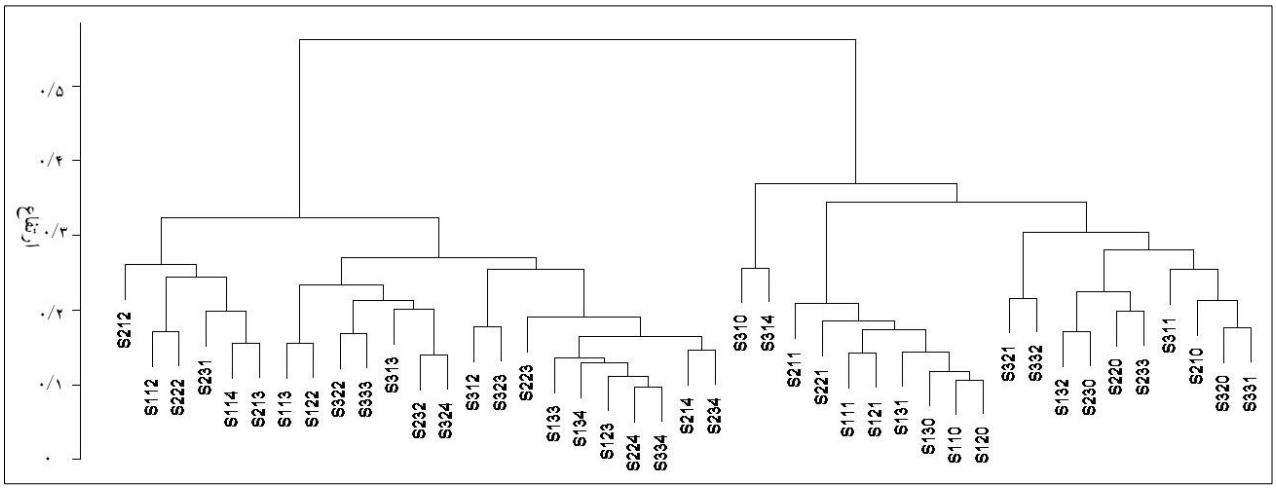
شکل ۲- طبقه‌بندی به‌دست‌آمده از تجزیه‌وتحلیل خوشه‌ای جوامع بلوط منطقه مورد مطالعه با استفاده از روش اتصال نزدیک‌ترین همسایه (ارتفاع نشان‌دهنده مقدار تشابه یا عدم تشابه بین دو قطعه‌نمونه است.)



شکل ۳- طبقه‌بندی به‌دست‌آمده از تجزیه‌وتحلیل خوشه‌ای جوامع بلوط منطقه مورد مطالعه با استفاده از روش اتصال دورترین همسایه

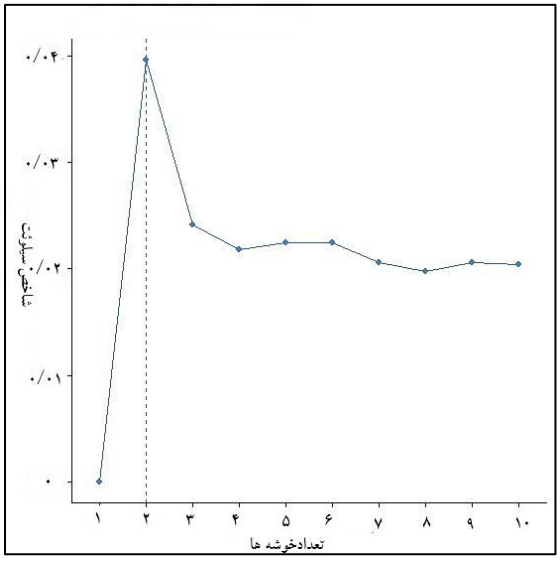


شکل ۴- طبقه‌بندی به‌دست‌آمده از تجزیه‌وتحلیل خوشه‌ای جوامع بلوط منطقه مورد مطالعه با استفاده از روش اتصال میانگین



شکل ۵- طبقه‌بندی به‌دست‌آمده از تجزیه‌وتحلیل خوشه‌ای جوامع بلوط منطقه مورد مطالعه با استفاده از روش اتصال وارد

شکل ۶ تعداد بهینه خوشه‌ها با استفاده از معیار سیلوئت را برای این مجموعه داده نشان می‌دهد. تعداد خوشه‌ها با معیار سیلوئت از یک تا ۱۰ متغیر است. براساس این معیار در جوامع بلوط منطقه مورد مطالعه، بهترین تعداد خوشه در مرحله دو قرار داشت.



