

خوشه‌بندی مصرف آب شرب شهری با رویکرد فازی

سیدعارف موسوی^۱، مسعود عسگرزاده طرقله^۲، امیرحسین حسینی^۳، سیدمحسن میرمفیدی^۴، علی عباسی^{۵*}

۱- دانشجوی کارشناسی ارشد مهندسی و مدیریت منابع آب، گروه مهندسی عمران دانشگاه فردوسی مشهد (arref3@gmail.com)

۲- دانشجوی کارشناسی ارشد مهندسی و مدیریت منابع آب، گروه مهندسی عمران دانشگاه فردوسی مشهد (as.masood30@gmail.com)

۳- دانشجوی دکتری مهندسی آب و سازه‌های هیدرولیکی، گروه مهندسی عمران دانشگاه فردوسی مشهد (amirhoseini632@gmail.com)

۴- دانشجوی کارشناسی ارشد مهندسی و مدیریت منابع آب، گروه مهندسی عمران دانشگاه فردوسی مشهد (mohsenmirmofidi99@gmail.com)

۵- عضو هیئت علمی گروه مهندسی عمران دانشگاه فردوسی مشهد (نویسنده مسئول) (aabbasi@um.ac.ir)

چکیده

با رشد روزافزون جمعیت، گسترش شهرنشینی و تغییرات اقلیمی، مدیریت تقاضای آب شهری به یک چالش اساسی در حوزه منابع آب تبدیل شده است. در این راستا، تحلیل الگوهای مصرف و شناسایی رفتار مصرف‌کنندگان نقش مهمی در بهینه‌سازی سامانه‌های آبرسانی ایفا می‌کند. پژوهش حاضر با هدف تحلیل رفتار مصرف آب در بازه زمانی ماهانه، از الگوریتم خوشه‌بندی فازی C-میانگین (FCM) بهره گرفته است. برای این منظور داده‌های مصرف آب مربوط به کنتورهای هوشمند در منطقه Sunshine Coast استرالیا در سه ماه متوالی (می، ژوئن و ژوئیه) برای تحلیل استفاده شد. در این پژوهش، ابتدا پیش‌پردازش بر روی داده‌ها با فواصل قرائت ۱۵ دقیقه‌ای انجام شده و سپس با تبدیل مقیاس زمانی داده‌ها به مقیاس زمانی ماهانه، الگوریتم FCM بر روی ۲۸۰ کنتور اجرا شد. تعداد خوشه‌ها به صورت تجربی برابر با سه در نظر گرفته شد تا گروه‌های کم‌مصرف، مصرف متوسط و پرمصرف تفکیک گردند. نتایج حاصل نشان داد که توزیع مصرف در سه ماه نسبتاً پایدار بوده و اکثر کنتورها در طول این دوره در یک دسته خاص باقی مانده‌اند. نمودارهای برش آلفا نیز به خوبی مرزهای سخت‌تر میان خوشه‌ها را نمایان ساخته‌اند. استفاده از خوشه‌بندی فازی، با در نظر گرفتن درجات عضویت، امکان تحلیل انعطاف‌پذیری از رفتار مصرف آب فراهم نموده و در سیستم‌های هوشمند آبرسانی می‌تواند ابزاری برای مدیریت هدفمند مصرف و شناسایی ناهنجاری‌ها تلقی شود. نتایج این تحقیق می‌تواند زمینه‌ساز توسعه ابزارهای پیش‌بینی، کنترل هوشمند و سیاست‌گذاری‌های دقیق‌تر در مدیریت پایدار منابع آب شهری باشد.

واژه‌های کلیدی: الگوی مصرف آب شرب، خوشه‌بندی فازی C-میانگین، درجه عضویت، برش آلفا

۱- مقدمه

با افزایش فشار بر منابع آبی ناشی از رشد جمعیت، تغییرات اقلیمی و گسترش شهرنشینی، مدیریت بهینه مصرف آب به یکی از چالش‌های اساسی در حوزه مهندسی منابع آب تبدیل شده است. در این راستا، تحلیل الگوهای مصرف و پیش‌بینی رفتار کاربران به کمک روش‌های داده‌کاوی، به‌ویژه خوشه‌بندی، جایگاه ویژه‌ای یافته است. خوشه‌بندی به عنوان یکی از ابزارهای مهم در تحلیل داده‌های بدون برچسب، امکان تفکیک کاربران، مناطق با ویژگی‌های مشابه و ... را فراهم می‌سازد، در نتیجه خوشه‌بندی پیچیدگی‌های موجود در داده‌ها را کاهش می‌دهد و آن‌ها را برای استفاده در الگوریتم‌های دیگر آماده می‌کند؛ این موضوع در سیستم‌های آبرسانی هوشمند، به‌ویژه در طراحی راهکارهای هدفمند برای مدیریت تقاضا و کاهش تلفات، اهمیت دوچندان می‌یابد.

در میان روش‌های مختلف خوشه‌بندی، خوشه‌بندی فازی (Fuzzy Clustering) نسبت به روش‌های سخت (Hard Clustering) مزایای قابل توجهی دارد. در خوشه‌بندی سخت، هر نمونه به‌صورت قطعی به یک خوشه تعلق می‌گیرد، در حالی که در خوشه‌بندی فازی، هر نمونه می‌تواند با درجات عضویت متفاوت در چند خوشه حضور داشته باشد. این ویژگی امکان تحلیل انعطاف‌پذیرتر، واقع‌بینانه‌تر و تطبیق‌پذیرتر با نوسانات طبیعی رفتار مصرف‌کنندگان را فراهم می‌سازد.

هدف از به‌کارگیری خوشه‌بندی فازی در تحلیل مصرف آب، شناسایی الگوهای مصرف، تعیین گروه‌های رفتاری مشترک و ارتقاء تصمیم‌گیری در زمینه تخصیص، قیمت‌گذاری و برنامه‌ریزی تقاضا است. این امر موجب بهبود کارایی سیستم‌های آبرسانی، کاهش هزینه‌های عملیاتی و افزایش پایداری در مدیریت منابع آب می‌شود. به‌طور خاص، استفاده از رویکردهای فازی در بستر سیستم‌های هوشمند، گامی مؤثر در جهت تحقق شهرهای هوشمند و توسعه پایدار منابع آبی تلقی می‌شود.

Guo و همکاران (۲۰۲۴) با هدف آشکارسازی الگوهای مصرف روزانه آب در شهر شانگهای، سه الگوریتم خوشه‌بندی غیرنظارتی شامل K-means، Agglomerative Hierarchical و Spectral Clustering را روی داده‌های ساعتی ۴۲ روز قبل و بعد از قرنطینه COVID-19 اعمال کردند. آن‌ها با استفاده از شاخص‌های سیلوئت ($SCI=0.6315$) و هاراباز-کالینسکی ($CHI=30.592$) نشان دادند که الگوریتم K-means بهترین تمایز و تراکم خوشه‌ها را ارائه می‌دهد. این مطالعه سه الگوی فصلی اصلی مصرف (بهار-زمستان، زمستان-پاییز و بهار-تابستان) را شناسایی کرد و بر تأثیر روزهای تعطیل و تغییر رفتار مصرف در دوره پاندمی تأکید نمود [۱].

Jalalian و Zolghadr در سال ۲۰۲۲ با معرفی یک معیار فاصله وزنی ترکیبی مکانی-زمانی، الگوریتم Fuzzy C-Means را برای خوشه‌بندی داده‌های حسگر پارکینگ در هنگام اوج پیک ترافیک در منطقه مرکزی تجاری ملبورن توسعه دادند. آن‌ها نشان دادند که با تنظیم جداگانه ضرایب مکانی و زمانی، FCM می‌تواند خطا و همگرایی بهتری نسبت به COBWEB، K-means و DBSCAN ارائه دهد. نتایج، تولید ۹ خوشه مکانی-زمانی متوازن با حداقل انرژی را تأیید کرد و نقش تنظیم λ را در افزایش پایداری خوشه‌ها برجسته ساخت [۲].

