نشریه علمی مهندسی فناوری اطلاعات مکانی

سال نهم، شماره دوم، تابستان ۱۴۰۰

Vol.9, No.2, Summer 2021

۱۰۵ -۱۲۸

مقاله پژوهشی

DOR: 20.1001.1.20089635.1400.9.2.1.0

توسعه یک الگوریتم خوشه‌بندی مبتنی بر تراکم مکانی و زمانی برای استخراج مکان‌های توقف از خط سیر کاربر

نگین مسن آبادی[۱]، فرهاد حسینعلی[۲*]، زهرا بهرامیان

۱- دانشجوی کارشناسی ارشد سیستم اطلاعات مکانی، دانشگاه تربیت دبیر شهید رجایی، تهران، ایران

۲- استادیار گروه مهندسی نقشه‌برداری، دانشگاه تربیت دبیر شهید رجایی، تهران، ایران

۳- دکترای مهندسی سیستم‌های اطلاعات مکانی، دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی، پردیس دانشکده‌های فنی، دانشگاه تهران، تهران، ایران

تاریخ دریافت مقاله: ۱۳۹۹/۰۹/۲۱ تاریخ پذیرش مقاله: ۱۴۰۰/۰۷/۲۴

چکیده

شناسایی مکان‌های توقف در خطوط سیر یک گام اولیه و ضروری در مطالعه اشیاء در حال حرکت است و تأثیر عمده‌ای در برنامه‌ها و خدمات مکانی دارد. برای استخراج نقاط توقف در این پژوهش از خوشه‌بندی خط سیر استفاده می‌شود. الگوریتم خوشه‌بندی مکانی مبتنی بر تراکم برنامه‌های کاربردی با نوفه (*DBSCAN*)، الگوریتم پایه روش‌های خوشه‌بندی مبتنی بر چگالی است که با وجود دارا بودن مزایایی، دارای مشکلاتی نظیر سخت بودن تعیین پارامترهای ورودی، عدم توانایی کشف خوشه‌های با چگالی متفاوت و عدم توجه به مشکل رفت و برگشت است. در روش پیشنهادی این تحقیق که مبتنی بر چگالی است با استفاده از شاخص‌های مکانی و زمانی و استفاده از چندین شعاع همسایگی، به استخراج نقاط توقف پرداخته می‌شود. حل مشکل رفت و برگشت، استخراج خوشه‌ها با چگالی متفاوت و کاهش میزان وابستگی نتایج به پارامترهای ورودی از مزایای روش پیشنهادی است. به منظور ارزیابی الگوریتم، این روش بر روی داده‌های خط سیر تولید شده در شهر اراک و نیز داده‌های مربوط به پروژه پژوهش ژئولایف پیاده‌سازی شد. نتایج اخذ شده با نتایج حاصل از پنج الگوریتم دیگر شامل *ST-DBSCANT*، *DVBSCAN*، *DBSCAN*، *VDBSCAN* و *K* میانگین، مورد مقایسه قرار گرفت. در مقایسه روی داده‌های خط سیر شهر اراک، مکان‌های توقف استخراج شده توسط الگوریتم پیشنهادی و الگوریتم‌های ذکر شده به ترتیب ۱۰۰٪ ، ۲۵٪ ، ۷۵٪ ، ۵۰٪ ، ۷۵٪ و ۵۰٪ به درستی استخراج شده‌اند که حاکی از برتری روش توسعه داده شده است. همچنین پس از استخراج نقاط توقف و حرکت، شاخص‌هایی از داده‌های *Geolife* برای شناسایی روز کاری و غیر کاری (تعطیل) تعیین گردید که با این شاخص‌ها، روش پیشنهادی تا ۹۴/۰۶٪ موفق عمل کرد. نتایج بیانگر کاهش میزان وابستگی نتایج به پارامترهای ورودی، استخراج نقاط توقف به طور صحیح، کاهش میزان انحراف معیار درون خوشه‌ها و افزایش فاصله مراکز خوشه‌ها می‌باشد.

کلید واژه‌ها : خط سیر، استخراج مکان‌های توقف، خوشه‌بندی مکانی-زمانی، *DBSCAN* .

* نویسنده مکاتبه کننده: تهران، لویزان، خیابان شعبانلو، دانشگاه تربیت دبیر شهید رجایی، دانشکده مهندسی عمران.

تلفن: ۲۲۹۷۰۰۲۱

Email: f.hosseinali@sru.ac.ir

۱- مقدمه

گذشت زمان و افزایش روزافزون و همه‌گیر شدن استفاده از دستگاه‌های متحرک و فناوری‌های تعیین موقعیت، منجر به تولید حجم انبوه و بی‌شماری از داده‌های خط سیر مکانی-زمانی و در پی آن ایجاد مشکلاتی در ذخیره‌سازی، انتقال، پردازش و پرس‌وجو از این داده‌ها شده است. خط سیر[1]، نمایش مسیر حرکت جسم متحرک است که معمولاً به صورت رشته‌ای از موقعیت‌های زمان‌دار تعریف می‌شود که به صورت موقعیت دوبعدی نقطه به همراه زمان ثبت مربوط به آن نقطه است[۱].

مسیرهای طی شده توسط کاربران حاوی اطلاعات رفتاری و علایق آن‌ها می‌باشد (مورد اول). این اطلاعات در زمینه‌های مختلف مفید است و مورد استفاده قرار می‌گیرد. به عنوان مثال، مشخص کردن خط سیر افراد مختلف در برنامه‌ریزی شهری برای مدیریت امکانات تفریحی، گردشگری[۲]، در کسب وکار برای برنامه‌ریزی فعالیت‌های تبلیغاتی[۳] و در حمل‌ونقل برای سازماندهی تاکسی‌ها استفاده می‌شود [۴]. رشد جمعیت و تغییر در نیازهای آن‌ها منجر به تغییر کاربری اراضی می‌شود[۵]. برنامه‌ریزان شهری نیز نیازمند شناخت تغییرات ویژگی‌ها در مناطق مختلف شهری و تعیین ویژگی‌های تعامل بین مناطق کارآمد هستند(مورد دوم). در مورد اول، توصیف منطقه به وسیله استخراج خط سیر برای شناسایی منطقه به دست می‌آید. در مورد دوم، برای کشف و مشخص کردن ارتباط بین مناطق، داده‌کاوی خط سیر برای بررسی تکامل مرز شهری و شناسایی مشکلات موجود در شبکه حمل‌ونقل مورد توجه قرار گرفته است.

علاوه بر این، گذشت زمان و افزایش روزافزون دستگاه‌های متحرک و فناوری‌های موقعیت‌یابی و همه‌گیر شدن آن، منجر به حجم انبوه و بی‌شماری از اطلاعات مکانی-زمانی و در پی آن مشکلاتی در ذخیره‌سازی، انتقال، پردازش و پرس‌وجو از این داده‌ها ایجاد شده است. از مشکلات اساسی موجود در کاربردهایی که از داده خط سیر استفاده می‌کنند، می‌توان افزایش حجم داده برای کاربران نهایی، افزونگی[۲] داده و اشغال حافظه بالا و افزایش مدت زمان بارگیری و بارگذاری برای ارائه‌کنندگان خدمات مکانی را نام برد [۲،۳و۴]. بنابراین لازم است تحلیل‌گران برای مطالعه چنین داده‌های حجیمی[۳] قدم بردارند و آن را به دانش مفید برای اهداف خود تبدیل کنند. درنتیجه ارائه راهکار مناسب برای مدیریت مؤثر این حجم عظیم داده‌ها ضروری است. برای مدیریت مؤثر داده‌های خط سیر، استخراج نقاط توقف و حرکت[۴] با استفاده از روش‌های داده‌کاوی[۵] مانند الگوریتم خوشه‌بندی مکانی مبتنی بر تراکم کاربردهای همراه با نوفه[۶][۶] *(DBSCAN)* به عنوان یک روش خوشه‌بندی و ابزاری برای تجزیه و تحلیل داده‌های خط سیر بسیار مؤثر است [۷]. الگوریتم‌های چگالی مبنا می‌توانند هر داده‌ای را دسته‌بندی کنند و سرعت بالایی در بین روش‌های خوشه‌بندی دارند [۸]. با وجود این، ماهیت مکانی-زمانی داده‌های خط سیر چالش‌های زیادی را برای مدیریت کارآمد آن‌ها ایجاد کرده است.

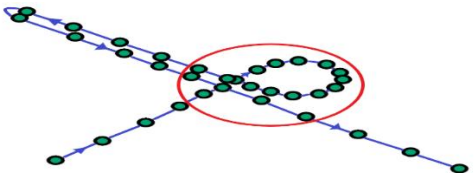
همچنین اغلب الگوریتم‌های خوشه‌بندی خط سیر، مانند الگوریتم *DBSCAN*، فقط بر اساس بعد مکان داده‌ها نقاط را خوشه‌بندی می‌کنند [۶، ۸، ۹، ۱۰، ۱۱، ۱۲، ۱۳، ۱۴] و به بعد زمان وارد نمی‌شوند. نتایج تجربی نشان می‌دهد الگوریتم خوشه‌بندی مکانی-زمانی از الگوریتم‌های سنتی مکانی عملکرد بهتری دارد [۱۵]. اگر برای داده‌های خط سیر، بعد زمان در نظر

[1] -*Trajectory*

[2] -*Redundancy*
[3] -*Big data*
[4] -*Stopping and moving point*
[5] -*Data mining*
[6] -*Density-Based Spatial Clustering of Applications with Noise*

[DOR: 20.1001.1.20089635.1400.9.2.1.0]
[DOI: 10.52547/jgit.9.2.105]

گرفته نشود و تنها بعد مکان در نظر گرفته شود، در الگوریتم‌های چگالی مبنا به اشتباه در مکان‌هایی که تردد از آن‌ها مکرر باشد رفت و برگشت به عنوان یک خوشه در نظر گرفته می‌شود (مانند شکل(۱)). مزیت استفاده از زمان در این الگوریتم جلوگیری از مشکل رفت و برگشت[1] است. البته الگوریتم‌های زیادی تا به امروز جهت بهبود الگوریتم *DBSCAN* ارائه شده است. الگوریتم‌های ارائه شده یا با توجه به ویژگی شیء متحرک یا بر اساس اطلاعات کاربر برنامه‌ریزی شده است.



شکل ۱: مشکل رفت و برگشت

در ادامه این مقاله، در بخش دوم به مبانی نظری خوشه‌بندی به خصوص الگوریتم‌های مبتنی بر تفکیک، مبتنی بر چگالی و توسعه‌های آن پرداخته می‌شود. بخش سوم به مرور مطالعات پیشین و بخش چهارم به بیان روش تحقیق می‌پردازد. بخش پنجم به پیاده‌سازی و تحلیل نتایج و بخش آخر هم به نتیجه‌گیری و پیشنهادها اختصاص دارد.

۲- مبانی نظری

خوشه‌بندی به معنای تقسیم داده‌ها به گروه‌هایی از اشیاء مشابه است. هرگروه، یک خوشه[2] نام دارد که شامل اشیایی است که بین آنها شباهت (یا شباهت‌هایی) وجود دارد و از طرفی عدم شباهت‌هایی بین اعضای یک خوشه و خوشه‌های دیگر دیده می‌شود. خوشه‌بندی یک روش کارآمد دسته‌بندی کردن بر اساس خصوصیت‌های داخلی و ذاتی داده است. به طور کلی، روش‌های خوشه‌بندی به پنج دسته تقسیم می‌شوند [۱۶]:

- خوشه‌بندی مبتنی بر تفکیک [۳] [۱۷]
- خوشه‌بندی مبتنی بر سلسه مراتب [۴] [۱۸]
- خوشه‌بندی مبتنی بر تراکم (چگالی)[۵] [۱۰]
- خوشه‌بندی مبتنی بر شبکه [۶][۱۹]
- خوشه‌بندی مبتنی بر مدل[۷][۱۹]

در خوشه‌بندی داده‌های خط سیر، این داده‌ها بر اساس الگوهای حرکت مشابه به گروه‌هایی تقسیم می‌شود. اشیاء متحرک بر اساس معیارهای مشابه حرکت مانند سرعت حرکت، جهت حرکت، فاصله مکانی، پراکندگی مکانی، مدت زمان و مفاهیم معنایی موقعیت خوشه‌بندی می‌شوند. به عنوان مثال، می‌توان از پیدا کردن مکان‌های مهم خط سیر به عنوان یکی از کاربردهای خوشه‌بندی خط سیر نام برد. در ادامه به خوشه‌بندی مبتنی بر تفکیک و مبتنی بر تراکم می‌پردازیم که از پرکاربردترین‌ها هستند و در این پژوهش از آن‌ها استفاده شده است.

۲-۱- خوشه‌بندی مبتنی بر تفکیک

فرض کنید n شیء در مجموعه داده اصلی وجود دارد. روش‌های تفکیکی، داده اصلی را به K خوشه متمایز تقسیم می‌کنند که هر خوشه با استفاده از مرکز آن خوشه مشخص می‌شود. یکی از محبوب‌ترین الگوریتم‌های این دسته *K-means* است.