شکل ۶- نمودار روند مقدار تغییرات معیار سیلوئت برای تعداد خوشه‌های مختلف

در جدول ۱ ضریب همبستگی کوفنتیک بین روش فاصله Gower و روش‌های مختلف اتصال خوشه ارائه شده است. براساس نتایج، بیشترین مقدار همبستگی بین روش فاصله Gower و روش اتصال خوشه میانگین مشاهده شد. کمترین مقدار همبستگی نیز متعلق به روش Gower و روش اتصال دورترین همسایه بود.

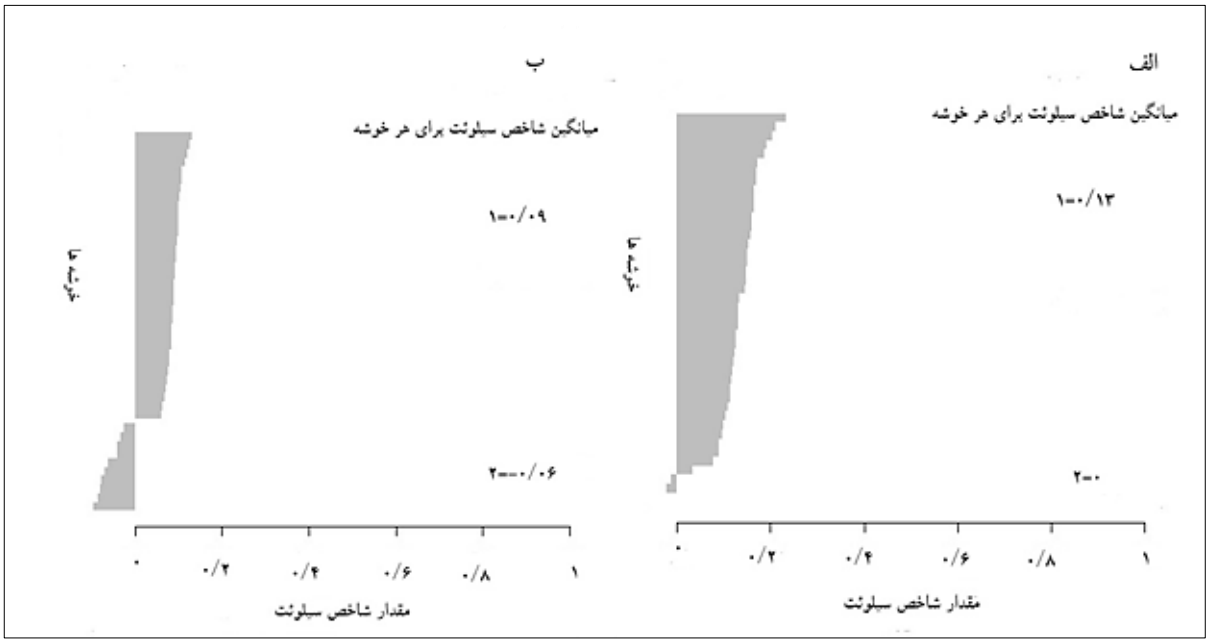
جدول ۱- ضریب همبستگی کوفنتیک بین روش فاصله و روش‌های مختلف خوشه‌بندی

مقدار ضریب همبستگی	روش اتصال خوشه	روش فاصله
۰/۸۷	روش نزدیک‌ترین همسایه	روش Gower
۰/۴۵	روش دورترین همسایه	
۰/۸۹	روش میانگین	
۰/۷	روش وارد	