Pereira و همکاران (۲۰۱۳) در مقاله خود، مدلی فازی برای تحلیل پاسخ به تقاضا (Demand Response) در شبکه‌های هوشمند برق ارائه کردند که با خوشه‌بندی و استخراج قوانین فازی در MATLAB ترکیب شده بود. آن‌ها سه نوع الگوی مصرف (اقتصادی، آزادمتوسط، آزادپرمصرف) را شناسایی کردند و نشان دادند که کنترل مبتنی بر قوانین فازی، انعطاف‌پذیری و انطباق مصرف با ظرفیت در دسترس را بهبود می‌بخشد. یک مثال از منطقه مسکونی نشان داد که مدل قادر به تضمین حداکثر استفاده اقتصادی از توان مجاز است. این کار پیش‌درآمد مهمی برای کاربرد خوشه‌بندی فازی در مدیریت تقاضای آب شهری فراهم می‌آورد [۳].

Goyal و Gupta (۲۰۱۴) با بهره‌گیری از Fuzzy C-Means و داده‌های بارش ماهانه ایستگاه‌های متعدد در شمال شرق هند (۱۹۵۱-۱۹۹۰)، پنج ناحیه بارش همگن را شناسایی کردند که مرزهای پیچیده اقلیمی را منعکس می‌کند. آن‌ها با استفاده

از معیار Xie-Beni و درجه عضویت نسبی، مشخص نمودند که روش فازی می‌تواند در تعیین دقیق نواحی جغرافیایی با تنوع اقلیمی و توپوگرافی ناهمگن، کارایی بالاتری نسبت به روشهای سخت مانند K-means داشته باشد. نتایج این مطالعه نشان می‌دهد که خوشه‌بندی فازی توان ایجاد تقسیم‌بندی‌های انعطاف‌پذیر و واقع‌گرایانه را در تحلیل داده‌های هیدرولوژیکی داراست [۴].

D'Urso و همکاران (۲۰۲۳) دو نسخه از خوشه‌بندی فازی داده‌های مکانی بازه‌ای را توسعه دادند: یک روش با ماتریس مجاورت مکانی و آنتروپی (EFMd-SpID) و یک روش بدون آنتروپی (FMD-SpID). آنها با ارزیابی سه مورد مطالعاتی در رم، نشان دادند که گنجاندن اطلاعات مجاورت و آنتروپی منجر به خوشه‌هایی همگن‌تر و تفسیرپذیرتر می‌شود. نتایج نشان‌دهنده قدرت روش فازی در کشف ساختارهای مکانی پیچیده و همسوسازی خوشه‌ها با ویژگی‌های واقعی داده است [۵].

Rasmani و Zaidi در سال ۲۰۱۶ پژوهشی را با هدف طبقه‌بندی مصرف‌مادر آب خانگی در مالزی با استفاده از الگوریتم Fuzzy C-Means (FCM) انجام دادند. در این مطالعه، داده‌های رفتاری مربوط به مصرف آب در ۴۳۹ واحد مسکونی شامل ۱۱ متغیر مانند تعداد دفعات استحمام، شستشوی دست، و استفاده از ماشین لباسشویی مورد استفاده قرار گرفت. نتایج نشان داد که الگوریتم FCM قادر به تفکیک مصرف‌کنندگان با احتمال مصرف بیش‌ازحد از سایرین بوده و شباهت ۸۶.۳۳٪ با روش آماری Hotelling's T-squared را از خود نشان داد. این پژوهش با ارائه چارچوبی برای خوشه‌بندی مصرف‌کنندگان بر اساس درجه عضویت فازی، نشان داد که داده‌های رفتاری مصرف‌کننده می‌تواند بدون نیاز به تجهیزات سنجش فیزیکی، پایه‌ای برای توسعه سیستم‌های هشدار یا کنترل هوشمند مصرف آب باشد. در عین حال، عدم وجود داده‌های برچسب‌خورده و تکیه صرف بر داده‌های رفتاری از جمله محدودیت‌های اصلی مطالعه تلقی شده است [۶].

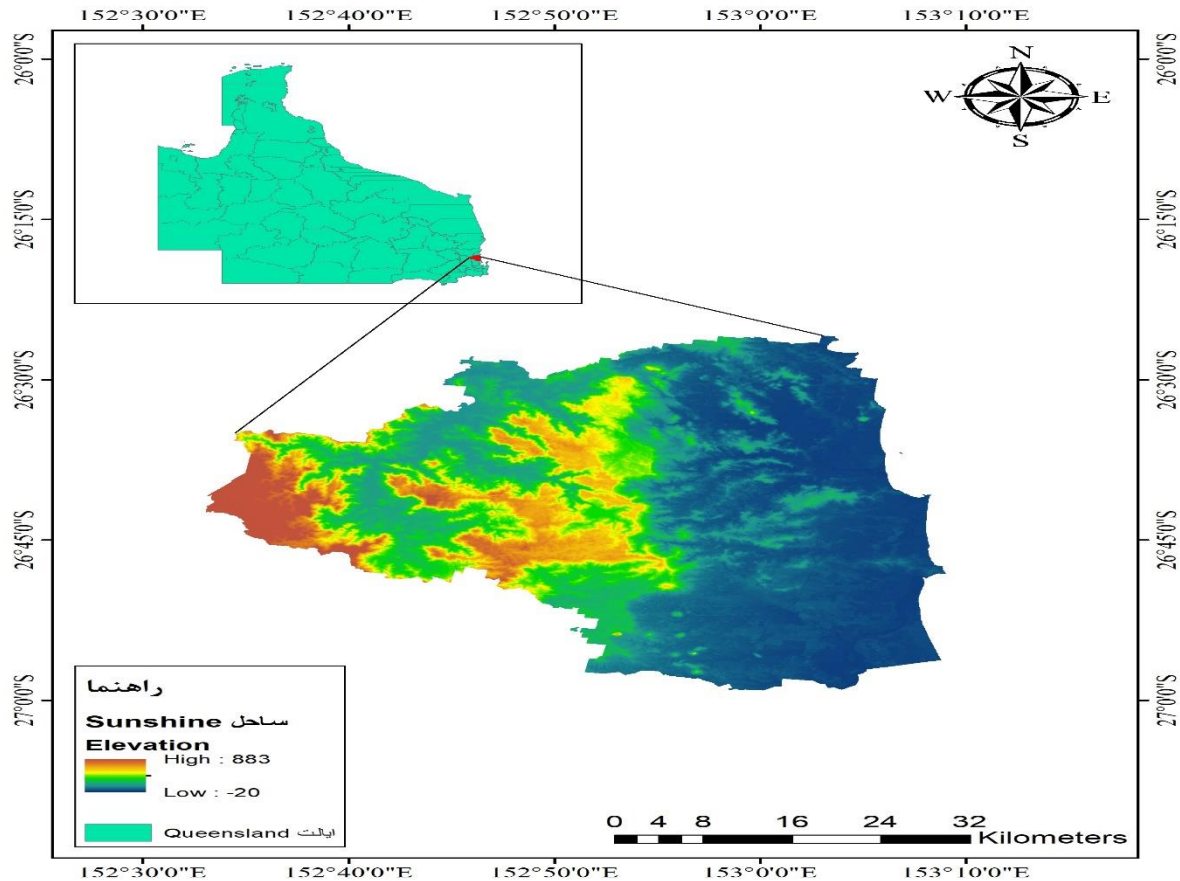
Viegas و همکاران (۲۰۱۵) در مطالعه‌ای نوآورانه، از الگوریتم Fuzzy C-Means برای خوشه‌بندی الگوی مصرف برق خانگی استفاده کردند. در این پژوهش، داده‌های کنتور هوشمند ۱۹۷۲ خانوار در کشور ایرلند با ویژگی‌هایی نظیر تعداد اعضای خانواده، سطح تحصیلات، نوع گرمایش و مالکیت وسایل خانگی پر مصرف به‌عنوان ورودی‌های خوشه‌بندی و مدل‌سازی فازی مورد استفاده قرار گرفت. پس از استخراج الگوهای مصرف ۲۴ ساعته، الگوریتم FCM توانست خانوارها را به دو خوشه «مصرف بالا» و «مصرف پایین» تفکیک کند. مزیت این روش در مقایسه با خوشه‌بندی سخت، توانایی آن در برخورد با داده‌های دارای همپوشانی و پیچیدگی رفتاری بود. همچنین سیستم استنتاج فازی تاگای-سوگنو (TS-FIS) مبتنی بر نتایج خوشه‌بندی، قادر به پیش‌بینی خوشه مصرف با دقت بالایی شد. این مدل‌ها علاوه بر دقت مناسب، امکان تفسیر رفتاری و تصمیم‌گیری مدیریتی برای شرکت‌های خدماتی را فراهم ساختند [۷].