۲-۱-۱- الگوریتم K میانگین *(K-Means)*

در این روش، ابتدا تعداد خوشه *(K)* توسط کاربر تعیین می‌شود. در مرحله اول از الگوریتم، K شیء از مجموعه داده‌ها به صورت تصادفی به عنوان مراکز خوشه انتخاب

[1] -*Return problem*
[2] -*Cluster*
[3]- *Partition-based method*
[4]- *Hierarchy-based method*
[5] -*Density-based method*
[6] -*Grid-based method*
[7] -*Model-based method*

می‌شوند. سپس هر شیء به خوشه‌ای که کمترین فاصله تا مرکز آن خوشه را دارد، تعلق می‌گیرد. پس از این مرحله، میانگین اشیاء تعلق یافته به یک خوشه، به عنوان مرکز جدید آن خوشه محاسبه می‌شود. این دو مرحله تا زمانی که مراکز جدید خوشه‌ها ثابت شوند و یا به تعدادی معین برسند، تکرار می‌شوند [۲۰].

مشکل اصلی الگوریتم‌های تفکیکی به‌دست آوردن تعداد بهینه خوشه‌ها است. روش‌های متفاوتی برای این کار پیشنهاد شده است. یکی از این روش‌ها البو[1] است. در این روش میانگین مجموع فواصل درون خوشه‌ای داده‌ها را به عنوان تابعی از تعداد خوشه‌ها در نظر می‌گیرد. به این ترتیب تعداد خوشه‌ها به نحوی انتخاب می‌شوند که افزودن یک خوشه دیگر، بهبودی در حداقل سازی مجموع مربعات فواصل درون خوشه‌ای ایجاد نکند. در واقع هدف از این روش یافتن مقدار K به نحوی است که برای خوشه واریانس کمتری داشته باشد [۲۱].

۲-۲-خوشه‌بندی مبتنی بر تراکم

ایده اصلی مطرح در خوشه‌بندی مبتنی بر تراکم استفاده از مفهوم فیزیکی چگالی می‌باشد. در واقع در این روش از میزان تراکم اشیاء در یک محدوده مکانی خاص، به عنوان معیاری برای تشخیص خوشه‌ها استفاده می‌شود. در روش خوشه‌بندی مبتنی بر تراکم یک خوشه با توجه به اشیاء همسایه آن (در یک شعاع خاص) شروع به رشد می‌کند و این کار تا زمانی ادامه می‌یابد که تعداد اشیاء موجود در همسایگی، بیشتر یا مساوی یک حد آستانه مشخص شود. دراین صورت، رشد خوشه‌های فعلی متوقف شده و خوشه‌های دیگری شکل خواهند گرفت. در این روش نقطه‌ها در سه دسته قرار می‌گیرند:

- نقاط مرکزی: این نقاط، نقاط درونی هستند.
- نقاط حاشیه‌ای: یک نقطه حاشیه‌ای، نقطه‌ای است که نقطه مرکزی نباشد یعنی تعداد کافی نقطه در همسایگی خود ندارد اما خودش جزء نقاط همسایه یک نقطه مرکزی است.
- نقاط نوفه: نقطه‌ای که نه مرکزی و نه حاشیه‌ای باشد.

در این روش خوشه‌ها، قسمت‌هایی از فضای داده با چگالی زیادی هستند که توسط نواحی با چگالی کمتر از یکدیگر جدا شده‌اند. این الگوریتم می‌تواند هر خوشه-ای را دسته‌بندی کند و سرعت بالایی دارد. الگوریتم-های خوشه‌بندی مبتنی بر تراکم رایج، شامل مرتب کردن نقاط برای تشخیص ساختار خوشه-بندی[2] *(OPTICS)* و خوشه‌بندی مکانی مبتنی بر تراکم برنامه‌های کاربردی همراه با نوفه هستند [۱۹]. در ادامه الگوریتم *DBSCAN* و چند توسعه این الگوریتم معرفی می‌شوند.

۲-۲-۱- الگوریتم *DBSCAN*

الگوریتم‌های خوشه‌بندی مبتنی بر چگالی یکی از روش‌های اصلی برای خوشه‌بندی در داده‌کاوی هستند. عدم محدودیت در شکل خوشه‌ها، سادگی و قابل فهم بودن از جمله مزایای این الگوریتم‌ها است. در واقع در این روش از میزان تراکم اشیاء در یک محدوده‌ی مکانی خاص، به عنوان معیاری برای تشخیص خوشه‌ها استفاده می‌شود. الگوریتم *DBSCAN* [۶] یک الگوریتم پایه در خوشه‌بندی مبتنی بر چگالی است. این الگوریتم نخستین بار توسط استر[3] و همکاران [۶] معرفی شد. الگوریتم *DBSCAN* نیاز به دو پارامتر حداقل تعداد نقاط[4] و شعاع همسایگی[5] دارد. در واقع این دو پارامتر تعیین‌کننده حداقل چگالی یک خوشه هستند. روند الگوریتم به این صورت است که ابتدا یک نقطه به صورت اختیاری انتخاب می‌شود که قبلاً بازدید نشده

[1] - *Elbow*

[2] - *Ordering Points to Identify the Clustering Structure*
[3] - *Ester*
[4] - *Minpts*
[5] - *Eps*

است. همسایگی این نقطه به شعاع *Eps* بررسی می‌شود و در صورتی که حداقل تعداد نقاط همسایگی لازم را داشت، خوشه ایجاد می‌شود و در غیر این صورت، به عنوان یک نقطه نوفه برچسب می‌خورد.

به عنوان یک کاربرد مهم، این الگوریتم برای استخراج مکان‌های با اهمیت از نظر کاربر مورد نظر مورد استفاده قرار گرفته است [۶ و ۲۲]. الگوریتم‌های خوشه‌بندی مبتنی بر چگالی می‌توانند بر بسیاری از محدودیت‌های رویکرد *K-means*، مانند وارد کردن تعداد خوشه‌ها به عنوان ورودی، غلبه کنند. با وجود این، آن‌ها فقط ابعاد مکانی را مورد توجه قرار می‌دهند و ویژگی‌های پی‌در-پی زمانی را نادیده می‌گیرند.

۲-۲-۲- الگوریتم *DVBSCAN*

الگوریتم خوشه‌بندی مکانی مبتنی بر تغییرات چگالی برنامه‌های کاربردی همراه با نوفه[1] [۱۴] *(DVBSCAN)* یکی از الگوریتم‌هایی است که به منظور رفع مشکل تغییرات چگالی الگوریتم *DBSCAN* ارائه شده است. این الگوریتم از مفهوم واریانس چگالی خوشه[2] و شاخص شباهت خوشه[3]، به منظور جلوگیری از بسط خوشه از ناحیه متراکم به ناحیه متراکم‌تر و برعکس استفاده می‌کند. الگوریتم با انتخاب یک نقطه مرکزی شروع به شکل‌دهی خوشه‌ها می‌کند. سپس همه نقاطی را که در همسایگی نقطه مرکزی انتخابی باشند، به یک صف وارد می‌کند. این نقاط در صورتی اجازه بسط پیدا می‌کنند که واریانس چگالی خوشه آن‌ها کمتر یا مساوی از حد آستانه باشد. برای به‌دست آمدن این پارامتر ابتدا از تعداد همسایگی نقاط موجود در خوشه در حال رشد میانگین‌گیری می‌شود (میانگین تراکم خوشه) و سپس واریانس چگالی این خوشه در حال رشد محاسبه می‌گردد. همچنین شاخص شباهت خوشه، یعنی اختلاف بین حداقل و حداکثر تعداد نقاط همسایگی برای هر نقطه قرار گرفته شده در خوشه در حال گسترش، نیز باید کمتر از حد آستانه باشد. در غیر این صورت، نقطه به طور ساده به خوشه افزوده می‌شود و دیگر بسط داده نمی‌شود. این الگوریتم علاوه بر دو پارامتر استفاده شده در الگوریتم *DBSCAN* نیاز به تعیین دو پارامتر حد آستانه شاخص شباهت خوشه و حد آستانه واریانس چگالی خوشه دارد که به منظور محدود کردن مقدار تغییر چگالی محلی مجاز در داخل خوشه‌ها استفاده می‌شوند. الگوریتم *DVBSCAN* قابلیت تشخیص خوشه‌های با اندازه، اشکال و چگالی متفاوت را دارد و در مقابل نوفه نیز مقاوم است ولی نیاز به تعیین چهار پارامتر دارد.

۲-۲-۳- الگوریتم *VDBSCAN*

الگوریتم خوشه‌بندی مکانی مبتنی بر چگالی متنوع برنامه‌های کاربردی همراه با نوفه[4] [۲۳] *(VDBSCAN)* به منظور رفع مشکل تجزیه و تحلیل خوشه‌های با چگالی متفاوت ارائه شده است. ایده این الگوریتم به این صورت است که قبل از اعمال الگوریتم *DBSCAN* با استفاده از مفهوم ترسیم چندفاصله‌ای[5] چگالی‌های مختلف را شناسایی کرده و برای هر چگالی یک مقدار شعاع همسایگی متناسب را برمی‌گزینند. بعد از تعیین مقادیر مختلف شعاع همسایگی، الگوریتم *DBSCAN* به تعداد چگالی‌های به دست آمده با استفاده از مقادیر مختلف شعاع همسایگی به‌دست آمده بر روی مجموعه داده اعمال می‌شود. منحنی-*k dist plot* از مرتب‌سازی نقاط مجموعه داده بر اساس فاصله‌ی هر نقطه از $k^{ام}$ین نزدیکترین همسایه‌اش ساخته می‌شود و هر تغییر شدید در این منحنی تعیین-کننده یک چگالی است.

الگوریتم *VDBSCAN* توانایی کشف اشکال اختیاری از خوشه‌ها را دارد و در مقابل نوفه قوی است. این

[1] -*Density Varied Based Spatial Clustering of Applications with Noise*
[2] -*Cluster Density Variance (CDV)*
[3] -*Cluster Similarity Index (CSI)*

[4]-*Varied Density Based Spatial Clustering of Applications with Noise*
[5]- *K-dist plot*

الگوریتم پیچیدگی زمانی مشابه *DBSCAN* دارد ولی برخلاف *DBSCAN* قابلیت تشخیص خوشه‌های با چگالی متفاوت را دارد و همچنین پارامتر شعاع همسایگی را نیز به صورت خودکار تعیین می‌کند. نیاز به پارامتر ورودی K یکی از نقاط ضعف این الگوریتم است به گونه‌ای که عدم انتخاب صحیح آن باعث تنزل دقت نتایج می‌شود. در پژوهش اسکویر و باروسو[1] و نیز شاما و آپادیا[2] برای تعیین مقدارK الگوریتم تجزیه و تحلیل خوشه مکانی مبتنی بر تراکم متنوع ارائه شده است به صورتی که این مقدار، پارامتر ورودی نیست و توسط یک الگوریتم به دست می‌آید [۲۴و۲۵].

۲-۲-۴- الگوریتم *ST-DBSCAN*

الگوریتم خوشه‌بندی مکانی-زمانی مبتنی بر تغییرات چگالی برنامه‌های کاربردی همراه با نوفه[3] [۲۶] *(ST-DBSCAN)* یکی از توسعه‌های الگوریتم *DBSCAN* است و برخلاف الگوریتم *DBSCAN* قابلیت کشف خوشه‌ها مطابق با مقادیر مکانی، غیرمکانی و زمانی اشیاء را دارد. این الگوریتم برخلاف *DBSCAN* در مواقعی که خوشه‌هایی با چگالی متفاوت در مجموعه داده وجود داشته باشند، نیز قابلیت تشخیص نوفه را دارد. از طرف دیگر، الگوریتم *DBSCAN* قابلیت تشخیص خوشه‌های با چگالی متفاوت را ندارد و پاسخ نهایی این الگوریتم به پارامتر ورودی وابسته است.

لازم به ذکر است که در این پژوهش روش *K-means* [۲۷] هم به عنوان یک روش غیر چگالی‌مبنا، پیاده‌سازی شد و نتایج آن با نتایج سایر روش‌ها مورد مقایسه گرفت. روش‌های ذکر شده می‌توانند با ویژگی‌های خاص و در برخی شرایط عملکرد مطلوبی کسب کنند. با نگاه کلی به این پژوهش‌ها می‌توان دریافت یکی از مهم‌ترین مسائلی که پژوهشگران در روش‌های پیشنهادی خود برای خوشه‌بندی داده‌های خط سیر به آن توجه داشتند تعداد پارامترهای ورودی مورد نیاز روش خوشه‌بندی، کم بودن میزان وابستگی پاسخ به این پارامترها و عدم پشتیبانی از خوشه‌های با چگالی متفاوت و مخصوصاً توانایی روش در تعیین تعداد خوشه‌ها در حین انجام عملیات خوشه‌بندی می‌باشد. حساسیت کمتر الگوریتم روش پیشنهادی به نوفه و قابلیت پیاده‌سازی بر روی مجموعه داده‌های حجیم نیز از دیگر مواردی است که در پیشنهاد یک روش خوشه‌بندی به آن توجه فراوان شده است. همچنین این الگوریتم از خوشه‌هایی با چگالی متفاوت پشتیبانی می‌کند. بنابراین با توجه به این توضیحات، سعی شد که روش پیشنهادی این تحقیق از نقاط قوت ذکر شده در بالا برخوردار باشد. در روش پیشنهادی از بعد زمان و مکان استفاده شده است. همچنین شعاع همسایگی که نتایج به آن وابسته‌تر هستند و وابستگی زیادتری به داده‌ها دارد از خود داده‌ها استخراج می‌شود. بدین ترتیب میزان وابستگی پاسخ نهایی به پارامترهای ورودی کاهش می‌یابد. علاوه بر این، روش پیشنهادی از داده‌ها شعاع‌های همسایگی متفاوتی را استخراج می‌کند. بدین ترتیب با در نظر گرفتن شعاع‌های همسایگی مختلف، توقف‌هایی که مدت زمان ماندن در آن‌ها متفاوت است به خوبی شناسایی می‌شوند.