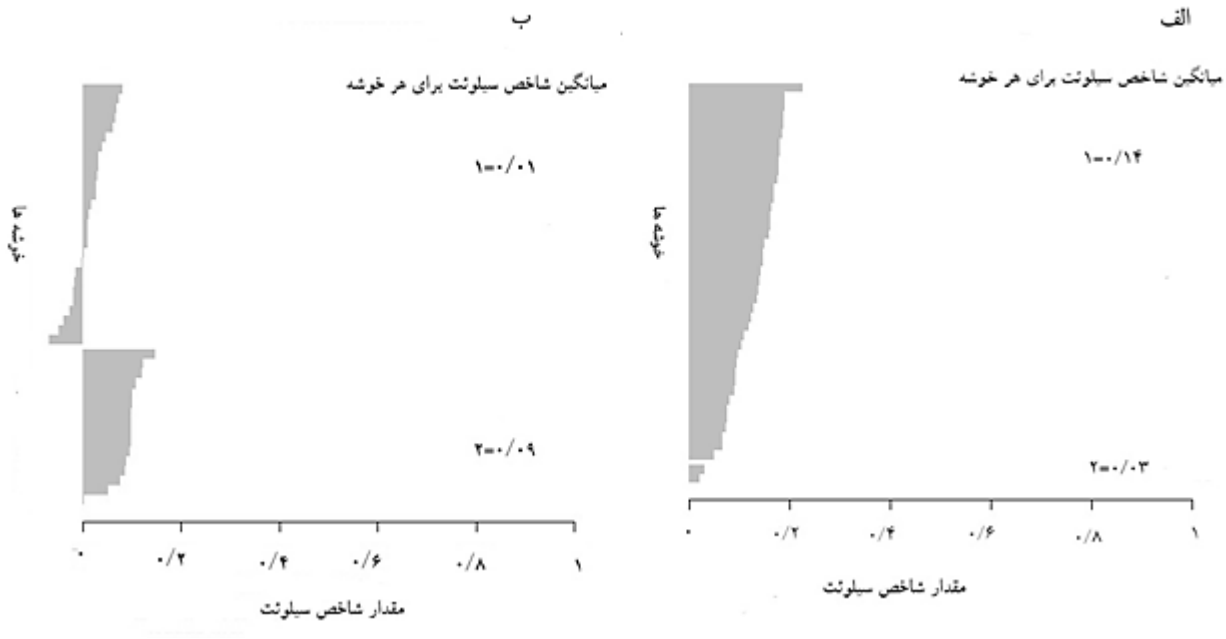
در شکل‌های ۷ و ۸، کیفیت روش‌های مختلف خوشه‌بندی مورد مطالعه براساس معیار سیلوئت ارائه شده است. باتوجه‌به نتایج، میانگین شاخص سیلوئت در روش‌های اتصال نزدیک‌ترین همسایه و اتصال میانگین برابر با ۰/۱۳ بود، بنابراین دو روش مذکور نسبت به روش اتصال دورترین همسایه با میانگین شاخص سیلوئت ۰/۰۶ و روش اتصال وارد با میانگین شاخص سیلوئت ۰/۰۴ ساختار خوشه‌بندی بهتری را ارائه دادند (جدول ۲).

جدول ۲- میانگین شاخص سیلوئت در روش‌های مختلف خوشه‌بندی

میانگین شاخص سیلوئت	روش اتصال خوشه	روش فاصله
۰/۱۳	روش نزدیک‌ترین همسایه	روش Gower
۰/۰۶	روش دورترین همسایه	
۰/۱۳	روش میانگین	
۰/۰۴	روش وارد	



شکل ۷- کیفیت خوشه‌بندی براساس معیار سیلوئت الف) روش Gower و روش اتصال نزدیک‌ترین همسایه ب) روش Gower و روش اتصال دورترین همسایه



شکل ۸- کیفیت خوشه‌بندی براساس معیار سیلوئت الف) روش Gower و روش اتصال میانگین ب) روش Gower و روش اتصال وارد