تاکنون مطالعات محدودی (به‌خصوص در ایران) به تحلیل و خوشه‌بندی مصرف آب شهری با بهره‌گیری از رویکرد فازی پرداخته‌اند. در راستای پر کردن این شکاف، در این پژوهش از رویکرد فازی برای خوشه‌بندی مصرف آب شرب شهری در مقیاس زمانی ماهانه استفاده شده است. با توجه به ماهیت پویای الگوهای مصرف آب در مناطق شهری، استفاده از خوشه‌بندی فازی این امکان را فراهم می‌سازد تا رفتار مصرف‌کنندگان به‌صورت انعطاف‌پذیرتر و واقع‌بینانه‌تر تحلیل شود. نتایج این تحلیل می‌تواند به‌عنوان ابزاری کارآمد در هوشمندسازی سیستم آبرسانی شهری، مدیریت هدفمند و پویای مصرف و همچنین کاهش ریسک در تصمیم‌گیری‌های مرتبط با تخصیص منابع آبی مورد استفاده قرار گیرد. از سوی دیگر، این روش می‌تواند زمینه‌ساز شناسایی خوشه‌های پنهان و ناهنجاری‌های مصرف باشد و نقش مهمی در بهینه‌سازی سیاست‌های آبی ایفا کند.

۲- مواد و روش‌ها

۲-۱- منطقه مورد مطالعه

منطقه مورد مطالعه در این پژوهش، ساحل Sunshine واقع در ایالت کوئینزلند در استرالیا می‌باشد. بر اساس آخرین آمار، جمعیت منطقه ساحل Sunshine در ژوئن ۲۰۲۱ حدود ۳۴۲,۵۴۱ نفر بوده است. این منطقه در امتداد ساحل شرقی استرالیا و در نزدیکی اقیانوس آرام قرار دارد (شکل ۱) و به دلیل توسعه سریع شهری، جمعیت روبه‌رشد و ترکیب متنوع کاربری‌های زمین (مناطق مسکونی پرتراکم، مراکز گردشگری، فضاهای تجاری و زمین‌های کشاورزی)، از جمله مناطق مهم در مطالعات مرتبط با مصرف و مدیریت منابع آب به شمار می‌رود. این تنوع کاربری‌ها منجر به الگوهای متفاوت مصرف آب در بخش‌های مختلف منطقه شده و تحلیل این الگوها برای برنامه‌ریزی پایدار و مدیریت بهینه منابع آب اهمیت ویژه‌ای دارد. اقلیم این منطقه نیمه‌گرمسیری مرطوب است و میانگین بارندگی سالیانه این منطقه در حدود ۱,۵۰۰ میلیمتر گزارش شده است. زیرساخت‌های توزیع آب در ساحل Sunshine شامل شبکه‌های توزیع آب شهری، سیستم‌های ذخیره‌سازی و منابع سطحی و زیرزمینی، امکان بررسی دقیق‌تر رفتار مصرف‌کنندگان در مقیاس محلی را فراهم می‌سازد. به دلیل این ویژگی‌ها، ساحل Sunshine نمونه‌ای مناسب برای اجرای روش‌های داده‌کاوی مانند خوشه‌بندی به منظور تحلیل الگوهای مصرف آب، شناسایی رفتارهای مشابه مصرف‌کنندگان، و پشتیبانی از تصمیم‌گیری‌های مدیریتی در زمینه بهینه‌سازی توزیع و استفاده از منابع آب می‌باشد. [۸].



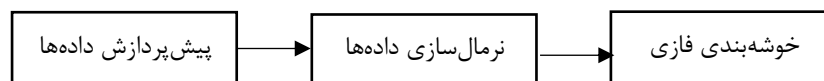
شکل شماره ۱ - نقشه منطقه مطالعاتی

۲-۲- داده‌های مورد استفاده

در این پژوهش از داده‌های واقعی مصرف آب ثبت شده در کنتورهای هوشمند نصب شده در منطقه‌ی ساحل Sunshine استفاده شده است. این داده‌ها به صورت پیوسته و با فواصل زمانی ۱۵ دقیقه‌ای ثبت شده‌اند و بازه‌ی زمانی سه ماهه می (May)، ژوئن (June) و جولای (July) سال ۲۰۲۲ را در بر می‌گیرند. ثبت داده‌ها با این سطح از تفکیک زمانی، امکان تحلیل دقیق نوسانات روزانه و الگوهای رفتاری مصرف‌کنندگان را فراهم می‌سازد. داده‌های خام شامل میزان مصرف لحظه‌ای آب بر حسب لیتر در هر بازه ۱۵ دقیقه‌ای برای هر مشترک می‌باشند. این داده‌ها از کیفیت بسیار بالایی برخوردار بوده و درصد بسیار ناچیزی از آن‌ها دارای مقادیر مفقود یا خطا بوده‌اند که با روش‌های پیش‌پردازش استاندارد اصلاح شده‌اند. استفاده از داده‌های با دقت بالا و پوشش زمانی مناسب، زمینه‌ساز اجرای دقیق الگوریتم‌های خوشه‌بندی و تحلیل رفتاری در این مطالعه شده است. همچنین رزولوشن زمانی این داده‌ها، قابلیت تعمیم‌پذیری نتایج به سایر بازه‌های زمانی را افزایش می‌دهد [۹].

۲-۳- روش انجام پژوهش

در این مطالعه، داده‌های قرائت آب به صورت ۱۵ دقیقه‌ای مربوط به ۷۹۹ کنتور منحصر به فرد از یک منطقه مطالعاتی جمع‌آوری شد. بازه زمانی داده‌ها شامل سه ماه می (May)، ژوئن (June) و ژوئیه (July) بود. این داده‌ها از فایل‌های CSV استخراج و در محیط برنامه‌نویسی پایتون پردازش شدند. در شکل ۲ مراحل کلی انجام پژوهش ارائه شده است.



شکل شماره ۲ - روند کلی انجام پژوهش

۲-۳-۱- پیش‌پردازش داده‌ها

در مرحله نخست، بررسی کامل داده‌های مفقوده (missing value) انجام شد و داده‌های مفقوده با استفاده از میانبایی پر شدند. سپس با هدف حذف نویزها و داده‌های غیرواقعی (noise filtering)، عملیات پاک‌سازی اولیه انجام شد و به تفکیک هر ماه، سه دیتاست مجزا تهیه گردید. پس از این مرحله، تعداد کنتورهای باقی‌مانده در ماه‌های May، June و July به ترتیب ۷۸۷، ۷۹۳ و ۷۷۴ مورد بودند.

تحلیل فراوانی میزان مصرف و شناسایی داده‌های پرت: در گام بعد، برای انتخاب روش مناسب جهت حذف داده‌های پرت، تحلیل فراوانی مصرف کنتورها در هر ماه صورت گرفت. هدف این بود که مشخص شود چند درصد از کنتورها از توزیع نرمال پیروی می‌کنند و آیا توزیع اکثریت آن‌ها متقارن است یا خیر. نتایج نشان داد که تنها ۲۰٫۶٪ از کنتورها در ماه May، ۲۰٫۹٪ در ماه June و ۲۱٫۲٪ در ماه July از توزیع نرمال پیروی می‌کنند و در هر سه ماه، توزیع اکثریت کنتورها نامتقارن بود. با توجه به این نتایج، روش‌های آماری مبتنی بر فرض نرمال بودن داده‌ها قابل استفاده نبوده و برای شناسایی داده‌های پرت، روش نمودار جعبه‌ای (Box Plot) مورد استفاده قرار گرفت.