۳- پیشینه پژوهش

در سال‌های اخیر موضوع خط سیر به واسطه اهمیت روزافزون آن، مورد توجه پژوهشگران متعددی بوده است. پژوهش‌های مربوط به خط سیر در حوزه‌های گوناگون توسعه یافته است که از آن جمله می‌توان به ایجاد روش‌های پایه برای بازسازی خط سیر خام، توسعه نوع داده جدید و پیچیده و عملگرها و زبان‌هایی برای پایگاه داده اشیاء متحرک و ایجاد الگوریتم‌های داده‌کاوی جدید برای داده‌های خط سیر اشاره کرد [۲۸، ۲۹، ۳۰و ۳۱]. اصطلاح تجزیه و تحلیل خوشه‌ای اولین بار توسط تریان[4] [۳۲] مورد استفاده قرار گرفت.

[1] *-Schoier and Borruso*
[2] *- Sharma and Upadhyay*
[3] *-Spatial-Temporal Density-Based Spatial Clustering of Applications with Noise*
[4] *-Tryon*

خوشه‌بندی تعدادی از الگوریتم‌ها و روش‌های مختلف را برای خوشه‌بندی اشیاء از نوع مشابه در دسته مربوطه در بر می‌گیرد. خوشه‌بندی یا تحلیل خوشه‌ای یکی از روش‌های متداول داده‌کاوی برای داده‌های با حجم بالا است [۱۷, ۳۳, ۳۴]. با استفاده از خوشه‌بندی، نمایشی سطح بالا از داده‌ها برای تحلیل‌گر فراهم می‌شود [۳۴].

جاین[1] و همکاران (سال۱۹۹۹ میلادی)، بیان می‌کنند که خوشه‌بندی، طبقه‌بندی غیرنظارتی الگوها، شامل مشاهدات، اقلام داده‌ها یا بردارهای خصیصه‌ها، به گروه‌ها و خوشه‌های منظم است. آنان شرح می‌دهند که مشکل خوشه‌بندی در بسیاری از زمینه‌ها، توسط پژوهشگران علوم مختلف مورد توجه قرار گرفته است. این امر، نشانگر فایده و درخواست وسیع آن به عنوان یکی از مراحل تحلیل داده‌کاوی است [۳۵].

تاکنون پژوهش‌های فراوانی در رابطه با مسیر حرکت شیء متحرک انجام گرفته است و بیشتر آنان بر روی استخراج الگوی حرکتی متمرکز شده‌اند که این امر، اهمیت تحقیق در این زمینه را نشان می‌دهد. به عنوان نمونه، آناگنس‌توپولس [2] و همکاران از داده‌های خط سیر برای پیش‌بینی مسیر حرکت شیء متحرک با استفاده از روش داده‌کاوی[3] و یادگیری ماشین[4] استفاده کردند [۳۶]. آنان پیش‌بینی مقصد بعدی کاربر را به دو صورت مکانی و مکانی-زمانی انجام دادند و مشاهده نمودند که پیش‌بینی مکانی نتایج بهتری ارائه می‌دهند. در تحقیقی دیگر، ایده و سیجیاما [5] با استفاده از مسیرهای طی شده توسط خودروها و بررسی میزان شباهت آن‌ها با یکدیگر، مسیرها را دسته‌بندی کردند و مسیر خودروها را با استفاده از این دسته‌بندی پیش‌بینی نمودند [۳۷].

با ارائه مدل توقف-حرکت، بسیاری از مطالعات بر همین اساس توسعه یافته است. مدل *SMoT*[6] یک الگوریتم تقاطع‌مبنا برای غنی‌سازی معنایی خط سیر است. در این مدل، مناطق توقف براساس تقاطع خطوط سیر با مناطق جغرافیایی از پیش تعریف شده توسط کاربر (به عنوان کاندیدای توقف) شناسایی می‌شوند. تعیین مناطق کاندیدای توقف براساس کاربرد انجام می‌شود. به عنوان مثال، در کاربرد گردشگری، کاندیداهای توقف می‌توانند فرودگاه، هتل و اماکن تفریحی و گردشگری باشند. در این مدل، کاندیداهای توقف توسط کاربر به الگوریتم معرفی می‌شود و احتمال این‌که توقفی از دید کاربر[7] دور بماند وجود دارد [۳۸].

روچا[8] و همکاران (۲۰۱۰ میلادی) شناسایی مناطق توقف و حرکت در خط سیر را بر اساس تغییر جهت نقاط تعریف کردند. خوشه‌هایی از نقاط که تغییر جهت بیشتری نسبت به نقاط دیگر داشته باشند، مناطق مهم خط سیر هستند. در روش آنان که بر روی داده‌های خط سیر کشتی‌های ماهی‌گیری (به عنوان شیء متحرک) پیاده‌سازی شد؛ هدف، یافتن مناطقی بود که فعالیت‌های ماهی‌گیری در آن بیشتر است [۳۹].

تاکنون چندین روش برای تشخیص نقاط توقف خط سیر استفاده شده است که می‌توان آن‌ها را به دو دسته کلی به شرح زیر تقسیم کرد:

- دسته اول از داده‌های خام و یا مشتقات آن برای تشخیص نقاط توقف استفاده می‌کنند. به عنوان مثال، آشبروک[9] و همکاران (سال ۲۰۰۳ میلادی) نقاط توقف را به عنوان مکان‌هایی که سرعت حاصل از داده‌ها صفر است، شناسایی نمودند [۴۰]. همچنین کروم[10] و همکاران (سال ۲۰۰۶ میلادی)، دو پارامتر را برای تشخیص نقاط توقف در نظر گرفتند که شامل سرعت پایین‌تر از ۲ مایل بر ساعت

[1]- *Jain*
[2] -*Anagnostopoulos*
[3] -*Data Mining*
[4] -*Machine Learning*
[5] *Ide and Sygiyama*

[6]-*Stop and Move of Trajectories*
[7] - اشتباه کاربر
[8]-*Rocha*
[9]-*Ashbrook*
[10] -*Krumm*

[DOR: 20.1001.1.20089635.1400.9.2.1.0]
[DOI: 10.52547/jgit.9.2.105]

یا عدم ثبت داده برای بیش از پنج دقیقه می‌باشد که در آن استفاده از مقادیر سرعت به دلیل محدودیت رویکردهای مبتنی بر تراکم *GPS*[1] قابل اعتماد نیست [۴۱]. نقاط توقف را می‌توان با تجزیه و تحلیل خوشه‌بندی نیز استخراج نمود که در آن نقاط بالقوه به عنوان مکان‌های کاندید تعمیم می‌یابند [۴۲]. همچنین تانگ[2] و همکاران (سال ۲۰۱۹ میلادی) با تغییر سرعت گره، نقاط توقف را استخراج کردند [۴۳]. یانگ[3] و همکاران (سال ۲۰۲۰ میلادی) روشی را ارائه دادند که از دو فاکتور توانایی حرکتی و زمانی و همچنین تحمل نوفه برای استخراج نقاط توقف استفاده می‌کند [۴۴]. در پژوهش مرادی و ملک (سال ۱۳۹۵) علاوه بر موقعیت مکانی گردشگر به جهت حرکت، سرعت حرکت، مسیر و زمان برای استنتاج شرایط گردشگر و سپس تطبیق سامانه پیشنهادی با این شرایط توجه شده است [۲].

- در دسته دوم، می‌توان از داده‌های خام خط سیر به همراه اطلاعات مرتبط دیگر (داده اضافی) برای مشخص کردن نقاط توقف استفاده کرد. به عنوان مثال، از نقاط مورد علاقه و نقاط دیدنی می‌توان برای استخراج نقاط توقف خط سیر مسافرت استفاده کرد [۴۵]. گنگ[4] و همکاران (سال ۲۰۱۵ میلادی) طی دو مرحله نقاط توقف با فعالیت و بدون فعالیت را شناسایی کردند. مرحله‌ی اول نقاط توقف و حرکت با استفاده از الگوریتم *DBSCAN* شناسایی می‌شود و در مرحله دوم نقاط توقف وارد روش ماشین‌های بردار پشتیبان *(SVM)* شده و توقف با فعالیت از توقف بدون فعالیت شناسایی می‌شود [۴۶]. لامب[5] و همکاران (سال ۲۰۲۰ میلادی) روش سلسله مراتبی تجمعی مبتنی بر فضا و زمان را ارائه دادند که بر روی داده‌های ردیابی حیوانات در محدوده خانه حیوانات اعمال و خوشه‌ها و یا نقاط فعالیت آن‌ها استخراج شد [۴۷]. وو[6] و همکاران (سال ۲۰۲۱ میلادی) در پژوهشی به رابطه پویای بین عناصر مکانی و رفتار-هایی که باعث توقف می‌شوند، پی بردند که با توجه به آن روشی برای استخراج نقاط توقف ارائه کردند [۱۵]. یکی از مشکلات در این دسته عدم وجود داده اضافی است.

از دیگر تحقیقاتی که به خوشه‌بندی داده‌های خط سیر با استفاده از نقاط خام پرداخته‌اند می‌توان به پژوهش کامی[7] و همکاران (سال ۲۰۱۰ میلادی) [۴۸] و تحقیق لو[8] و همکاران(سال ۲۰۱۷ میلادی) اشاره نمود [۴۹]. تحقیقات حسین‌پور میلاگردان [9] و همکاران (سال ۲۰۱۸ میلادی) [۵۰]، زیمرمان[10] و همکاران (سال ۲۰۰۹ میلادی) [۵۱] و ژوات[11] و همکاران (سال ۲۰۰۷ میلادی) [۵۲] نیز در زمرهٔ پژوهش‌هایی می-گنجند که از داده‌های اضافی بهره برده‌اند. خوشه‌بندی مکانی علاوه بر کاربردهای رایج مورد اشاره، برای کشف انواع الگوهای مکانی [۵۳] نیز مورد استفاده قرار می-گیرد. برای مثال پیله فروش‌ها و کریمی (سال ۲۰۲۰ میلادی) در تحقیقی برای تشخیص الگوهای ساختمانی و تعمیم[12] آن، یک الگوریتم بهبود یافته *DBSCAN* به نام *LA-DBSCAN*[13] ارائه نمودند. این الگوریتم با توجه به نوع داده مورد استفاده از پارامترهای جدیدی استفاده می‌کند و دارای شعاع همسایگی محلی[14] و سرتاسری[15] است و همچنین بعد مکان را در نظر می-گیرد [۵۴].

[1] -*Global Positioning System*
[2] -*Tang*
[3] -*Yang*
[4] - *Gong*
[5] -*Lamb*
[6] -*Wu*
[7] -*Kami*
[8] - *Luo*
[9] - *Hosseinpoor Milaghardan*
[10] - *Zimmermann*
[11] - *Zhouet*
[12] - *Generalization*
[13] -*Local Adaptive DBSCAN*
[14] -*Local*
[15] -*Global*

با نگاه کلی به این پژوهش‌ها می‌توان دریافت که یکی از مهم‌ترین مسائلی که پژوهشگران در روش‌های پیشنهادی خود برای خوشه‌بندی داده‌های خط سیر به آن توجه داشتند تعداد پارامترهای ورودی مورد نیاز روش خوشه‌بندی، کم بودن میزان وابستگی پاسخ به این پارامترها و عدم پشتیبانی از خوشه‌های با چگالی متفاوت و مخصوصاً توانایی روش در تعیین تعداد خوشه‌ها در حین انجام عملیات خوشه‌بندی می‌باشد. حساسیت کمتر روش به نوفه و قابلیت پیاده‌سازی بر روی مجموعه داده‌های حجیم نیز از دیگر مواردی است که در پیشنهاد یک روش خوشه‌بندی باید به آن توجه فراوان شود. بنابراین در این پژوهش سعی شده است روشی توسعه یابد که دارای این نقاط قوت باشد.

۴- روش پژوهش

در روش پیشنهادی به منظور استخراج نقاط توقف از تلفیقی از الگوریتم‌های خوشه‌بندی *DVBSCAN* و *VDBSCAN* استفاده می‌شود که با اضافه کردن بعد زمان بهبود می‌یابد. به همین دلیل، این روش خوشه-بندی مکانی-زمانی مبتنی بر چگالی متنوع در کاربرد-های دارای نوفه[1] یا به اختصار *VDBSTCAN* نامیده می‌شود. این تلفیق منجر به حل مشکل رفت و برگشت و بهبود نتایج می‌شود. در ادامه به شرح مراحل روش *VDBSTCAN* پرداخته می‌شود. هرکدام از این مرحله‌ها بر اساس ملاک و معیارهایی که برای آن‌ها در نظر گرفته شده، تکرار می‌شوند.