بحث

در پژوهش پیش‌رو، روش‌های مختلف خوشه‌بندی برای دستیابی به روش مناسب برای داده‌های ترتیبی ارائه شد. باتوجه‌به نتایج به‌دست‌آمده، روش‌های میانگین و نزدیک‌ترین همسایه، ضریب همبستگی بیشتری نسبت به دو روش دیگر داشتند. همان‌طور که در منابع مختلف بیان شده است، از ویژگی‌های مهم روش اتصال نزدیک‌ترین همسایه، تغییرناپذیری آن است، به‌این‌معنی که در هنگام اتصال خوشه‌ها، این روش منجر به تغییر فاصله محاسبه‌شده داده‌ها نمی‌شود (Everitt *et al.*, 2011). همچنین، روش اتصال میانگین یک روش بسیار مؤثر برای یافتن خوشه‌ها است که بسیار توصیه می‌شود (Ken *et al.*, 2008)، اما روش اتصال خوشه دورترین همسایه، کمترین مقدار همبستگی را داشت. این موضوع نشان می‌دهد که در مسیر انتقال اطلاعات از ماتریس فاصله به مرحله بعد و نمایش دندروگرام، اطلاعات زیادی از دسترس خارج شده یا به اطلاعات افزوده شده است. روش اتصال وارد نیز همبستگی کمتر از ۰/۸ را نشان داد. این روش یکی از روش‌های مؤثر در خوشه‌بندی است، اما فقط زمانی‌که مجموعه داده‌ها، توزیع کروی داشته باشند، عملکرد قابل قبولی دارد (Peet & Robert, 2013). در گام بعدی این پژوهش، تعداد بهینه خوشه‌ها با معیار سیلوئت تعیین شد. سپس اعتبار نتایج خوشه‌بندی به‌عنوان یکی از مسائل مهم طبقه‌بندی داده‌ها (Hämäläinen *et al.*, 2017) بررسی شد. برای این منظور، کیفیت خوشه‌بندی روش‌های مختلف براساس تعداد بهینه خوشه‌های تعیین‌شده، ارزیابی شد. باتوجه‌به نتایج، میانگین شاخص سیلوئت برای روش فاصله Gower و روش‌های اتصال خوشه نزدیک‌ترین همسایه و میانگین بیشتر از روش‌های اتصال دورترین همسایه و وارد بود. در پژوهشی که Lengyel و Botta-Dukát (۲۰۱۸) درمورد روش‌های مختلف خوشه‌بندی سلسله‌مراتبی براساس معیار سیلوئت انجام دادند، دو روش اتصال دورترین همسایه و اتصال میانگین را برای داده‌های کمی، روش‌های مؤثرتری معرفی کردند. Shakeri و همکاران (۲۰۱۱) نیز براساس شاخص سیلوئت نتیجه گرفتند که بهترین عملکرد متعلق به روش خوشه‌بندی سلسله‌مراتبی تجمعی با روش اتصال میانگین بود. در

پژوهشی که Ken و همکاران (۲۰۰۸) برای مقایسه روش‌های مختلف خوشه‌بندی انجام دادند، اغلب روش نزدیک‌ترین همسایه و روش میانگین، عملکرد خوبی داشتند. روش اتصال میانگین نسبت به روش تجزیه‌وتحلیل دوطرفه گونه‌های شاخص نیز عملکرد بهتری داشت (Belbin & McDonald, 1993; Cao *et al*., 1997).

باتوجه‌به نتایج مربوط به میانگین شاخص سیلوئت برای هر خوشه مشاهده می‌شود که خوشه دوم در روش اتصال نزدیک‌ترین همسایه، کیفیت خوشه‌بندی بسیار اندکی دارد، اما درمقابل، روش اتصال میانگین هر دو خوشه، کیفیت خوشه‌بندی قابل قبولی را ارائه می‌کند. از این‌رو، می‌توان روش میانگین را به‌عنوان روش مناسب اتصال خوشه همراه با ماتریس فاصله Gower برای داده‌های ترتیبی پیشنهاد کرد تا بدون از دست دادن یا افزودن اطلاعات، مسیر درستی را برای طبقه‌بندی داده‌های ترتیبی طی کرد.