حذف داده‌های پرت با استفاده از نمودار جعبه‌ای: روش نمودار جعبه‌ای بر پایه چارک‌ها (quartiles) عمل می‌کند. در این روش، برای هر مجموعه داده چارک اول (Q_1) و چارک سوم (Q_3) محاسبه می‌شود و سپس دامنه بین چارک‌ها (IQR) به صورت رابطه ۱ تعیین می‌گردد.

$$Q_1 - Q_3 = IQR \quad (1)$$

هر داده‌ای که خارج از بازه زیر باشد، پرت تلقی می‌شود:

$$[Q_1 - 1.5 IQR, Q_3 + 1.5 IQR] \quad (2)$$

در روابط ۱ و ۲ Q_1 چارک اول و Q_3 چارک سوم و IQR دامنه بین چارکی می‌باشد. این روش به دلیل استقلال از فرض نرمال بودن و مقاومت در برابر داده‌های نامتقارن، برای حذف داده‌های پرت در این مطالعه مناسب تشخیص داده شد. در جدول شماره ۱ درصد داده‌های حذف شده در هر ماه نشان داده شده است.

جدول شماره ۱: درصد داده‌های پرت در هر ماه

ردیف	ماه	تعداد کنتورها بررسی شده	متوسط درصد داده‌های پرت شناسایی شده (درصد)
۱	may	۷۸۷	۱۰.۰
۲	june	۷۹۳	۱۰.۱
۳	july	۷۷۴	۱۰.۲

تغییر مقیاس زمانی داده‌ها: به منظور آماده‌سازی داده‌ها برای خوشه‌بندی و با توجه به هدف این پژوهش (خوشه‌بندی مصرف آب در مقیاس ماهانه) داده‌ها به مقیاس ماهانه (monthly) تغییر مقیاس داده شدند تا مصرف کل هر کنتور در هر ماه محاسبه شود. پس از تغییر مقیاس زمانی داده‌ها به ماهانه، تعداد کنتورهای باقی‌مانده با داده ماهانه در هر ماه در جدول شماره ۲ ارائه شده است:

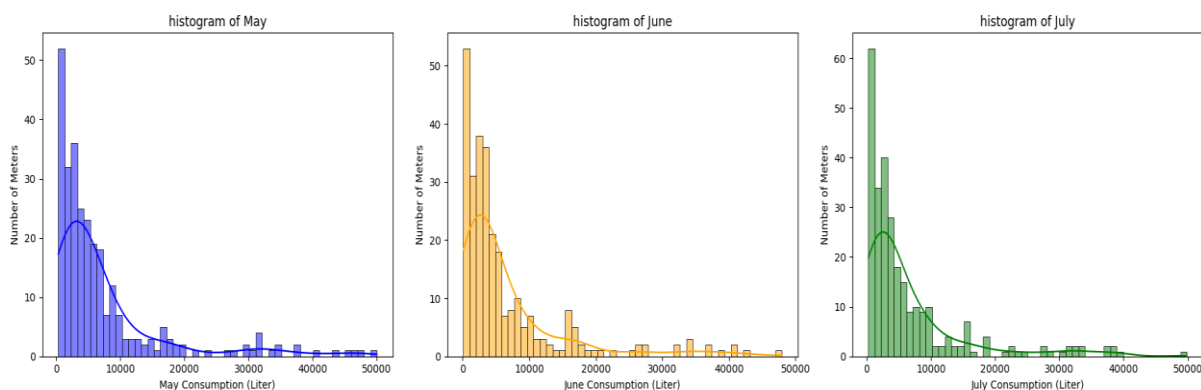
جدول شماره ۲: تعداد کنتورها در مقیاس ماهانه

ردیف	ماه	تعداد کنتور
۱	May	۴۶۷
۲	June	۴۵۱
۳	July	۴۴۴

با توجه به هدف مطالعه، که مقایسه خوشه‌بندی مصرف در هر سه ماه است، تنها کنتورهایی انتخاب شدند که در هر سه ماه داده کامل داشتند؛ در نتیجه ۴۰۱ کنتور مشترک برای تحلیل نهایی مورداستفاده قرار گرفتند. حذف کنتورهای پرت (ناهنجار): با بررسی دقیق‌تر این ۴۰۱ کنتور، مشخص شد که برخی از کنتورها در مقیاس ماهانه دارای مصارف غیرعادی بسیار بالا هستند که احتمالاً مربوط به کاربری‌های خاص یا اشتباهات سیستمی بوده‌اند. این موارد می‌توانند در فرآیند خوشه‌بندی اثر منفی بگذارند. برای رفع این مسئله، با بررسی آمارهای موجود از متوسط مصرف در منطقه موردنظر به ازای هر خانوار، کنتورهایی که مصرف بیش از ۵۰,۰۰۰ لیتر در ماه داشتند حذف شدند. در نهایت، ۲۸۰ کنتور پاک‌سازی شده برای اجرای تحلیل فازی انتخاب شدند. مشخصات آماری ۲۸۰ کنتور باقی‌مانده به تفکیک ماه در جدول شماره ۳ آمده است و همچنین هیستوگرام فراوانی مصرف آن‌ها در شکل شماره ۳ نشان داده شده است.

جدول شماره 3: ویژگی‌های آمار توصیفی کنتورها به تفکیک ماه پس از حذف کنتورهای پرت

ردیف	عنوان	واحد	may	june	july
۱	تعداد کنتورها	تعداد	۲۸۰	۲۸۰	۲۸۰
۲	میانگین	لیتر	۷۲۶۹.۲۷	۶۵۳۲.۰۷	۶۴۶۲.۴۴
۳	انحراف از معیار	لیتر	۹۳۵۴.۷۸	۸۵۵۷.۳	۸۴۳۶.۰۳
۴	مینیمم	لیتر	۳۲۳	۱۰۷	۲۶۰.۵
۵	ماکسیمم	لیتر	۴۹۹۸۲.۵	۴۷۹۴۰	۴۹۸۴۵



شکل شماره ۳ - هیستوگرام مصرف ماهانه کنتورها به تفکیک ماه پس از حذف کنتورهای پرت

۲-۳-۲-۲- نرمال‌سازی داده‌ها

نرمال‌سازی داده‌ها یکی از مراحل کلیدی پیش‌پردازش در تحلیل داده‌ها و یادگیری ماشین محسوب می‌شود. این فرآیند با هدف یکسان‌سازی مقیاس ویژگی‌ها، استفاده می‌شود تا تفاوت در واحد اندازه‌گیری یا دامنه عددی ویژگی‌ها در نتایج تحلیل دخالت نداشته باشد. در مسائل مربوط به خوشه‌بندی، به‌ویژه خوشه‌بندی فازی، نرمال‌سازی یک ضرورت است. الگوریتم‌هایی مانند Fuzzy C-Means مبتنی بر محاسبه فاصله بین نمونه‌ها هستند و اگر مقیاس ویژگی‌ها متفاوت باشد، ویژگی‌هایی با دامنه بزرگ‌تر، تأثیر بیشتری بر محاسبه فاصله و در نتیجه بر نتایج خوشه‌بندی خواهند گذاشت.

در این مطالعه، از رایج‌ترین روش نرمال‌سازی، یعنی Min-Max Scaling استفاده شده است. این روش، هر مقدار را نسبت به حداقل و حداکثر موجود در همان ویژگی، به بازه $[0, 1]$ تبدیل می‌کند. فرمول این روش به صورت رابطه (۳) تعریف می‌شود:

$$X_{Normal} = \frac{X - X_{Max}}{X_{Max} - X_{Min}} \quad (3)$$

که در این پژوهش X_{Normal} مصرف آب ماهانه نرمال شده؛ X_{Max} حداکثر مصرف آب ماهانه؛ X_{Min} حداقل مصرف آب ماهانه و X مقدار مصرف آب ماهانه می‌باشد.

۲-۳-۳- خوشه‌بندی فازی C- میانگین (Fuzzy C-Means Clustering)

الگوریتم خوشه‌بندی فازی C- میانگین (Fuzzy C-Means یا FCM) یکی از روش‌های معروف خوشه‌بندی است که در آن، برخلاف خوشه‌بندی سخت (مانند K-Means)، هر نمونه می‌تواند به بیش از یک خوشه تعلق داشته باشد و این تعلق به صورت یک مقدار عضویت فازی بین ۰ و ۱ بیان می‌شود. این ویژگی به الگوریتم FCM توانایی مدل‌سازی دقیق‌تر و واقع‌گرایانه‌تری از داده‌ها می‌دهد، به ویژه در مسائلی که مرز بین خوشه‌ها واضح نیست، مانند داده‌های مصرف آب، الگوهای رفتاری یا سیستم‌های پیچیده.