۴-۱- مرحله اول: محاسبه مقدار یا مقادیر شعاع همسایگی

برای یافتن مقادیر ورودی شعاع همسایگی ابتدا برای هر نقطه از مجموعه نقاط ورودی، متوسط فاصله با بیست درصد از از کل نقاط [۵۰] که نزدیک نقطه مورد نظر قرار دارند محاسبه می‌شود. فاصله متوسط بین نقاط ورودی از رابطه (۱) به دست می‌آید:

رابطه(۱)
$$d(P_i)=\sum_{\substack{j=1\\ x_j\in N}}^{\left[\frac{1}{5}\times n\right]}\frac{dist(p_i,x_i)}{\left[\frac{1}{5}\times n\right]}$$

که در آن Pi نقطه انتخابی، n تعداد کل نقاط ورودی به الگوریتم، $dist(Pi,Xj)$ فاصله‌ی اقلیدسی نقطه‌ی iاُم از نقطه‌ی Xj، و $d(Pi)$ تابع متوسط فاصله است. با به‌دست آمدن متوسط فاصله هر نقطه ابتدا این فاصله‌ها به صورت صعودی مرتب و شماره‌گذاری شده، با توجه به این اطلاعات نموداری برحسب شماره نقاط و متوسط فاصله نقاط رسم می‌گردد. سپس از نمودار مقادیر تکراری به عنوان شعاع همسایگی به الگوریتم وارد می‌شود زیرا هر تغییر در این منحنی تعیین‌کننده یک چگالی است. برای هر یک از مقادیر شعاع همسایگی که در مرحله اول یافت می‌شود، مراحل دوم تا چهارم که در ادامه آمده است، تکرار می‌شوند.

۴-۲- مرحله دوم: صف کردن نقاط

در این مرحله ابتدا یک نقطه در نظر گرفته می‌شود. اگر در شعاع مورد بررسی تعداد نقاطی که در همسایگی نقطه انتخابی بود از حد آستانه ورودی تعداد نقاط (*Minpts*) الگوریتم بیشتر بود نقطه انتخابی به عنوان مرکز خوشه در نظر گرفته می‌شود. تمامی نقاطی که در اطراف نقطه مورد نظر هستند در یک صف قرار می-گیرند و با بررسی شرط‌هایی که در ادامه مطرح شده‌اند، اجازه تشکیل و گسترش خوشه پیدا می‌کنند.

۴-۳- مرحله سوم: گسترش خوشه با توجه به بعد مکان

در این مرحله، قبل از اینکه اجازه گسترش نقطه انتخابی پردازش نشده گرفته شود، دو شرط مکانی بر روی واریانس تراکم خوشه و شاخص شباهت خوشه بررسی می‌شود:

شرط واریانس تراکم خوشه: ابتدا میانگین تراکم خوشه در حال گسترش از رابطه (۲) [۱۴] به‌دست می-آید:

رابطه(۲)
$$CDM(C)=\frac{\sum_{O\in C}|N_\varepsilon(O)|}{|C|}$$

[1]-*Varied Density-Based Spatial-Temporal Clustering of Application with Noise (VDBSTCAN)*

که در رابطه(۲)، C شماره خوشه مورد نظر، $N_\varepsilon(O)$ نشان‌دهنده تعداد اشیاء موجود در شعاع همسایگی ε نقطه انتخابی O در خوشه C و $|C|$برابر با تعداد نقاط موجود در کل خوشه مورد نظر است. سپس واریانس تراکم خوشه با توجه به میانگین تراکم خوشه با استفاده از رابطه(۳) [۱۴] به‌دست می‌آید:

رابطه(۳)

$$CDV(C)=\frac{\sum_{O\in P}\{|N_\varepsilon(O)|-CDM(C)\}^2}{|P|}$$

در رابطه(۳)، $|P|$ نماد تعداد اشیاء موجود در اطراف نقطه مرکزی خوشه است. صورت این کسر بیانگر مجموع مربع اختلاف تعداد اشیاء (نقاط) موجود در اطراف نقاط موجود در خوشه با میانگین تراکم خوشه می‌باشد. اگر واریانس تراکم خوشه کمتر از مقدار حد آستانه واریانس چگالی خوشه (α) که از ورودی دریافت شده باشد (رابطه (۴) [۱۴]) نقطه مورد نظر ممکن است امکان گسترش یافتن را در خوشه داشته باشد. در غیر این صورت خوشه امکان گسترش ندارد.

رابطه(۴) $CDV(C)\le\alpha$

شرط شاخص شباهت خوشه: شاخص شباهت خوشه[1] که تفاوت بین حداقل و حداکثر اشیاء در خوشه است با استفاده از رابطه (۵) [۱۴] به دست می‌آید:

رابطه(۵)

$$CSI(C)=MAX_\varepsilon\{|N_\varepsilon(O)|\}-MIN\{|N_\varepsilon(O)|\}$$

$MAX_{o\in p}\{|N_\varepsilon(O)|\}$ نشان‌دهنده بیشترین تعداد همسایگی میان نقاط موجود در خوشه، در شعاع ε و $MIN_{o\in p}\{|N\varepsilon(O)|\}$ نشان‌دهنده کمترین تعداد همسایگی میان نقاط موجود در خوشه، در شعاع ε است. برای برقراری این شرط باید مقدار شاخص شباهت خوشه برای خوشه درحال گسترش کمتر از حدآستانه شاخص شباهت خوشه (β) باشد (رابطه (۶) [۱۴]).

رابطه(۶) $CSI(C)\le\beta$

در صورت برقراری دو شرط واریانس تراکم خوشه و شاخص شباهت خوشه، خوشه از لحاظ مکانی اجازه گسترش می‌یابد.

۴-۴- مرحله چهارم: گسترش خوشه با توجه به بعد زمان

در این مرحله امکان گسترش خوشه با توجه به بعد زمان بررسی می‌شود. با توجه به چگالی زمانی خوشه پیوستگی زمانی نقاط موجود در خوشه در حال گسترش بررسی می‌شود. برای این منظور، میانگین چگالی زمانی خوشه[2] با توجه به رابطه(۷) محاسبه می‌شود.

رابطه(۷) $$CTDM(C)=\frac{\sum_{O\in C}|T(O)|}{|C|}$$

که در رابطه(۷)، $T(O)$ نشان‌دهنده پارامتر زمانی نقاط موجود در خوشه و $|C|$ برابر تعداد اشیاء موجود در خوشه است. سپس مطابق با رابطه (۸) انحراف معیار چگالی زمانی خوشه به دست می‌آید:

رابطه(۸)

$$CTDD(C)=\sqrt{\frac{\sum_{O\in P}\{|T_\varepsilon(O)|-CDTM(C)\}^2}{|C|}}$$

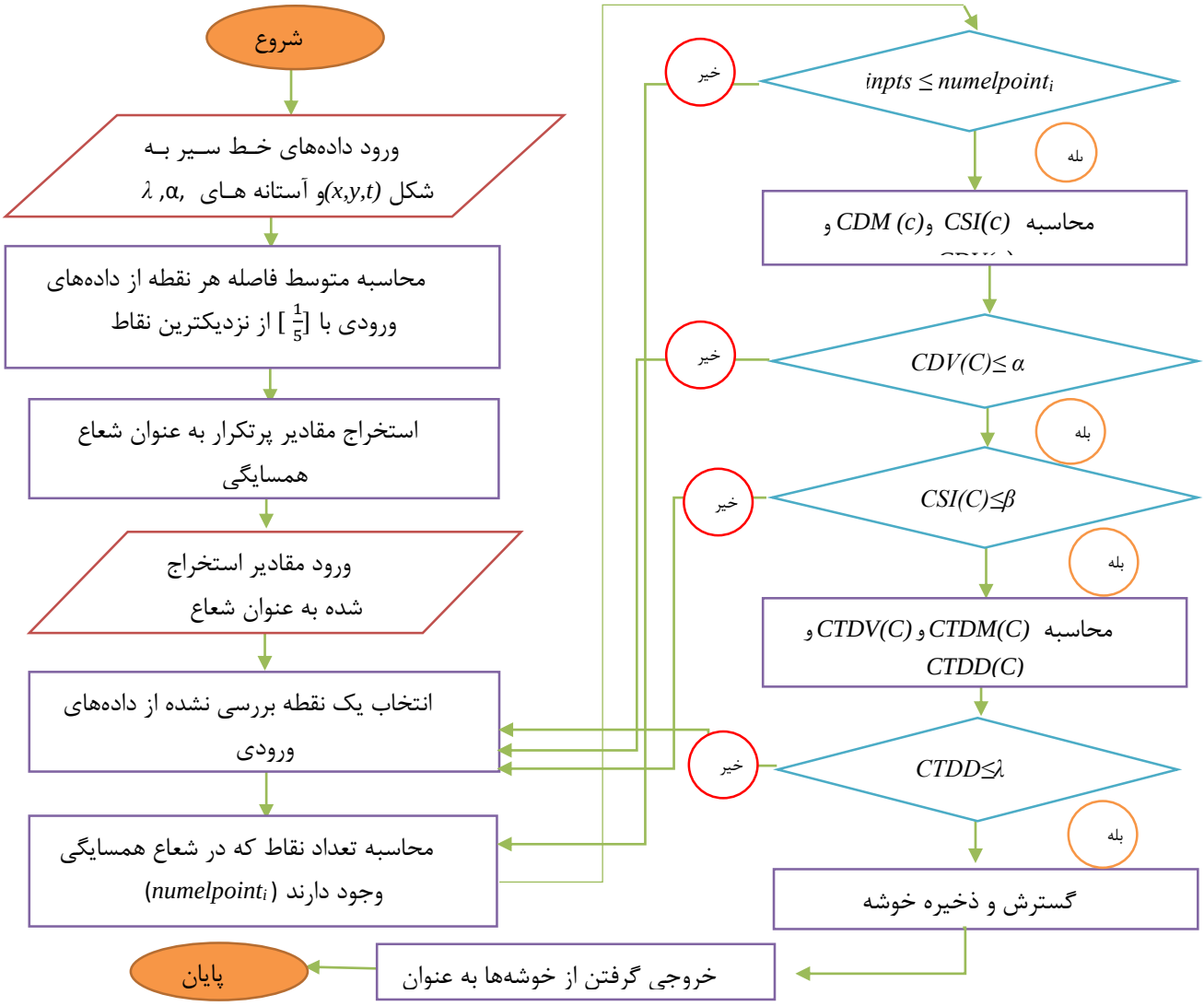
که در رابطه(۸)، $T_\varepsilon(O)$ نشان‌دهنده پارامتر زمانی نقاط موجود در خوشه است. از لحاظ زمانی نیز اگر انحراف معیار چگالی زمانی کوچکتر از حد آستانه انحراف معیار چگالی زمانی خوشه (λ) باشد (رابطه (۹)) نقطه مورد نظر امکان گسترش یافتن را در خوشه دارد. در غیر این صورت خوشه امکان گسترش ندارد.

رابطه(۹) $CTDD\le\lambda$

بدین ترتیب در صورت برقراری شرایط مکانی و زمانی، خوشه اجازه گسترش می‌یابد. مراحل بالا برای تمامی داده‌ها اجرا می‌شود و پس از هر تکرار مرحله چهارم، اگر خوشه اجازه گسترش یافته بود به عنوان یک خوشه از الگوریتم خارج می‌شود (استخراج خوشه‌های گسترش یافته). شکل(۲) روندنمای روش پیشنهادی را نمایش می‌دهد.

[1] *Cluster Similarity Index (CSI)*

[2] *Cluster time density mean (CTDM)*



شکل۲ : روندنمای روش پیشنهادی *(VDBSTCAN)*

۵- پیاده‌سازی و ارزیابی نتایج

برای ارزیابی عملکرد روش پیشنهادی، ابتدا روش‌ها روی داده‌های برداشت شده توسط نویسندگان در شهر اراک آزموده و مقایسه شدند. سپس کارایی روش پیشنهادی روی یک مجموعه داده بزرگ مربوط به شهر پکن محک زده شد.