منابع مورد استفاده

- Alamgir, M., Turton, S.M., Macgregor, C.J. and Pert, P.L., 2016. Ecosystem services capacity across heterogeneous forest types: understanding the interactions and suggesting pathways for sustaining multiple ecosystem services. Science of the Total Environment, 566-567: 584-595
- Belbin, L. and McDonald, C., 1993. Comparing three classification strategies for use in ecology. Journal of Vegetation Science, 4(3): 341-348.
- Cao, Y., Bark, A.W. and Williams, W.P., 1997. A comparison of clustering methods for river benthic community analysis. Hydrobiologia, 347(1-3): 25-40.
- Damgaard, C., 2014. Estimating mean plant cover from different types of cover data: a coherent statistical framework. Ecospher, 5(2): 1-7.
- De Cáceres, M., Font, X. and Oliva, F., 2010. The management of vegetation classifications with fuzzy clustering. Journal of Vegetation Science, 21(6): 1138-1151.
- El-Serag, H.B., 2012. Epidemiology of viral hepatitis and hepatocellular carcinoma. Gastroenterology, 142(6):1264-1273.
- Eshaghi Rad, J., Soleimani, F. and Khodakarami, Y., 2017. Comparison of flora at the edge and within oak forests in southern slopes of Kermanshah forests. Applied Biology, 30(1): 19-35 (In Persian).
- Euskirchen, E.S., Chen, J. and Bi, R., 2001. Effects of edges on plant communities in a managed landscape in northern Wisconsin. Forest Ecology and Management, 148(1-3): 93-108.
- Everitt, B.S., Landau, S., Leese, M. and Stahl, D., 2011. Cluster Analysis, 5th Edition. John Wiley & Sons, Ltd., Chichester, UK, 346p.
- Gehlhausen, S.M., Schwartz, M.W. and Augspurger, C.K., 2000. Vegetation and microclimatic edge effects in two mixed-mesophytic forest fragments. Plant Ecology, 147(1): 21-35.
- Gill, D. and Tipper, J.C., 1978. The adequacy of non-metric data in geology: tests using a divisive omnithetic clustering technique. The Journal of Geology, 86(2): 241-259.
- Grabherr, G., Reiter, K. and Willner, W., 2003. Towards objectivity in vegetation classification: the example of the Austrian forests. Plant Ecology, 169(1): 21- 34
- Hall, M. and Richardson, T., 2016. Basic statistics for comparing categorical data from 2 or more groups. Hospital Pediatrics, 6(6): 383-385.
- Hämäläinen, J., Jauhiainen, S. and Kärkkäinen, T., 2017. Comparison of internal clustering validation indices for prototype-based clustering. Algorithms, 10(3): 105.
- Hüllbusch, E., Brandt, L.M., Ende, P. and Dengler, J., 2016. Little vegetation change during two decades in a dry grassland complex in the Biosphere Reserve Schorfheide-Chorin (NE Germany). Tuexenia, 36: 395-412.
- Ken, A., 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
- Lechner, A.M., McCaffrey, N., McKenna, P., Venables, W.N. and Hunter, J.T., 2016. Ecoregionalization classification of wetlands based on a cluster analysis of environmental data. Applied Vegetation Science, 19(4): 724-735.
- Lengyel, A. and Botta-Dukát, Z., 2018. Silhouette width using generalized mean – a flexible method for assessing clustering efficiency. Biorixv, DOI: 10.1101/434100.
- Lengyel, A. and Podani, J., 2015. Assessing the relative importance of methodological decisions in classification of vegetation data. Journal of Vegetation Science, 26(4): 804-815.
- Lewis, K.P., 2004. How important is the statistical approach for analyzing categorical data? A critique using artificial nests. Oikos, 104(2): 305-315.
- Mahmoodi, M., Ramezani, E., Eshaghi-Rad, J. and