هدف اصلی الگوریتم FCM، یافتن موقعیت بهینه مراکز خوشه‌ها و درجه عضویت هر داده به این مراکز است، به گونه‌ای که تابع هدف الگوریتم حداقل مقدار ممکن را داشته باشد. در این الگوریتم، داده‌ها بر اساس میزان شباهتشان به مراکز خوشه‌ها طبقه‌بندی می‌شوند و خروجی نهایی شامل موقعیت مراکز خوشه‌ها و ماتریس عضویت فازی داده‌ها به خوشه‌ها می‌باشد. مراحل خوشه‌بندی فازی C- میانگین:

(۱) انتخاب تعداد خوشه‌ها (C): انتخاب تعداد خوشه‌ها یکی از تصمیمات کلیدی پیش از اجرای الگوریتم است. این مقدار باید از پیش تعیین شود و معمولاً با استفاده از روش‌های زیر تعیین می‌گردد: الف) دانش مسئله (مثلاً تقسیم‌بندی مصرف‌کنندگان به سه دسته کم‌مصرف، متوسط و پرمصرف؛ ب) شاخص‌های ارزیابی درونی نظیر: Fuzzy Partition Coefficient (FPC), Xie-Beni index و Silhouette Score؛

در این مطالعه، با توجه به هدف پژوهش و تجربه، تعداد خوشه‌ها برابر ۳ انتخاب شد تا امکان تحلیل الگوهای مصرف در سه گروه اصلی (کم‌مصرف، مصرف متوسط و پرمصرف) فراهم گردد.

(۲) مقداردهی اولیه ماتریس عضویت (U): ماتریس اولیه به صورت تصادفی مقداردهی می‌شود.

$$U = [u_{ij}] \quad (4)$$

به صورتی که:

$$\sum_j u_{ij} = 1 \quad \text{برای } i = 1, 2, 3 \text{ و } \dots \text{ و } N = 401 \quad (5)$$

که در آن u_{ij} نشان دهنده میزان تعلق داده i به خوشه j ام و N تعداد کل داده‌ها (کنتورها) می‌باشد.

(۳) تعیین تابع هدف: این تابع هدف به منظور حداقل کردن فاصله بین هر نقطه در خوشه تا مرکز آن خوشه و حداکثر کردن فاصله بین مراکز دسته‌ها تعریف می‌گردد.

$$J(M) = \sum_{i=1}^N \sum_{j=1}^C u_{ij} d_{ij}^2 \quad (6)$$

که در آن $J(M)$ تابع هدف و u_{ij} مقدار عضویت داده i در خوشه j و d_{ij} فاصله داده i از مرکز خوشه j می‌باشد.

(۴) محاسبه مراکز اولیه خوشه‌ها: با استفاده از ماتریس عضویت اولیه، مراکز اولیه خوشه‌ها v_{ij} با رابطه زیر محاسبه می‌شود:

$$v_{jk} = \frac{\sum_{i=1}^N u_{ij}^m x_{ik}}{\sum_{i=1}^N u_{ij}^m} \quad (7)$$

که در آن v_{kj} مرکز k امین ویژگی تاثیرگذار در خوشه‌بندی در j امین خوشه می‌باشد، x_{ik} داده نرمال شده i ، u_{ij} درجه عضویت داده i در گروه j و m ضریب فازی بودن می‌باشد. ضریب فازی بودن بین ۱.۲۵ تا ۲ قرار دارد که در این پژوهش برای افزایش نقش فازی بودن در خوشه‌بندی برابر با ۲ در نظر گرفته می‌شود.

(۵) بروزرسانی ماتریس عضویت: با داشتن مراکز جدید خوشه‌ها، درجه عضویت هر داده به هر خوشه طبق رابطه ۸ به‌روزرسانی می‌شود:

$$u_{ij} = \left[\sum_{b=1}^c \left(\frac{d_{ij}}{d_{ib}} \right)^{\frac{2}{m-1}} \right] \quad (۸)$$

که در آن u_{ij} درجه عضویت داده i به خوشه j ام در تکرار جدید، d_{ij} فاصله داده i ام از مرکز خوشه j ام، و d_{ib} فاصله داده i ام از مرکز خوشه b ام و m ضریب فازی بودن می‌باشد. برای محاسبه فاصله از فاصله اقلیدسی استفاده شده است:

$$d_{ij} = \sqrt{\sum_{k=1}^Z (x_{ik} - v_{jk})^2} \quad (۹)$$

که در آن d_{ij} فاصله داده i از مرکز خوشه j و x_{ik} نشان‌دهنده k امین ویژگی تاثیرگذار در خوشه‌بندی از داده i ام v_{jk} و مقدار ویژگی k ام در مرکز دسته j می‌باشد. باتوجه به اینکه در این پژوهش خوشه‌بندی تنها بر اساس مصرف ماهانه صورت گرفته $Z = 1$ می‌باشد.

با توجه به فرآیند تکراری بودن محاسبات قید شده در بالا، الگوریتم تا زمانی اجرا می‌شود که یکی از شرایط زیر برقرار شود: الف) تغییرات ماتریس عضویت یا مراکز خوشه‌ها کمتر از یک مقدار آستانه (در این پژوهش $\epsilon = 5 \times 10^{-4}$) باشد؛ ب) تعداد تکرارها تا یک مقدار مشخص (در این پژوهش ۱۰۰۰ تکرار) انجام شود. خلاصه مقادیر انتخاب شده برای روش خوشه‌بندی فازی C-میانگین در جدول شماره ۴ ارائه شده است.

جدول شماره ۴: مقادیر پارامترهای مورد استفاده در روش FCM

ردیف	پارامتر	شرح	مقدار
۱	c	تعداد دسته‌ها	۳
۲	m	ضریب فازی بودن	۲
۳	N	تعداد کنتورها	۲۸۵
۴	Z	تعداد ویژگی‌های مورد بررسی در هر کنتور	۱
۵	ϵ	آستانه تغییرات ماتریس عضویت در هر تکرار	5×10^{-4}
۶	epoch	حداکثر تعداد بروزرسانی ماتریس عضویت	۱۰۰۰

۳- نتایج پژوهش

به‌منظور ارزیابی نتایج حاصل از خوشه‌بندی فازی در هر ماه، از ترکیبی از ابزارهای عددی و تصویری استفاده شده است. این ابزارها شامل جدول فراوانی کنتورها در هر خوشه، هیت مپ عضویت، نمودارهای برش آلفا و همچنین جدول تخصیص نهایی داده‌ها به خوشه‌ها بر اساس بیشترین درجه عضویت می‌باشند. در ادامه، نتایج برای ماه‌های مه، ژوئن و ژوئیه به‌صورت جداگانه ارائه و تحلیل شده‌اند.