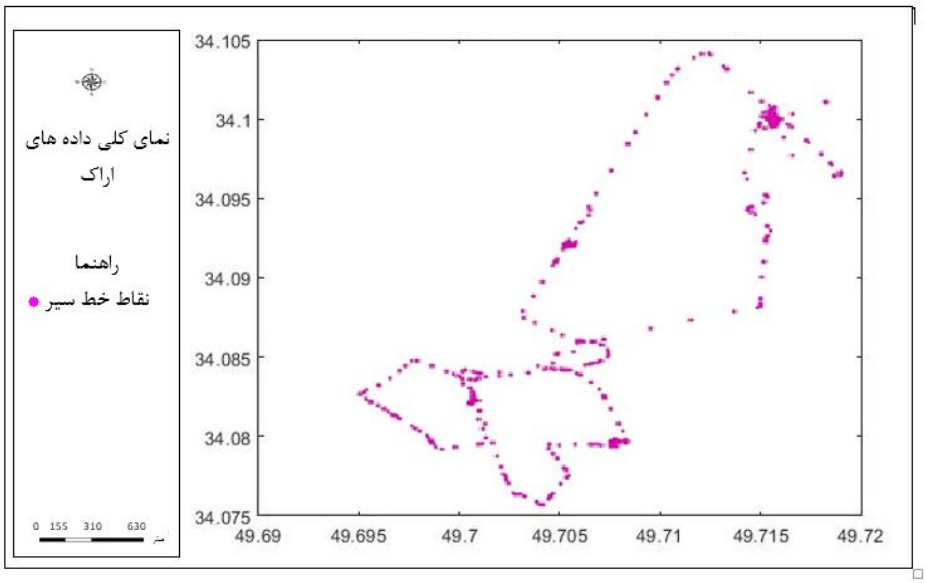
۵-۱-ارزیابی توسط داده‌های *GPS* برداشت شده

برای جمع‌آوری داده‌ها از دستگاه تعیین موقعیت جهانی (*GPS*) دستی گارمین مدل مُنتانا[1]۶۸۰ استفاده شده است. این دستگاه سیگنال ماهواره‌های *GPS* و گلوناس[2] را دریافت می‌کند و خطای برآورد آن ۲۴پا[3](در حدود ۷ متر) می‌باشد. این داده مربوط به خط سیر یک روز از

[1] *Montana680*
[2] *-GLONASS*
[3] *-Foot*

زندگی روزمره پژوهشگر این مقاله در شهر اراک در تاریخ ۱۴ شهریور ماه ۱۳۹۹ مصادف با ۴ سپتامبر ۲۰۲۰ میلادی می‌باشد. در این خط سیر سعی شده مدت زمان توقف متفاوت باشد. طول خط سیر در حدود ۳۰ کیلومتر است. نرخ برداشت این داده‌ها کمتر از یک دقیقه است. در شکل(۳) نمای کلی داده‌ها مشاهده می‌شود.

برای اجرای روش *VDBSTCAN* پارامترها به صورت ورودی توسط کاربر وارد می‌شوند تا تنظیم پارامتر صورت گیرد. در جدول (۱) مقادیر مورد استفاده در تنظیم پارامترها آورده شده است. با توجه به مقادیر میانگین انحراف معیار درون خوشه‌ای و انحراف فاصله مراکز خوشه‌ها در جدول (۱)، مقادیر پارامترهای این روش بر اساس کمتر شدن میانگین انحراف معیار درون خوشه‌ای و بیشتر شدن انحراف فاصله مراکز خوشه‌ها تعیین شد که در جدول (۲) آمده است.



شکل۳: نمای کلی داده‌های خط سیر برداشت شده در شهر اراک

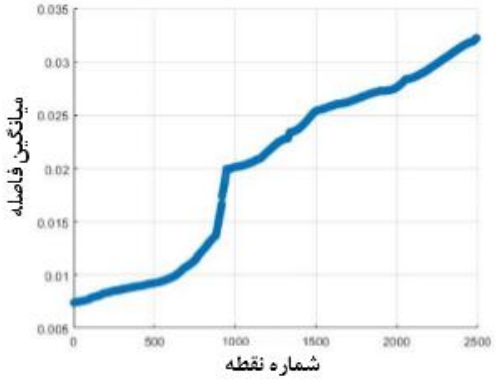
جدول ۱: مقادیر پارامترهای ورودی

حداقل تعداد نقطه (*MinPts*)	حدآستانه واریانس چگالی خوشه (α)	حدآستانه شباهت خوشه (β) (درجه)	حدآستانه انحراف معیار زمانی خوشه (λ) (ساعت)	میانگین انحراف معیار درون خوشه‌ای (سانتی-متر)	انحراف فاصله مراکز خوشه‌ها (متر)
۲۰۰	۰/۰۰۵	۰/۰۲	۰/۰۹	۱۲۲	۸۹۵/۱۷
۱۰۰	۰/۰۰۵	۰/۰۲	۰/۰۹	۷۳	۱۰۱۲/۷۳
۵۰	۰/۰۰۵	۰/۰۲	۰/۰۹	۶۹	۱۰۶۸/۸۶
۵۰	۰/۰۰۵	۰/۰۰۲	۰/۰۹	۵۱	۱۰۲۷/۲۴
۵۰	۰/۰۰۵	۰/۰۵	۰/۰۹	۸۵	۱۱۲۵/۲۹
۵۰	۰/۰۰۵	۰/۰۵	۰/۰۰۹	۱۴۹	۲۲۶/۸۹
۵۰	۰/۰۵	۰/۰۵	۰/۰۹	۹۰	۱۰۷۰/۷۸
۵۰	۰/۰۰۲	۰/۰۰۲	۰/۲	۴۷	۱۰۸۹/۲۳

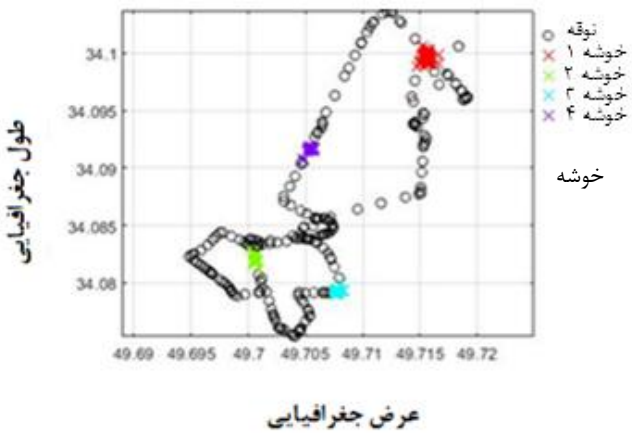
جدول۲ : مقدار انتخابی پارامترهای ورودی

نام پارامتر	مقدار پارامتر
حداقل نقطه $(MinPts)$	۵۰
حدآستانه واریانس چگالی خوشه (α)	۰٫۰۰۲
حدآستانه شاخص شباهت خوشه (β)	۰٫۰۰۲
حدآستانه انحراف معیار زمانی خوشه (λ)	۰٫۲

بعد از وارد کردن خطوط سیر و پارامترهای ورودی الگوریتم ابتدا میانگین فاصله هر نقطه با یک پنجم نقاط که به نقطه مورد نظر نزدیکتر هستند، به دست آمده، به صورت نزولی مرتب می‌شوند. در شکل (۴) این نمودار برای خط سیر مشاهده می‌شود. سپس از این نمودار مقادیر تکراری استخراج می‌شود. این مقادیر، به عنوان پارامتر شعاع همسایگی وارد الگوریتم می‌شوند. برای این داده ۱۸ شعاع متفاوت توسط الگوریتم استخراج شد. در شکل (۵) نقاط توقف استخراج شده توسط روش پیشنهادی مشاهده می‌شوند که در تطابق با فعالیت کاربر در این خط سیر مشخص شد که الگوریتم *VDBSTCAN* توانسته است کلیه نقاط توقف موجود در خط سیر را به درستی تشکیل داده است.

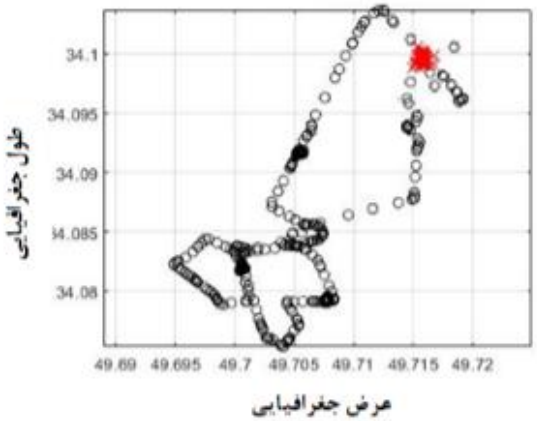


شکل۴: تقسیم‌بندی میانگین فاصله

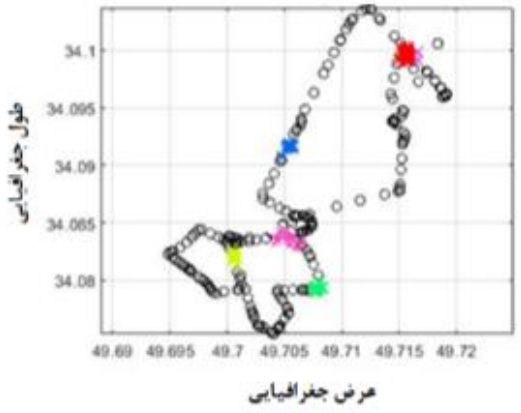


شکل۵: نقاط توقف استخراج شده توسط روش *VDBSTCAN* با حداقل نقطه برابر ۵۰

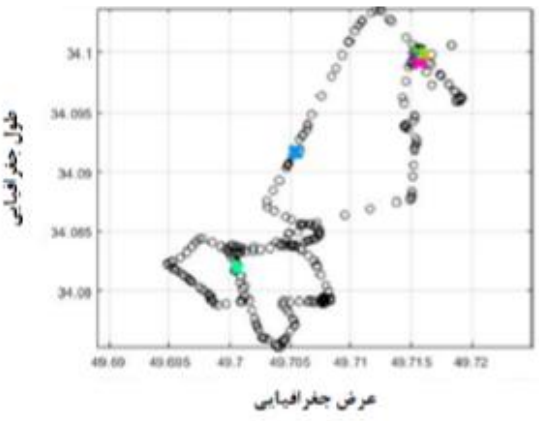
در مقام مقایسه، نتایج حاصل از برداشت خط سیر مابین همه روش‌های چگالی‌مبنا یعنی *DBSCAN*، *VDBSCAN* و*DVBSCAN* و *ST-DBSCAN* و نیز الگوریتم تفکیکی *K-means* مقایسه شد. نتایج در شکل (۶) قابل مشاهده است. در جدول (۳) و شکل (۷) مقایسه نتایج حاصل از اعمال روش‌های مختلف بر روی داده مورد نظر ارائه شده است. همچنین در شکل (۸) خط زمانی نتایج روش‌های مختلف نمایش داده می شود. همان طور که مشاهده می‌شود نتایج حاصل از الگوریتم *VDBSTCAN* با در نظر گرفتن سه معیار انحراف معیار کم‌تر و فاصله بین مراکز خوشه‌ای بیشتر و همچنین نداشتن همپوشانی زمانی در مقایسه با سایر الگوریتم‌ها روشی بهتر است. نداشتن همپوشانی زمانی در خوشه‌ها یعنی بازه زمانی شروع و پایان هر خوشه (توقف)، همپوشانی با سایر خوشه‌ها نداشته باشد، یعنی جسم متحرک نمی‌تواند در یک زمان در دو مکان متفاوت باشد. داده خط سیر استفاده شده دارای تعداد توقف‌های متفاوت و با مدت زمان‌های توقف متفاوت و در طول مسیرهای متفاوت است و الگوریتم پیشنهادی توانست تمامی مکان‌های توقف را به درستی استخراج نماید.



