
بسم الله الرحمن الرحيم

دانشکده مهندسی عمران

پایان‌نامه کارشناسی ارشد مهندسی راه و ترابری

ارائه‌ی روشی برای خوشه‌بندی تصادفات با استفاده از داده‌های مکانی

نگارنده

امین گنجعلی خسروشاهی

استاد راهنما

دکتر ایمان آقایان

استاد مشاور

دکتر عبدالاحد چوپانی

بهمن ۱۳۹۷

تقدیم به

پدر و مادر عزیزم

که همواره در تمام مراحل زندگی‌ام برایم تکیه‌گاهی محکم بودند.

تقدیر و تشکر

این‌جانب لازم می‌دانم که بدین‌وسیله از زحمات بی‌دریغ و نیز همراهی و راهنمایی‌های مستمر اساتید راهنمای این پایان‌نامه جناب آقای **دکتر ایمان آقایان** و جناب آقای **دکتر عبدالاحد چوپانی** تشکر و قدردانی نمایم.

تعهدنامه

این‌جانب امین گنجعلی خسروشاهی دانشجوی دوره کارشناسی ارشد رشته عمران- راه و ترابری دانشکده مهندسی عمران دانشگاه صنعتی شاهرود نویسنده پایان‌نامه ارائه‌ی روشی برای خوشه‌بندی و شناسایی عوامل مؤثر در تصادفات با استفاده از داده‌های مکانی، تحت راهنمایی دکتر ایمان آقایان و دکتر عبدالاحد چوپانی متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط این‌جانب انجام‌شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورداستفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافتهای آن‌ها) استفاده‌شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده‌شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته‌شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.

چکیده

با مشاهده آمار تصادفات به‌وضوح قابل‌درک است که تصادفات ترافیکی آثار زیان‌باری را به‌سلامت جامعه وارد می‌کند و هزینه‌های اقتصادی نامطلوبی را به افراد تحمیل می‌کند، ازاین‌رو محققین را به مطالعه‌ی تمام جوانب و روش‌های ممکن برای شناسایی تصادفات وادار می‌سازد. امروزه روش‌های داده‌کاوی بالأخص خوشه‌بندی برای تحلیل داده‌های تصادف و پیدا کردن الگوهای مکانی آن‌ها مورداستفاده قرار می‌گیرند. در این تحقیق از روشی ترکیبی برای خوشه‌بندی داده‌های تصادفات با عنوان الگوریتم GriDBSCAN استفاده کردیم. خوشه‌بندی DBSCAN به علت در نظر گرفتن تراکم داده‌های مکانی و GRID با شبکه‌بندی به علت بالا بردن دقت و سرعت محاسبات در داده‌های انبوه همچون داده‌های تصادفات قابل توجیه است. همچنین از روش‌های دیگر خوشه‌بندی همچون K-Means ، Nnh، KDE و خودهمبستگی مکانی استفاده شد تا با روش GriDBSCAN مقایسه شود. نتایج تحقیق نشان داد که الگوریتم Nnh دقیق‌ترین روش برای شناسایی مکانی نقاط حادثه‌خیز می‌باشد و الگوریتم GriDBSCAN نیز برای تفکیک مناطق پرتراکم تصادفات در داده‌های با حجم بالا کاربردی می‌باشد. خوشه‌های به‌دست‌آمده از این الگوریتم، تصادفات مناطق مختلف شهری را با توجه به تراکم آن‌ها تفکیک کرد. با بررسی پارامترهای موجود در این خوشه‌ها می‌توان در جهت کاهش تصادفات و درک عوامل مؤثر بر آن‌ها قدم برداشت.

کلمات کلیدی: تصادفات درون‌شهری، الگوریتم‌های داده‌کاوی، تحلیل خوشه‌بندی مکانی، الگوریتم GriDBSCAN