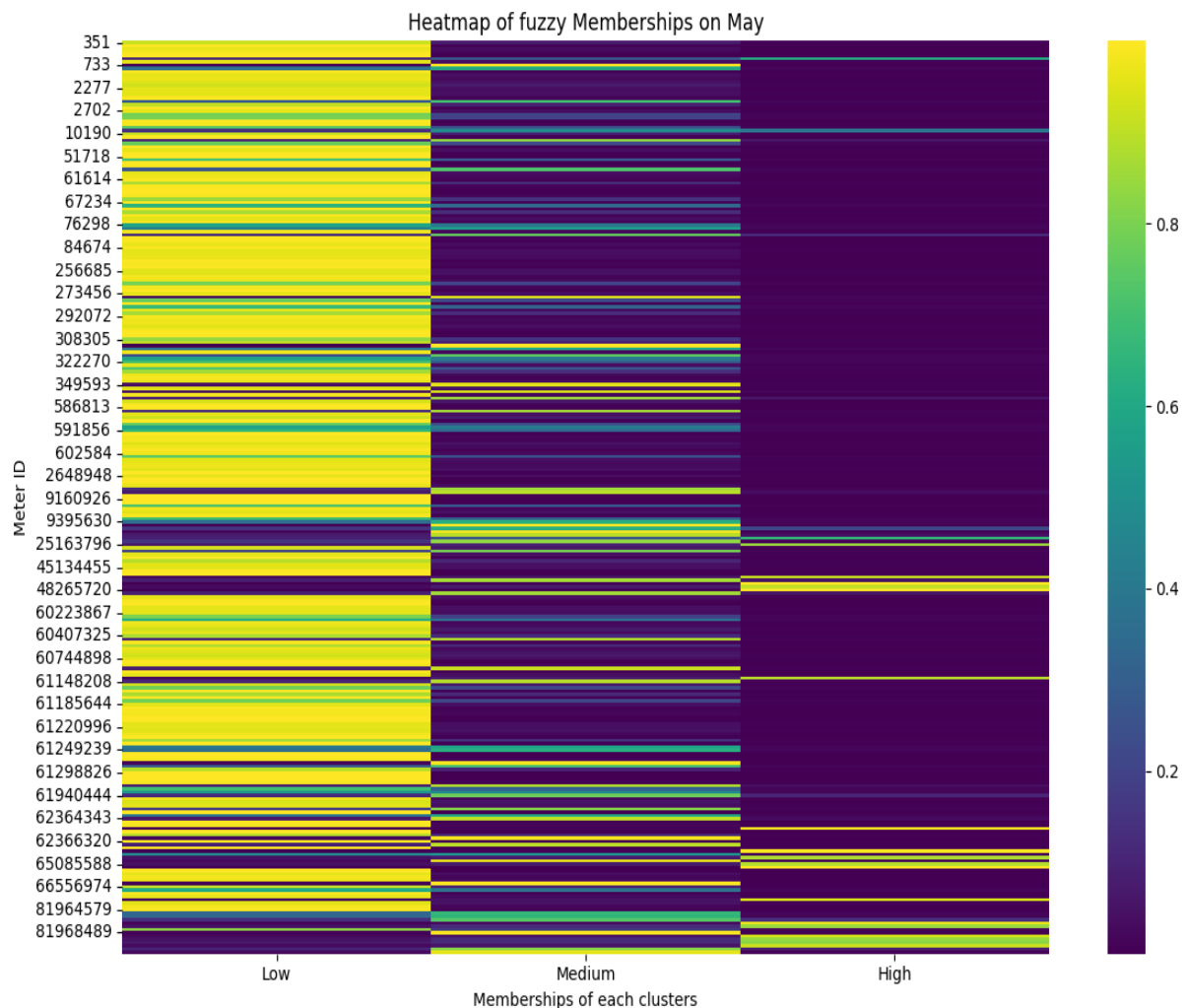
پس از اجرای الگوریتم خوشه‌بندی فازی C-میانگین با تعداد خوشه‌های مشخص (۳ دسته)، با محاسبه تعداد کنتورهای اختصاص یافته به هر خوشه بر اساس بیشترین درجه عضویت، جدول شماره ۵ به‌دست می‌آید. این اطلاعات، توزیع کلی الگوهای مصرف در هر ماه را نمایش می‌دهد.

جدول شماره ۵: فراوانی تعداد کنتورها در هر خوشه

ردیف	عنوان دسته	may	june	july
۱	کم مصرف	۲۱۰	۲۱۳	۲۰۲
۲	مصرف متوسط	۵۰	۵۰	۵۹
۳	پرمصرف	۲۰	۱۷	۱۹

1-3- نمودار درجه عضویت

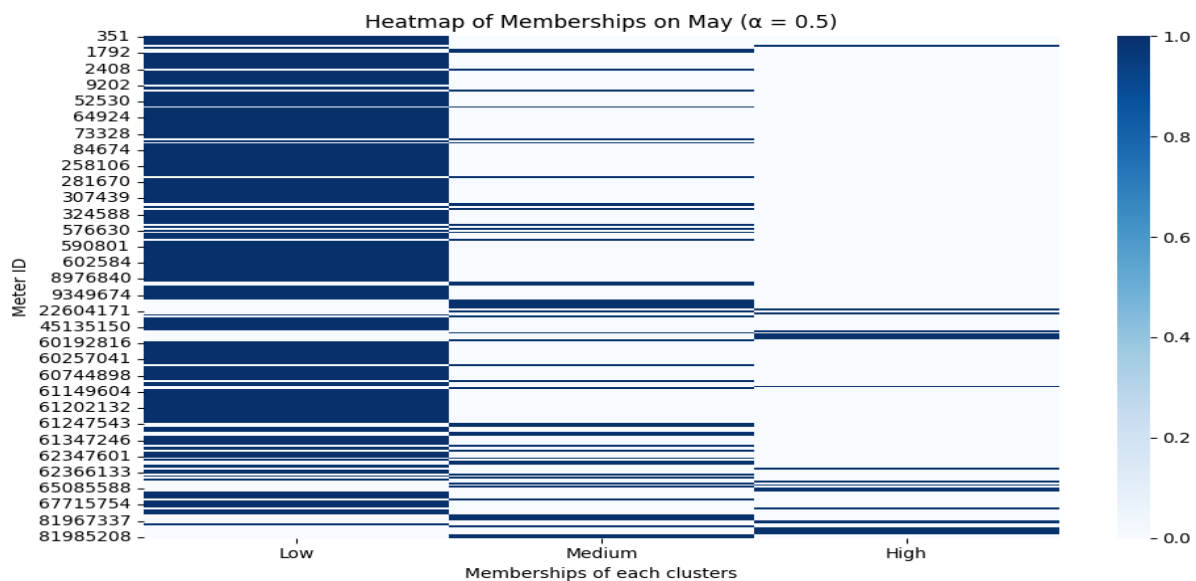
این نمودار درجه عضویت هر کنتور به هر دسته را نشان می‌دهد. با بررسی این نمودار مشخص می‌شود که برخی کنتورها با اطمینان بالا به یک خوشه خاص تعلق دارند، درحالی‌که برخی دیگر بین دو یا چند خوشه دارای تعلق نسبی هستند که بیانگر رفتارهای مصرفی نامشخص‌تر آن‌هاست. این نمودار می‌تواند در مدیریت مصارف و سیاست‌گذاری‌های مربوطه بسیار کمک‌کننده باشد و به‌عنوان ابزاری برای تصمیم‌گیری مدیران در اختیار ایشان قرار گیرد. در شکل شماره ۴ نمودار درجه عضویت مربوط به ماه may به‌عنوان نمونه نشان داده شده‌است.



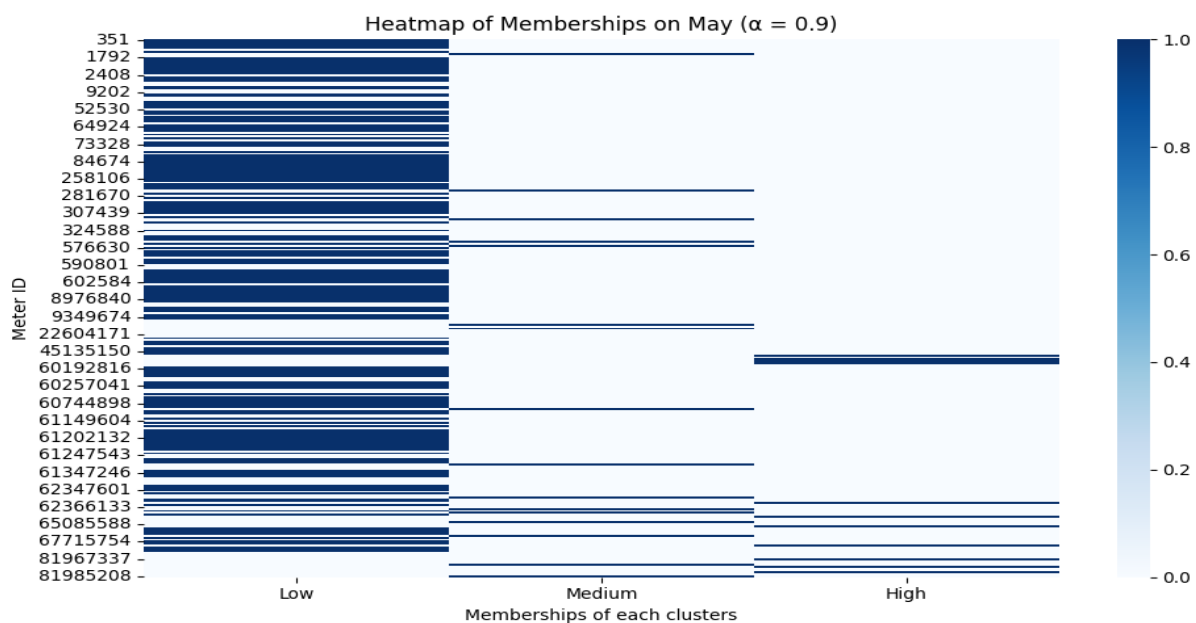
شکل شماره ۴- نمودار درجه عضویت کنتورها در هر دسته