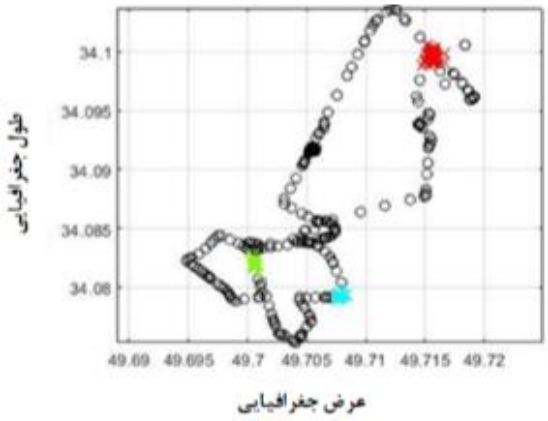
الف) *DBSCAN (ε=0.005, MinPts=50)*



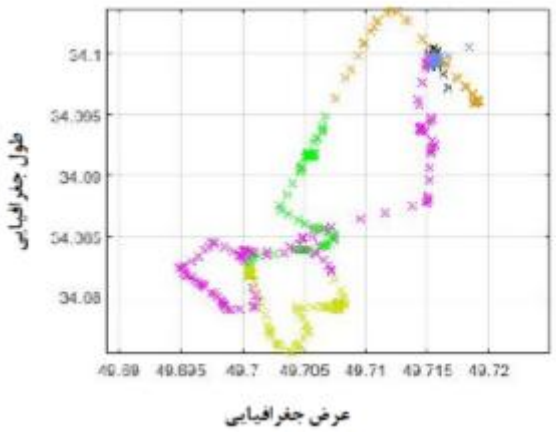
ب) *VDBSCAN (MinPts=50)*



ج) *DVBSCAN (ε =0.005, MinPts=50, β=0.02, λ=0.05)*



د) *ST-DBSCAN (ε1 =0.0006, ε2=0.002, MinPts=50)*



ه) *K-Means (K=6)*

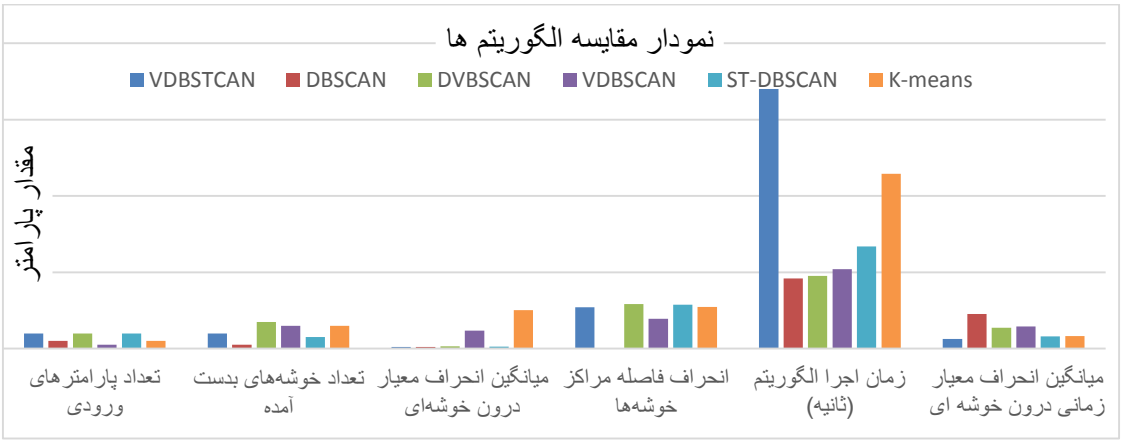
شکل ۶ : نقاط توقف استخراج شده توسط الگوریتم‌های (الف) الگوریتم*DBSCAN* ، (ب) الگوریتم*VDBSCAN* ، (ج) الگوریتم *DVBSCAN*، (د) الگوریتم *ST-DBSCAN* ، (ه) الگوریتم*K-means* ، دایره‌های توخالی، نقاط نوفه و بدون خوشه هستند و هر رنگ معرف یک خوشه می‌باشد. (تعداد خوشه‌های شناسایی شده در هر روش مطابق جدول ۳ است)

همچنین برای بررسی دقیق‌تر خوشه‌ها از شاخص انحراف معیار زمانی درون خوشه‌ای استفاده شد که این معیار همبستگی زمانی نقاط موجود در خوشه را به ما نشان می‌دهد. همان‌طور که در جدول(۳) دیده می‌شود، انحراف معیار زمانی الگوریتم پیشنهادی از سایر روش‌ها کمتر است. الگوریتم *DVBSCAN* با مشکل رفت و برگشت مواجه شد. این الگوریتم به اشتباه تردد‌های تکراری را به عنوان یک خوشه در نظر گرفته است. الگوریتم *DBSCAN* تنها نقاطی را که مدت زمان توقف در آن زیاد بود، به دست آورد ولی الگوریتم‌های دیگر توانستند سایر نقاط توقف را نیز کم و بیش به دست آورند. بنابراین نتایج حاکی از کارآمدتر و دقیق‌تر بودن الگوریتم *VDBSTCAN* به عنوان روش پیشنهادی این تحقیق است. البته ذکر این نکته ضروری است که با توجه به پردازش‌های بیشتر، طبعاً سرعت اجرای الگوریتم *VDBSTCAN* نسبت به سایر الگوریتم‌ها اندکی کمتر است ولی از آنجا که یافتن نقاط توقف کاربرد آنی ندارد این زمان، مسأله‌ساز نیست.

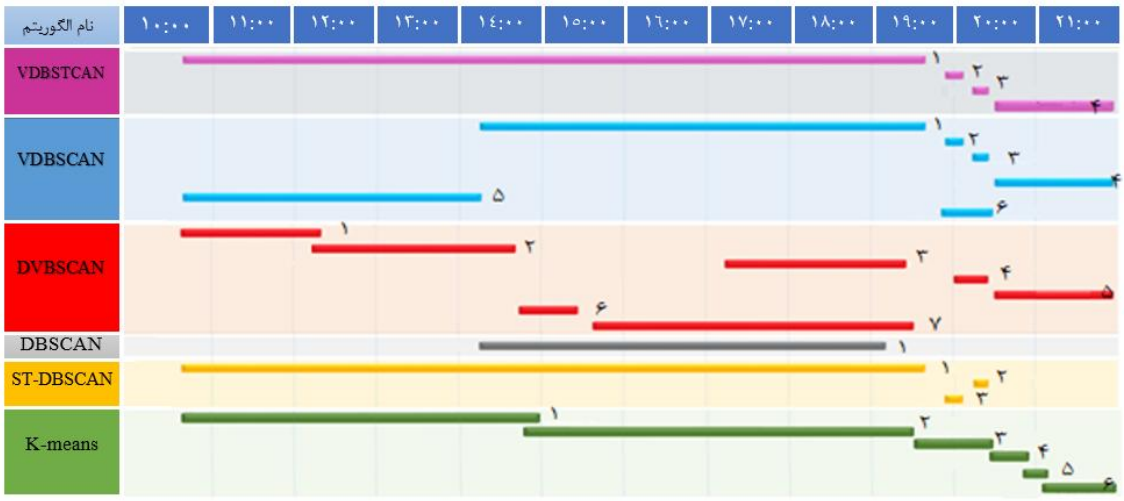
جدول ۳ : مقایسه نتایج روش‌های مختلف استخراج نقاط توقف

نام الگوریتم	تعداد پارامترهای ورودی	تعداد خوشه‌های به دست آمده	میانگین انحراف معیار درون خوشه‌ای(سانتی متر)	انحراف فاصله مراکز خوشه‌ها (متر)	زمان اجرای الگوریتم (ثانیه)	میانگین انحراف معیار زمانی درون خوشه‌ای	پارامترهای در نظر گرفته شده
VDBSTCAN	۴	۴	۴۷٫۰	۱۰۸۹٫۲۳	۶۸٫۰۶۳	۰:۳۶:۰۳	مکان و زمان
DBSCAN	۲	۱	۴۷٫۲	-	۱۸٫۴۰۸	۲:۱۰:۵۰	زمان
DVBSCAN	۴	۷	۶۴٫۳	۱۱۶۴٫۰۲	۱۹٫۰۷۷	۱:۱۸:۳۶	زمان
VDBSCAN	۱	۶	۴۶۹٫۸	۷۸۴٫۸۸	۲۰٫۸۰۷	۱:۲۴:۰۸	زمان
ST-DBSCAN	۴	۳	۴۹٫۱	۱۱۵۴٫۰۳	۲۶٫۸۰۷	۰:۴۶:۲۱	مکان و زمان
K-means	۲	۶	۱۰۱۲٫۱	۱۰۸۶٫۲۳	۴۵٫۸۵	۰۰:۴۷:۰۶	مکان

[DOR: 20.1001.1.20089635.1400.9.2.1.0]
[DOI: 10.52547/jgit.9.2.105]



شکل۷ : مقایسه نتایج روش‌های مختلف استخراج نقاط توقف



شکل۸: نمایش خط زمانی نتایج روش‌های مختلف استخراج نقاط توقف

همان طور که ذکر شد خط سیر برداشت شده است، لذا مکان‌های توقف مشخص و در اختیار بوده است. بر این اساس نقاط توقف استخراج شده توسط الگوریتم‌های *ST-DBSCAN*، *DBSCAN*، *VDBSCAN*، *DVBSCAN* ، *ST-DBSCAN* و *K-means* به ترتیب ۱۰۰٪ ، ۲۵٪ ، ۲۵٪ ، ۵۰٪ ، ۷۵٪ و ۵۰٪ به درستی استخراج شده است. میانگین انحراف معیار درون خوشه‌ای الگوریتم پیشنهادی *(VDBSTCAN)* نسبت به الگوریتم‌های *ST-DBSCAN*، *DVBSCAN*، *VDBSCAN*، *DBSCAN* و *K-means* به ترتیب حدود ۱٪ ، ۲۷٪ ، ۹۰٪ ، ۴٪ و ۹۵٪ و انحراف فاصله مراکز خوشه‌های این الگوریتم نسبت به *ST-DBSCAN* ، *DVBSCAN*، *VDBSCAN* و *K-means* به ترتیب ۶٪، ۳۹٪، ۶٪ ، ۹۹٪ و همچنین انحراف معیار زمانی درون خوشه‌ای الگوریتم پیشنهادی نسبت به الگوریتم‌های *DBSCAN*، *VDBSCAN*، *ST-DBSCAN*، *DVBSCAN*، و *K-means* به ترتیب ۷۲٪، ۵۴٪، ۵۷٪، ۲۲٪، ۷۸٪ بهبود یافته است. ولی الگوریتم پیشنهادی زمان بیشتری را برای استخراج نتایج احتیاج داشت.

۵-۲- داده‌های ژئولایف (پکن)

پس از آزمون الگوریتم *VDBSTCAN* جهت استخراج نقاط توقف و موفقیت آن، با یک هدف کاربردی‌تر از

داده‌های پروژه *Geolife* مربوط به شهر پکن، پایتخت کشور چین استفاده شد. در شکل (۹) موقعیت پکن نشان داده شده است.

شکل۹: موقعیت شهر پکن

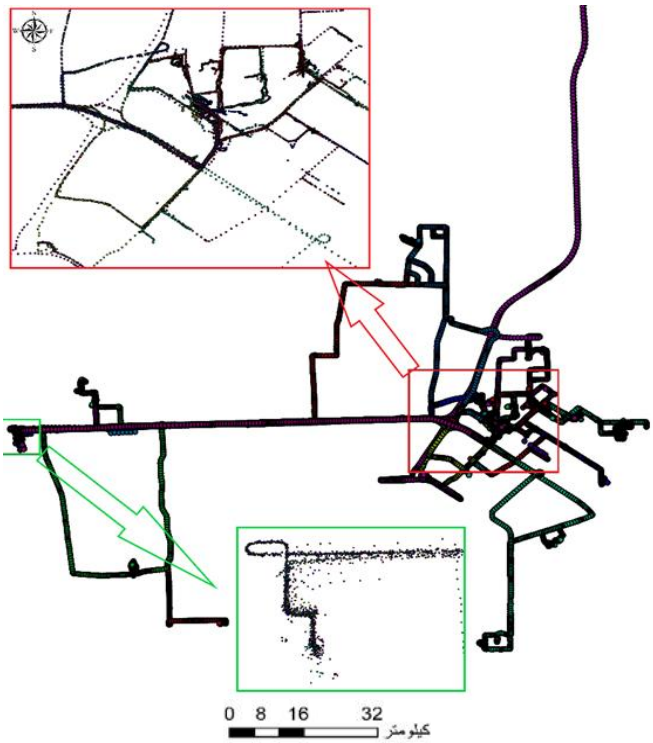
این مجموعه داده، خط سیر *GPS* در پروژه ژئولایف[1] توسط ۱۸۲ کاربر در مدت زمان ۵ سال از آوریل ۲۰۰۷ تا اوت ۲۰۱۲ میلادی جمع‌آوری شده است. خط‌سیر *GPS* این مجموعه داده توسط رشته‌ای از نقاط زمانی که هر کدام شامل اطلاعات طول و عرض جغرافیایی، ارتفاع و زمان هستند نمایش داده می‌شود. این مجموعه شامل ۱۷۶۲۱ خط سیر با مسافت کلی ۱۲۹۲۹۵۱ کیلومتر و مدت زمان کلی ۵۰۱۷۶ ساعت است. این خطوط سیر توسط دستگاه‌های *GPS* و گوشی‌های تلفن مجهز به *GPS* ضبط شده است و نرخ‌های نمونه‌برداری متنوعی دارد. ۹۱/۵ درصد از این خطوط سیر به صورت متراکمی اخذ شده‌اند یعنی هر ۱تا ۵ ثانیه یا هر ۵ تا ۱۰ متر، داده ثبت و جمع‌آوری شده است. این مجموعه داده محدوده وسیعی از حرکات و جابه‌جایی‌های خارجی کاربر، شامل نه تنها روال عادی زندگی مانند رفتن به منزل و محیط کار بلکه همچنین سرگرمی‌ها و تفریحات و فعالیت‌های ورزشی مانند دوچرخه‌سواری را هم شامل می‌شود.

در این پژوهش از مجموعه داده خط سیر فرد (پوشه) ۱۶۳ که از تاریخ ۱ دسامبر ۲۰۱۱ تا ۲۹ فوریه ۲۰۱۲ مصادف با فصل زمستان و از تاریخ ۱ ژوئن ۲۰۱۲ تا ۳۱ آگوست ۲۰۱۲ مصادف با فصل تابستان است استفاده گردید. در شکل (۱۰) نمایش کلی داده‌های فصل زمستان برای نمونه نشان داده شده است. در این بخش از تحقیق هدف، بررسی این موضوع است که الگوریتم توسعه داده شده تا چه حد قادر است با استفاده از داده‌های خط سیر، الگوهایی را از زندگی افراد استخراج نماید. در مجموعه داده مورد استفاده نقاط توقف افراد وجود ندارد بنابراین استخراج آنها بر عهده الگوریتم است و پس از آن سعی می‌شود با استفاد از شاخص-هایی، روزهای تعطیل و غیر تعطیل استخراج شود. شاخص‌ها نقش کلیدی را در شناخت انسان از منطقه و محیط‌های اطراف آن ایفا می‌کنند [۵۵].