Heidari Rikan, M., 2015. On the relationship between vegetation cover and physiographic factors in a gallery forest in southern Urmia, NW Iran. Iranian Journal of Forest and Poplar Research, 23(2): 279-293 (In Persian).
- McGranahan, D.A., Engle, D.M., Fuhlendorf, S.D., Miller, J.R. and Debinski, D.M., 2013. Multivariate analysis of rangeland vegetation and soil organic carbon describes degradation, informs restoration and conservation. Land, 2(3): 328-350.
- Mokaram Kashtiban, S., Mousavi Mirkala, S.R. and Eshaghi Rad, J., 2018. Effect of traditional utilization on woody species composition and diversity through Detrended Correspondence Analyses in Sardasht Forests (West Azerbaijan Province). Journal of Forest Research and Development, 4(3): 363-376 (In Persian).
- Pazouki, M., Sepehri, M.M. and Saberifiroozi, M., 2014. Discovering hidden cluster structures in patients with cirrhosis based on laboratory data. Govaresh, 19(3): 191-197 (In Persian).
- Peet, R.K. and Roberts, D.W., 2013. Classification of natural and semi-natural vegetation: 26-62. In: van der Maarel, E. and Franklin, J. (Eds.). Vegetation Ecology, 2nd Edition. Wiley-Blackwell, Chichester, UK, 572p.
- Podani, J., 1999. Extending Gower's general coefficient of similarity to ordinal characters. Taxon, 48(2): 331-340.
- Podani, J., 2005. Multivariate exploratory analysis of ordinal data in ecology: Pitfalls, problems and solutions. Journal of Vegetation Science, 16(5): 497-510.
- Shakeri, M.T., Sabaghian, E. and Esmaeili, H., 2011. CCK (Clustering-Classification-Kappa); a new validation index to assessing clustering results of gene expression data. Journal of North Khorasan University of Medical Sciences, 3: 67-78 (In Persian).
- Suh, J.P., Roh, J.H., Cho, Y.C., Han, S.S., Kim, Y.G. and Jena, K.K., 2009. The pi40 gene for durable resistance to rice blast and molecular analysis of pi40-advanced backcross breeding lines. Phytopathology, 99(3): 243-250.
- Vavrek, M.J., 2016. A comparison of clustering method for biogeography with fossil datasets. PeerJ, 4: e1720.
- Yang, S.Z., Feng, Y.Y. and Yeh, F.Y., 2007. Application of ordinal clustering to the taxonomy of the genus Entada (Fabaceae) in Taiwan. Bangladesh Journal of Plant Taxonomy, 14(2): 93-100
- Zuur, A.F., Leno, E.N. and Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution, 1(1): 3-14.

Comparison of different methods for cluster analysis (Case study: Kermanshah oak forests)

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

1- Ph.D. Student, Department of Forestry, Faculty of Natural Resources, Urmia University, Urmia, Iran
2*- Corresponding author, Associate Prof., Department of Forestry, Faculty of Natural Resources, Urmia University, Urmia, Iran
E-mail: javad.eshaghi@yahoo.com
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: 12.03.2019 Accepted: 10.06.2019

Abstract

Vegetation classification is an essential tool to describe, understand, predict and manage ecosystems. The aim of this study was to compare different types of hierarchical clustering. Three forest patches with similar slope and altitude gradients located on the southern slopes of Chahar Zebar forests, Kermanshah province, were selected. Vegetation sampling in each patch was conducted at 0, 25, 50, 100 and 150-meter distances along three transects that were 200 m apart. Cluster analysis was used for the classification of samples. Amongst the applied methods, Gower's distance (or similarity) initially computes distances between pairs of variables over data sets and then merges those distances with the nearest neighbor, complete neighbor, average neighbor, and Ward's method. The optimal number and quality of clusters were evaluated with silhouette criteria. In addition, the Cophenetic correlation coefficient was computed for evaluating the correlation between the dendrogram and the distance matrix. Results showed that two was the optimal number of clustering for oak stands. Moreover, the Cophenetic correlation coefficient between the distance matrix and the nearest neighbor and average method was higher than that returned between complete neighbor and Ward's method. Based on silhouette criteria, the nearest neighbor and average methods were associated with higher cluster quality compared with two other methods. However, the mean value of the silhouette index was low for the second cluster of the nearest neighbor method. Considering the disadvantages of the nearest neighbor, the average method is suggested for clustering categorical data.

Keywords: Classification, clustering, Gower distance, ordinal number.