۳-۲- نمودار برش آلفا (α -cut)

برای بررسی حدود سخت تر هر خوشه، از مفهوم برش آلفا استفاده شد. در این روش، تنها داده‌هایی که درجه عضویت آن‌ها به یک خوشه خاص بیش از مقدار آلفا باشد، به عنوان عضو آن خوشه در نظر گرفته می‌شوند. این دسته نمودارها به عنوان راهی برای غیرفازی سازی نتایج استفاده می‌شود و به درک مرزهای هر خوشه کمک می‌کند. به منظور تحلیل بهتر نتایج و بررسی تغییرات نتایج دسته‌بندی فازی، مقادیر آلفا برابر با ۰.۵، ۰.۷ و ۰.۹ در نظر گرفته شده‌اند؛ در شکل‌های شماره ۵ و ۶ نمودارهای برش آلفا ($\alpha = 0.5$ و $\alpha = 0.9$) مربوط به ماه may به عنوان نمونه نمایش داده شده است.



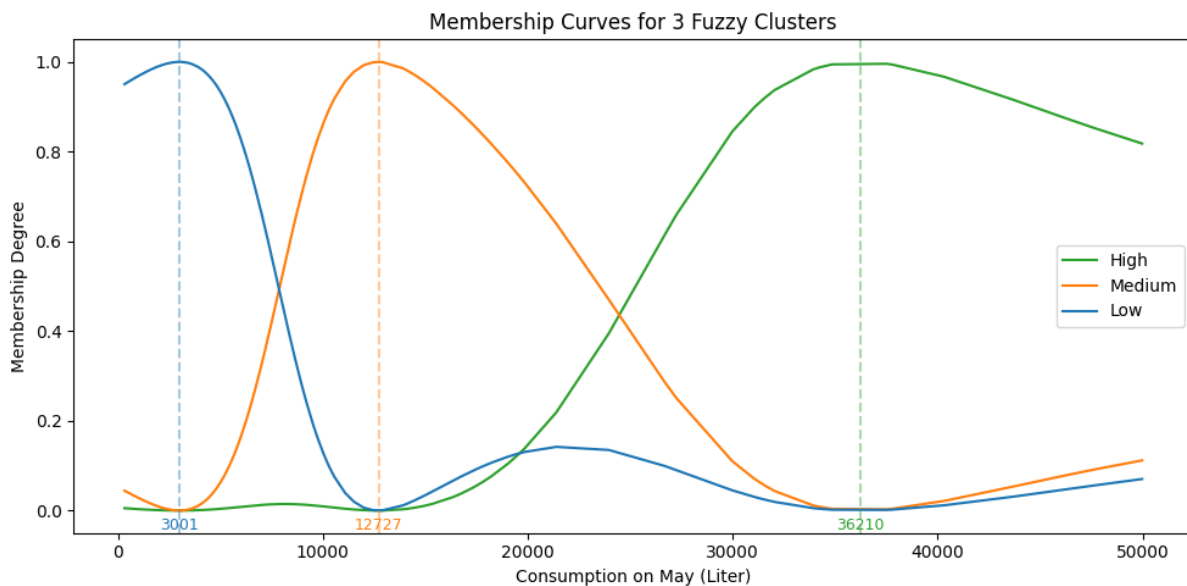
شکل شماره ۵ - نمودار برش آلفا برای $\alpha = 0.5$



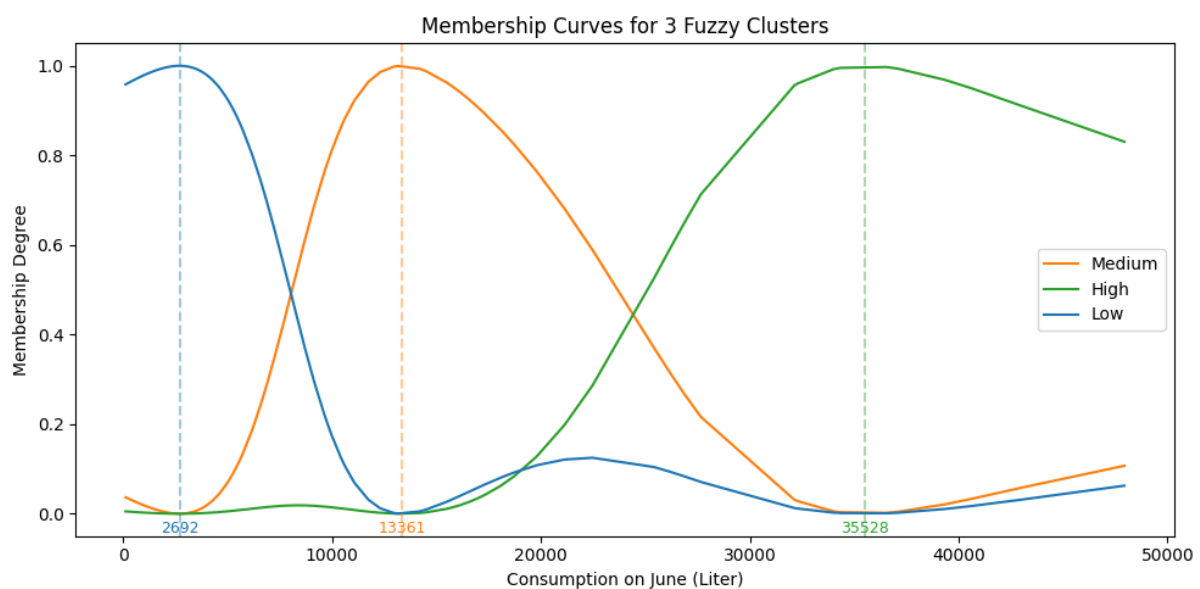
شکل شماره ۶ - نمودار برش آلفا برای $\alpha = 0.9$

۳-۳- منحنی عضویت

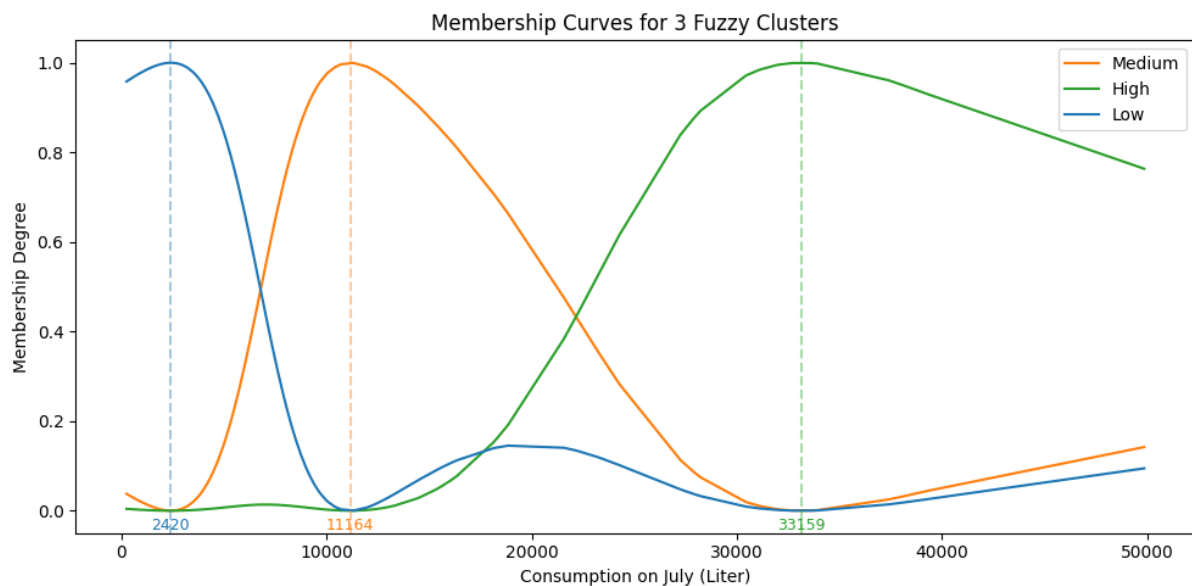
این دسته منحنی‌ها اطلاعات ارزش‌مندی را دارا می‌باشند؛ در این نمودار درجه عضویت مربوط به مصارف مختلف در هر دسته نشان داده شده و همچنین نقاطی که درجه عضویت آنها برابر با ۱ می‌باشد نشان دهنده مراکز دسته‌ها می‌باشند. در شکل‌های شماره ۷ تا ۹ منحنی عضویت سه ماه may و june و july نشان داده شده‌است.