در ابتدا برای پیاده‌سازی الگوریتم بر روی داده‌ها، خط-سیر کاربر ۱۶۳ با توجه به فصل (زمستان و تابستان) و روز هفته (روز کاری و روز تعطیل) از روی تقویم به چهار دسته شامل فصل زمستان-روزهای کاری، فصل زمستان-روزهای تعطیل، فصل تابستان-روزهای کاری، و فصل تابستان -روزهای تعطیل تقسیم‌بندی شد.

[1] -*http://research.microsoft.com/en-us/people/yuzheng/default.aspx*

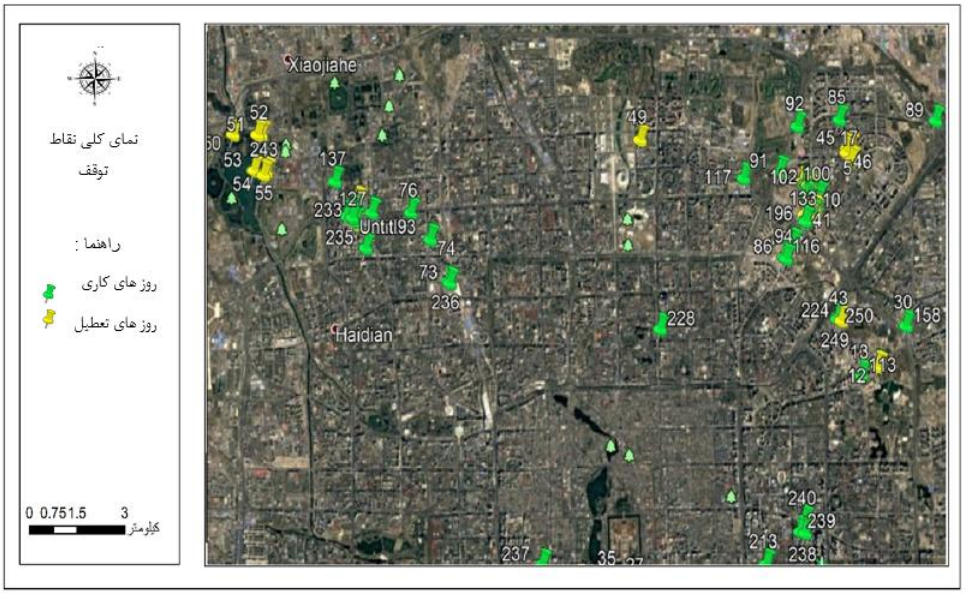
در جدول (۴) تعداد روزهای کاری و تعطیل مورد استفاده در تحقیق به تفکیک فصل مشاهده می‌شود.
سپس فایل خط سیر روزانه این کاربر وارد الگوریتم شد و نقاط توقف این خطوط سیر برای هر یک از این چهار حالت استخراج شد. شکل (۱۱) به عنوان نمونه، نمای کلی نقاط توقف فصل زمستان را به تفکیک روزهای تعطیل و کاری نشان می‌دهد.



شکل ۱۰: نمایش کلی داده‌های *Geolife* فصل زمستان

جدول ۴: تعداد روزهای کاری و تعطیل به تفکیک فصل

	روزهای تعطیل	روزهای کاری	مجموع
فصل تابستان	۱۴	۴۱	۵۵
فصل زمستان	۱۵	۵۹	۷۴
مجموع	۲۹	۱۰۰	۱۲۹



شکل ۱۱ : نمای کلی نقاط توقف فصل زمستان در روزهای کاری و تعطیل

با مقایسه خوشه‌ها در دو فصل مشاهده شد که در روز-های تعطیل فصل تابستان، تعداد و فاصله نقاط توقف بیشتر است. همچنین از بررسی الگوی رفتاری فرد در روزهای تعطیل و کاری مشاهده شد که فرد مورد نظر در روزهای کاری حداقل ۲ الی ۳ ساعت روز را در یک مکان خاص سپری کرده است ولی در روزهای تعطیل مکان خاصی برای توقف در روز مشاهده نشد (داده ها در سیستم تصویر *UTM* ثبت شده‌اند). با در نظر گرفتن فصول، روزهای کاری و تعطیل مشاهده شد که در روزهای تعطیل فصل تابستان فاصله نقاط توقف بیشتر و تعداد نقاط توقف هم بیشتر است. بعد از استخراج نقاط توقف، با بررسی‌های مکانی و زمانی، شاخص‌های زیر برای تشخیص روز کاری و تعطیل استخراج شد:

- مسافتی طی شده در طول روزهای کاری در بازه ۲۹ تا ۳۵ کیلومتر است.
- در ساعات ۱۰ صبح تا ۲ بعد از ظهر نقطه توقف در شعاع ۱۵ متری با مرکزیت (۳۹/۹۸۵۵۳۷۸۵۲ و ۱۱۶/۴۴۷۸۸۰۶۳۷) می‌باشد (محل کار برای روزهای کاری).
- در ساعات ۱۲ شب تا ۵ صبح نقطه توقف در شعاع ۱۰ متری با مرکزیت (۳۹/۹۸۶۷۸۷۷۷۶ و ۱۱۶/۴۴۸۹۵۵۲۹۹) می‌باشد.

شاخص‌های استخراج شده با تمام روزها به جز ۹ روز که با این شاخص‌ها هماهنگ نبودند تطابق داشت که دقت ۹۴٫۰۶ درصد را نشان می‌دهد. با توجه به در دسترس نبودن داده‌ها در فصل مشابه در سال بعد، این شاخص‌ها با داده‌های فصول پاییز و بهار همان سال مقایسه شد که برای بهار ۷۷ و برای پاییز تا ۶۹ درصد پیش‌بینی درست بود. همچنین از تحلیل داده‌ها نتایج زیر به‌دست آمد:

- کاربر مورد نظر در ساعت ۹ صبح تا ۴ بعدازظهر در محل ثابتی حضور داشته است. وی در این بازه مورد بررسی شغل ثابتی داشته است و احتمالاً کارمند بوده است.
- کاربر مورد نظر مسافت خانه تا محل کار را با وسیله نقلیه (ماشین شخصی یا اتوبوس یا تاکسی) طی می‌کرده است زیرا فاصله نقاط برداشت شده بیشتر شده است.

- در روزهای کاری، کاربر اکثر اوقات بدون توقف از محل کار به خانه بازمی‌گشته است و گاه در بین راه مدت زمان کوتاهی توقف می‌کرده است.
- در بازه شش ماهه بررسی شده، تردد به شهر-های اطراف نداشته ولی گهگاه به مکان‌های سرسبز اطراف شهر تردد داشته است.
- در روزهای تعطیل فصل تابستان در محیط اطراف (نزدیک به خانه) پیاده‌روی می‌کرده است و برای تفریح به استخرهای و زمین‌های بازی رو باز و پارک‌ها می‌رفته است. (بر روی نقشه نقاط توقف در مکان‌هایی بوده است که فقط محدوده مشخص شده است و ساختمان در آن دیده نشده است.)
- در روزهای تعطیل فصل زمستان به اماکن سربسته (بر روی نقشه نقاط توقف در مکان‌هایی بوده است که ساختمان در آن دیده شده است) تردد داشته است و در اواخر فصل برای تفریح به دریاچه و استادیوم تردد داشته است.

این اطلاعات از مشاهده خط سیر فرد مورد نظر که نقاط توقف آن با استفاده از الگوریتم *VDBSTCAN* استخراج شده بود به دست آمد. این اطلاعات کاربرد زیادی در عرصه‌هایی نظیر تبلیغات و برنامه‌ریزی‌های عمومی دارد. لذا در اختیار داشتن الگوریتم‌های پرقدرت برای تجزیه و تحلیل خط سیر افراد از جمله استخراج نقاط توقف بسیار ارزشمند خواهد بود.

۶- نتیجه‌گیری و پیشنهادها

در این تحقیق یک روش خوشه‌بندی برای استخراج نقاط توقف و حرکت داده‌های خط سیر ارائه گردید. روش توسعه داده شده با عنوان *VDBSTCAN* با بهره-گیری از بعد زمان توانست مشکل رفت و برگشت را نیز که در روش *VDBSCAN* دیده شده است، حل کند. بر اساس آزمایش روی داده‌های خط سیر برداشت شده، مشخص گردید روش پیشنهادی از نظر معیارهای انحراف معیار کم‌تر و انحراف فاصله بین مراکز خوشه‌ای بیشتر و نداشتن همپوشانی زمانی و همچنین تعداد خوشه‌های استخراج شده در مقایسه با الگویتم‌های *DBSCAN*، *VDBSCAN*، *DVBSCAN*، *ST-DBSCAN* و *K-means* بهتر عمل کرده است و با وجود اینکه سرعت اجرای آن نسبت به سایر الگوریتم‌ها کمتر می-باشد ولی دقت و کارایی بیشتری دارد. ترکیب استفاده از بعد زمان و مکان و به دست آوردن شعاع‌های همسایگی از خود داده‌ها روش خوبی برای بهبود عملکرد روش‌های خوشه‌بندی چگالی‌مبنا می‌باشد. از ویژگی‌های روش پیشنهادی می‌توان به کاهش میزان وابستگی نتایج به پارامترهای ورودی، استخراج نقاط توقف به طور کامل و دقیق‌تر، کاهش میزان انحراف معیار درون خوشه‌ها، افزایش فاصله مراکز خوشه‌ها، تشخیص خوشه‌ها با چگالی متفاوت، مقداردهی پارامتر-های ورودی و همچنین جلوگیری از مشکل رفت و برگشت اشاره کرد. در ادامه این تحقیق از الگوریتم توسعه داده شده برای استخراج الگوی رفتاری یکی از ساکنان شهر پکن، پایتخت چین که داده‌های خط سیر وی توسط پروژه *Geolife* ثبت شده بود، استفاده شد. با کمک خط سیر این فرد و استخراج نقاط توقف، الگوهای رفتار حرکتی وی از جمله محل کار، محل زندگی و تفریحات استخراج شد و برای آزمودن روش، روزهای کاری و تعطیل فرد با استفاده از شاخص‌هایی تنها با تحلیل خط سیر تفکیک شدند که با مقایسه با تقویم، دقت خوبی را نشان دادند.

گرچه الگوریتم پیشنهادی با استفاده از داده برداشت شده (یک روز)، نتایج بهتری از الگوریتم‌های دیگر نشان داد ولی این داده نمی‌تواند ملاک کافی برای برتری آن باشد. لذا این تحقیق با ارائهٔ الگوریتم *VDBSTCAN* محققان را برای آزمودن بیشتر آن و مقایسه‌اش با سایر روش‌ها فرامی‌خواند. همچنین پیشنهاد می‌شود با بهبود روش‌های پرس و جوی انجام گرفته در الگوریتم، از پیچیدگی زمانی آن کاست تا بتوان از آن در پایگاه داده‌های حجیم به منظور صرفه جویی در زمان و افزایش دقت استفاده کرد. در الگوریتم ارائه شده سعی شد که میزان وابستگی نتایج به پارامترهای ورودی تا حدودی کاهش یابد، ولی با استفاده از تکنیک‌هایی

[DOR: 20.1001.1.20089635.1400.9.2.1.0]
[DOI: 10.52547/jgit.9.2.105]

برای به دست آوردن مقدار حداقل نقاط از داده‌ها، میزان وابستگی نتایج به پارامترهای ورودی باز هم کاهش چشم‌گیری می‌یابد و نتایج دقیق‌تر و بهتری حاصل می‌شود که این مورد نیز به عنوان پیشنهادی برای تحقیقات آینده ارائه می‌گردد.

مراجع

[1] Q. Yu, Y. Luo, C. Chen, and X. Zheng, "Road Congestion Detection Based on Trajectory Stay-Place Clustering," ISPRS International Journal of Geo-Information, vol. 8, no. 6, pp. 264, 2019.

[2] A. MORADI, and M. MALEK, "Design and implementation of a context-aware ubiquitous GIS for tourists Case study: Maragheh City," Geographical data, vol. 27, no. 106 pp. 71-85, 2018 , (Persian).

[3] M. Azizkhani, and M. Malek, "Design and Implementation of Location-based Service for Targeted Advertising," Geospatial Engineering Journal, vol. 9, no. 2, pp. 11-16, 08/01, 2018, (Persian).

[4] J. Bian, D. Tian, Y. Tang, and D. Tao, "A survey on trajectory clustering analysis," ArXiv, vol. abs/1802.06971, 2018.

[5] M. Karimi, M. S. Mesgari, M. A. Sharifi, and P. Pilehforooshha, "Developing a methodology for modelling land use change in space and time," Journal of Spatial Science, vol. 62, no. 2, pp. 261-280, 2017.

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." Presented at the KDD, Oregon ,Portland, 1996.

[7] K. N. Ahmed, and T. Razak, "An Overview of Various Improvements of DBSCAN Algorithm in Clustering Spatial Databases," International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 2, pp. 360-363, 2016.