شکل شماره ۷- منحنی عضویت مصارف در ماه may

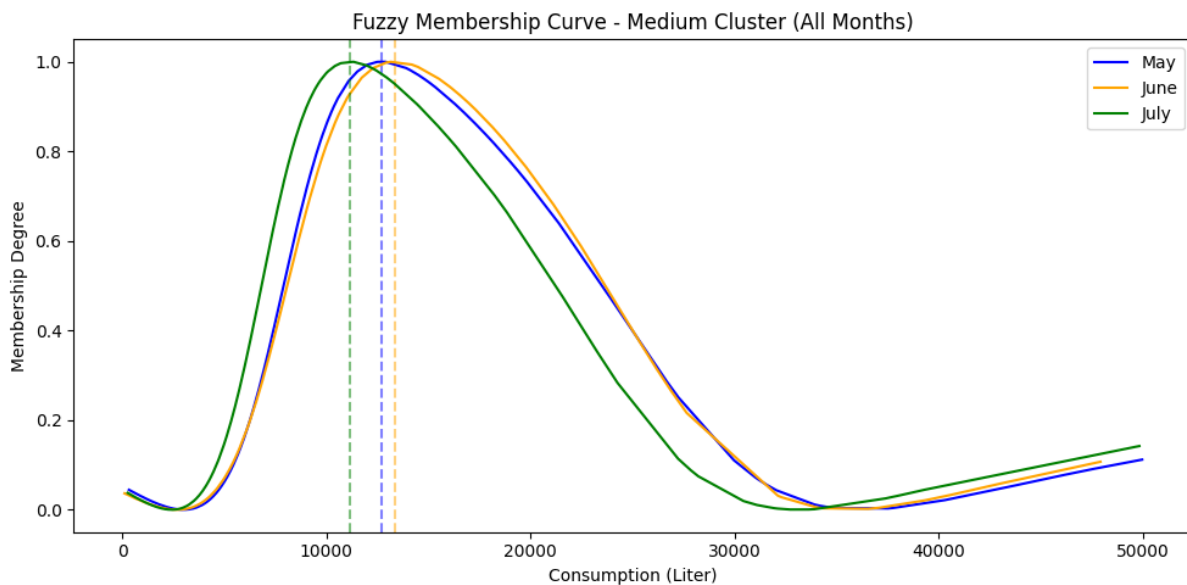


شکل شماره ۸- منحنی عضویت مصارف در ماه june



شکل شماره ۹ - منحنی عضویت مصارف در ماه July

در شکل شماره 10 منحنی عضویت دسته متوسط در هر سه ماه نشان داده شده است.



شکل شماره ۱۰ - منحنی عضویت مصارف در ۳ ماه

باتوجه به شکل ۱۰ می توان شباهت زیاد دسته ها در ماه های مختلف را متوجه شد.

۴- نتیجه گیری

باتوجه به نتایج این پژوهش، می توان دریافت بیشتر کنتورها در خوشه کم مصرف قرار گرفته اند که این امر نشان دهنده غالب بودن الگوی مصرف پایین در میان مشترکان منطقه مطالعه است (در بازه زمانی مورد مطالعه). همچنین با بررسی دسته های بدست آمده، می توان نتیجه گرفت که در هر ۳ ماه، تعداد کنتورهای متعلق به هر دسته تغییر معناداری نداشته و حول یک عدد مشخص نوسانی اندک داشته اند، که این موضوع می تواند نشان دهنده شباهت میان الگو مصرف در هر ۳ ماه باشد.

با توجه به نمودارهای درجه عضویت در هر سه ماه مشخص شد الگوی مصرف کنتورها تغییر زیادی نداشته و اکثر کنتورها در هر سه ماه به یک دسته خاص مصرفی تعلق داشته اند، که این ثبات می تواند حاکی از پایدار بودن رفتار مصرفی در بسیاری از خانوارها یا وجود عادت های تثبیت شده مصرف باشد. در عین حال، درجه عضویت متغیر برخی کنتورها به بیش از یک خوشه، بیانگر وجود عدم قطعیت در رفتار مصرفی آن ها می باشد و این دسته از کنتورها می توانند اهداف کلیدی در اجرای سیاست های مداخله ای، پاداش دهی یا هشداردهی باشند.

نمودارهای برش آلفا نشان می دهند با افزایش مقدار آلفا، خوشه بندی سخت تر شده و تمایز میان کنتورهای متعلق به هر دسته بیشتر می شود، و در این حالت تعداد کنتورهای متعلق به هر دسته کاهش می یابد.

این پژوهش نشان داد الگوریتم فازی-C میانگین قادر است با وجود تنها یک ویژگی ورودی (مصرف ماهانه)، ساختار رفتاری معناداری در بین کنتورها کشف کند. این قابلیت، امکان توسعه آن را با ترکیب ویژگی های بیشتری مانند ویژگی های مکانی، جمعیتی، اقلیمی یا اجتماعی در آینده فراهم می سازد؛ پیشنهاد می شود در ادامه مطالعات بیشتری بر روی آن صورت گیرد و تاثیر ترکیب این خصوصیات با مصرف کنتور بر خوشه بندی بررسی گردد.

۵- مراجع

1. Guo H., Liu X., Zhang Q. 2024. Identification of daily water consumption patterns using clustering algorithms. AQUA — Water Infrastructure, Ecosystems and Society, 73(5):870. <https://doi.org/10.2166/aqua.2024.294>
2. Zolghadr A, & Jalalian A. 2022. A fuzzy clustering based method for the spatio-temporal data analysis, Journal of Computer Science, 18(6):480–495. <https://doi.org/10.3844/jcssp.2022.480.495>.
3. Pereira R., Fagundes A., Melício R., Mendes V. F., Figueiredo J., et al. 2013. Demand Response Analysis in Smart Grids Using Fuzzy Clustering Model. In 4th Doctoral Conference on Computing, Electrical and Industrial Systems (DoCEIS), Costa de Caparica, Portugal, pp.403–412. https://doi.org/10.1007/978-3-642-37291-9_43
4. Goyal, M.K., Gupta, V. Identification of Homogeneous Rainfall Regimes in Northeast Region of India using Fuzzy Cluster Analysis. Water Resour Manage 28, 4491–4511 (2014). <https://doi.org/10.1007/s11269-014-0699-7>
5. D’Urso P., De Giovanni L, Federico L, & Vitale V. 2023. Fuzzy clustering of spatial interval-valued data, Spatial Statistics, 57:1–42. <https://doi.org/10.1016/j.spasta.2023.100764>
6. Zaidi, A. Z., & Rasmani, K. A. 2016. Classification of excessive domestic water consumption using fuzzy clustering method. 5th International Conference on Mathematical Modeling in Physical Sciences (IC-MSquare 2016), Seremban, Malaysia. <https://doi.org/10.1088/1742-6596/738/1/012081>

The First National Conference on Hydroinformatics and Artificial Intelligence in Water Engineering



7. Viegas J L, Vieira S M, & Sousa J M C. 2015. Fuzzy clustering and prediction of electricity demand based on household characteristics. 16th World Congress of the International Fuzzy Systems Association (IFSA) and 9th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), Gijón, Spain.
<https://doi.org/10.2991/ifs-eusflat-15.2015.147>
8. Bureau of Meteorology. (2025, May 13). Climate statistics for Australian locations: Sunshine Coast Airport (040861). Retrieved June 23, 2025, from http://www.bom.gov.au/climate/averages/tables/cw_040861.shtml
9. Queensland Government. (2022, August 20). Unitywater – Digital Water Meter data. Retrieved June 23, 2025, from <https://www.data.qld.gov.au/dataset/digital-water-meter-dataset-explanation>