[8] A. Moayedi, R. Ali Abbaspour, and A. R. Chehreghan, "Assessment of the Performance of Clustering Algorithms in the Extraction of Similar Trajectories," Journal of Geomatics Science and Technology, vol. 8, no. 4, pp. 135-149, 2019 , (Persian).

[9] L. Duan, L. Xu, F. Guo, J. Lee, and B. Yan, "A local-density based spatial clustering algorithm with noise," Information Systems, vol. 32, no. 7, pp. 978-986, 2007.

[10] B. Borah, and D. K. Bhattacharyya, "DDSC : A Density Differentiated Spatial Clustering Technique," Journal of Computers, vol. 3, no. 2, pp. 72-79, 2008.

[11] C.-F. Tsai, and C.-T. Wu, "GF-DBSCAN: A new efficient and effective data clustering technique for large databases.", Presented at the MUSP'09: Proceedings of the 9th WSEAS international conference on Multimedia systems & signal processing, Hangzhou, China,2009.

[12] H. Peter, and A. A, "An Optimised Density Based Clustering Algorithm," International Journal of Computer Applications, vol. 6, no. 9, pp. 16-19, 2010.

[13] W. Ashour, and S. Sunoallah, "Multi Density DBSCAN," Presented at the Intelligent Data Engineering and Automated Learning , Berlin, Heidelberg, 2011.

[14] A. Ram, J. Sunita, A. Jalal, and K. Manoj, "A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases," International Journal of Computer Applications, vol. 3, no. 6, pp. 1-4, 2010.

[15] T. Wu, H. Shen, J. Qin, and L. Xiang, "Extracting Stops from Spatio-Temporal Trajectories within Dynamic Contextual Features," Sustainability, vol. 13, no. 2, p. 690, 2021.

[16] P. Sun, S. Xia, G. Yuan, and D. Li, "An overview of moving object trajectory compression algorithms," Mathematical

Problems in Engineering, vol. 2016, no. 3, pp. 1-13, 2016.

[17] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in 13th ACM SIGKDD international conference on Knowledge discovery and data mining, San Jose, California, USA, 2007, pp. 330–339.

[18] T. S. Madhulatha, "An overview on clustering methods," IOSR Journal of Engineering, vol. 2, no. 4, pp. 719-725, 2012.

[19] G. Yuan, P. Sun, J. Zhao, D. Li, and C. Wang, "A review of moving object trajectory clustering algorithms," Artificial Intelligence Review, vol. 47, no. 1, pp. 123-144, 2017.

[20] P.Khalife, S.Niazmardi, and R.A.Abbaspour, "Evaluation of Partitioning Methods for Clustering of Spatial Trajectories " presented at the The 1st National Conference on Data Mining in Earth Sciences, Arak, Iran, 2020, (Persian).

[21] M. Syakur, B. Khotimah, E. Rohman, and B. Dwi Satoto, "Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster," IOP Conference Series: Materials Science and Engineering, vol. 336, no. 1, pp. 12-17, 2018.

[22] G.Schoier, G.Borruso, "Individual movements and geographical data mining". In Proceedings of the International Conference on Computational Science and ITS Applications, p. 11, 2011.

[23] P. Liu, D. Zhou, and N. Wu, "VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise," presented at the 2007 International Conference on Service Systems and Service Management, Chengdu, China, 2007.

[24] A. Sharma and D. Upadhyay, "VDBSCAN Clustering with Map-Reduce Technique," presented at the Recent Findings in Intelligent Computing Techniques, Singapore, 2018.

[25] A. W. M. M. Parvez, "Data set property based 'K'in VDBSCAN Clustering Algorithm," World of Computer Science and Information Technology Journal (WCSIT), vol. 2, no. 3, pp. 115-119, 2012.

[26] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial–temporal data," Data & knowledge engineering, vol. 60, no. 1, pp. 208-221, 2007.

[27] G. Kautsar and S. Akbar, "Trajectory pattern mining using sequential pattern mining and k-means for predicting future location," in Journal of Physics: Conference Series, Medan, Indonesia, 2017, vol. 801, no. 1, pp. 12-17: IOP Publishing.

[28] A. Nasiri, S. Azimi, and R. A. Abbaspour, "Data Reduction of Spatio-temporal Trajectories using a Modified Online Compression Algorithm," Engineering Journal of Geospatial Information Technology, vol. 6, no. 3, pp. 23-38, 2018.

[29] S. Aghel Shahneshin, S. S. Mirvahabi, and R. A. Abbaspor, "An Algorithm for Compression of a Spatio-Temporal Trajectory Preserving Its Semantic Nature," Engineering Journal of Geospatial Information Technology, vol. 3, no. 4, pp. 83-95, 2016, (Persian).

[30] R. Shourouni and M. Malek, "Route recommendation based on local users' trajectories," (in eng), Journal of Geospatial Information Technology, vol. 4, no. 4, pp. 53-67, 2017, (Persian).

[31] A. Hosseinpoor Milaghardan, R. A. Abbaspour, and A. Chehreghan, "A Framework for Exploring the Frequent Patterns based on Activities Sequence," Engineering Journal of Geospatial Information Technology, vol. 7, no. 4, pp. 101-114, 2020, (Persian).

[32] R. C. Tryon, "Cumulative communality cluster analysis," Educational and Psychological Measurement, vol. 18, no. 1, pp. 3-35, 1958.

[33] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[34] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern recognition letters, vol. 31, no. 8, pp. 651-666, 2010.

[35] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM computing surveys (CSUR), vol. 31, no. 3, pp. 264-323, 1999.

[36] T. Anagnostopoulos, C. B. Anagnostopoulos, S. Hadjiefthymiades, A. Kalousis, and M. Kyriakakos, "Path prediction through data mining," in IEEE International Conference on Pervasive Services, Istanbul, Turkey 2007, pp. 128-135, 2007.

[37] T. Idé and M. Sugiyama, "Trajectory regression on road networks," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 25, no. 1, 2011.

[38] L. O. Alvares, V. Bogorny, B. Kuijpers, J. A. F. de Macedo, B. Moelans, and A. Vaisman, "A model for enriching trajectories with semantic geographical information," in Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems, Seattle, Washington, pp. 1-8, 2007

[39] J. A. M. Rocha, V. C. Times, G. Oliveira, L. O. Alvares, and V. Bogorny, "DB-SMoT: A direction-based spatio-temporal clustering method," in 2010 5th IEEE international conference intelligent systems, London, UK 2010, pp. 114-119: IEEE,2010.

[40] D. Ashbrook, and T. Starner, "Using GPS to learn significant locations and predict movement across multiple users," Personal and Ubiquitous Computing, vol. 7, no. 5, pp. 275-286, 2003.

[41] J. Krumm and E. Horvitz, "Predestination: Inferring Destinations from Partial Trajectories," in International Conference on Ubiquitous Computing, Berlin, Heidelberg, 2006, pp. 243-260: Springer Berlin Heidelberg.

[42] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen, "Discovering personal gazetteers: an interactive clustering approach," in Proceedings of the 12th annual ACM international workshop on Geographic information systems, Washington, DC, USA, pp. 266-273, 2004.

[43] J. Tang, L. Liu, and J. Wu, "A trajectory partition method based on combined movement features," Wireless Communications and Mobile Computing, vol. 2019, no. 2, pp. 1-13, 2019.

[44] Y. Yang, J. Cai, H. Yang, J. Zhang, and X. Zhao, "TAD: A trajectory clustering algorithm based on spatial-temporal density analysis," Expert Systems with Applications, vol. 139, pp. 112846, 2020.

[45] S. Shang, K. Xie, K. Zheng, J. Liu, and J.-R. Wen, "VID Join: Mapping Trajectories to Points of Interest to Support Location-Based Services," Journal of Computer Science and Technology, vol. 30, no. 4, pp. 725-744, 2015.

[46] L. Gong, T. Yamamoto, and T. Morikawa, "Identification of activity stop locations in GPS trajectories by DBSCAN-TE method combined with support vector machines," Transportation Research Procedia, vol. 23, no. 3, pp. 146-154, 2018.

[47] D. S. Lamb, J. Downs, and S. Reader, "Space-time hierarchical clustering for identifying clusters in spatiotemporal point data," ISPRS International Journal of Geo-Information, vol. 9, no. 2, pp. 85-103, 2020.

[48] N. Kami, N. Enomoto, T. Baba, and T. Yoshikawa, "Algorithm for Detecting Significant Locations from Raw GPS Data," in International Conference on Discovery Science, Berlin, Heidelberg, pp. 221-235, 2010.

[49] T. Luo, X. Zheng, G. Xu, K. Fu, and W. Ren, "An Improved DBSCAN Algorithm to Detect Stops in Individual Trajectories," ISPRS International Journal of Geo-Information, vol. 6, no. 3, pp. 63, 2017.

[50] A. H. Milaghardan, R. A. Abbaspour, and C. Claramunt, "A Dempster-Shafer based approach to the detection of trajectory stop points," Computers, Environment and Urban Systems, vol. 70, pp. 189-196, 2018.

[51] M. Zimmermann, T. Kirste, and M. Spiliopoulou, "Finding Stops in Error-Prone Trajectories of Moving Objects with Time-Based Clustering," in International Conference on Intelligent Interactive Assistance and Mobile Multimedia Computing, Berlin, Heidelberg, pp. 275-286, 2009.

[52] C. Zhou, D. Frankowski, P. Ludford Finnerty, S. Shekhar, and L. Terveen, "Discovering personally meaningful places: An interactive clustering approach," ACM Transportation Infformation Syststem., vol. 25, no. 3, pp. 12-es, 2007.

[53] J. Gudmundsson, M. van Kreveld, and B. Speckmann, "Efficient detection of motion patterns in spatio-temporal data sets," in Proceedings of the 12th annual ACM international workshop on Geographic information systems, Washington, DC, USA, pp. 250-257, 2004.

[54] P. Pilehforooshha and M. Karimi, "A local adaptive density-based algorithm for clustering polygonal buildings in urban block polygons," Geocarto International, vol. 35, no. 2, pp. 141-167, 2020.

[55] G. Gartner, and W. Hiller, "Impact of Restricted Display Size on Spatial Knowledge Acquisition in the Context of Pedestrian Navigation," Location Based Services and TeleCartography II: From Sensor Fusion to Context Models, G. Gartner and K. Rehrl, eds., pp. 155-166, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009..

[DOR: 20.1001.1.20089635.1400.9.2.1.0]
[DOI: 10.52547/jgit.9.2.105]

K. N. TOOSI UNIVERSITY OF TECHNOLOGY
FACULTY OF GEODESY AND GEOMATICS ENGINEERING

Journal of Geospatial Information Technology
Vol.9 No.2, Summer 2021

Research Paper

Developing a spatial and temporal density-based Target detection clustering algorithm to extract stop locations from the user's trajectory

Negin Masnabadi [1], Farhad Hosseinali [2*], Zahra Bahramian [2]

1- MSc Student, Department of Surveying Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.
2- Assistant Professor, Department of Surveying Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.
3- PhD, Faculty of Surveying and Geospatial Information Engineering, College of Engineering, University of Tehran, Tehran, Iran.

Abstract

Identifying stopping points of trajectories is a preliminary and necessary step in the study of moving objects and has a major impact on spatial plans and services. In this study we use trajectory clustering to extract stopping points. DBSCAN algorithm (spatial clustering based on density of applications with noise) is the basic algorithm of density-based clustering methods, which despite its advantages has some shortcommings such as difficulty in determining input parameters, inability to detect clusters with different densities and not paying attention to round trip problem. In the proposed method, which is based on density, we use of spatial and temporal indices and several neighborhood radii to extract stop points. Solving the round trip problem, extracting clusters with different densities and reducing the degree of dependence of the results on input parameters are the advantages of the proposed method.In order to evaluate the proposed method, this method was implemented on the data obtained by handheld GPS in Arak city and the data related to the Geolife research project. The obtained results were compared with the results of five other algorithms including DBSCAN, ST-BDSCAN, VDBSCAN, DVBSCAN and K-means. Compared to the manual GPS route data in Arak city, the stop locations extracted by the proposed algorithm and the mentioned algorithms are 100%, 25%, 75%, 50%, 75% and 50%, respectively, which are correctly extracted and show the superiority of the developed method. Also, after extracting the stopping and moving points, indicators from Geolife data were determined to identify working and non-working days (holidays) with which the proposed method was able to act successfully up to 94.06%.The results show a decrease in the dependence of the results on input parameters, the accurate extraction of stopping points, a reduction in the standard deviation within the clusters, and an increase in the distance between the centers of the clusters.

Key words : *Trajectory, Extraction of stop locations, Spatiotemporal clustering, DBSCAN, summer and winter, working and non-working.*

Correspondence Address*: Faculty of Civil Engineering, Shahid Rajaee Teacher Training University, Lavizan, Tehran.*
Tel*: : +98 21 22970021*
Email : *f.hosseinali@sru.ac.ir*

[DOR: 20.1001.1.20089635.1400.9.2.1.0]
[DOI: 10.52547/jgit.9.2.105]