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل ۱

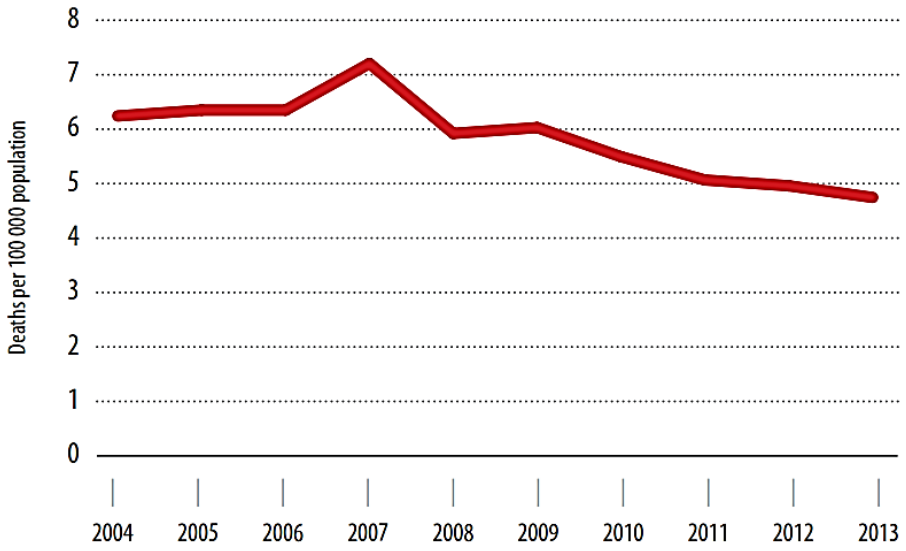
مقدمه

۱-۱- مقدمه

در دنیای امروز به علت افزایش جمعیت و نیاز به استفاده از راه‌ها مسئله‌ی حمل‌ونقل دارای اهمیت بالایی است. در هر کشور بر اساس زیرساخت‌های توسعه‌یافته در آن، یکی از شیوه‌های حمل‌ونقل صورت متداول‌تری به خود می‌گیرد. در کشور ترکیه نیز مانند بسیاری از کشورها، سفرهای جاده‌ای کاربرد بیشتری نسبت به سایر شیوه‌های حمل‌ونقل دارند؛ به همین دلیل جاده‌ها شریان‌های حیاتی برقرارکننده ارتباط بین مناطق مختلف هستند. از طرفی صدمات جانی و مالی ناشی از تصادفات رانندگی در جاده‌ها جوامع را با مشکلات عدیده‌ای روبه‌رو ساخته‌اند. حوادث رانندگی و به دنبال آن مرگ‌ومیر در حال حاضر به‌عنوان یکی از مشکلات اصلی ایمنی و سلامت عمومی در سراسر جهان مطرح است.

طبق آمار جهانی از سازمان جهانی بهداشت[1] در سال ۲۰۱۵، در سال‌های ۲۰۰۰ تا ۲۰۱۳ میزان مرگ‌ومیر ناشی از تصادفات در سراسر کشورها ۱۳ درصد افزایش‌یافته است. در سال ۲۰۱۳، به دلیل تصادفات جاده‌ای در سراسر جهان، حدود ۱,۲۵ میلیون نفر کشته‌شده‌اند و همچنین ۵۰ میلیون نفر دچار جراحت شده‌اند. علاوه بر این، تصادفات یکی از دلایل اصلی مرگ‌ومیر به‌ویژه در بین جوانان بین ۱۵ تا ۲۹ سال بوده است (سازمان بهداشت جهانی ،۲۰۱۵). با توجـه به آخرین گزارش سازمان بهداشت جهانی در سال ۲۰۱۵ برای کشور ترکیه ۳۶۸۵ فوتی(۷۷٪ مردان و ۲۳٪ زنان) گزارش‌شده است که در چند سال اخیر روند کاهشی داشته است. شکل شماره۱-۱ در گزارش سازمان بهداشت جهانی آمده است.

[1] World Health Organization



شکل ۱- ۱: روند آمار افراد فوتی ناشی از تصادفات در کشور ترکیه (WHO,2015)

گزارش کامل‌تر افراد فوتی و جراحت‌دیده در سامانه رسمی اینترنتی سازمان آمار ملی پلیس ترکیه[1] در سال ۲۰۱۳ و ۲۰۱۴ آمار تصادفات مربوط به مناطق درون‌شهری و برون‌شهری در جدول ۱-۱ آمده است.(سایت آمار تصادفات ترکیه،۲۰۱۸)

جدول ۱-۱: تعداد تصادفات شامل کشته‌ها و زخمی‌ها بر اساس نوع منطقه در کوجائلی در سال ۲۰۱۳-۲۰۱۴

سال	تعداد کل تصادفات منجر به فوت و جراحت			تعداد افراد کشته‌شده			تعداد افراد جراحت‌دیده		
	منطقه برون‌شهری	منطقه درون‌شهری	مجموع	منطقه برون‌شهری	منطقه درون‌شهری	مجموع	منطقه برون‌شهری	منطقه درون‌شهری	مجموع
۲۰۱۳	۴۱ ۲۱۱	۱۲۰ ۰۹۵	۱۶۱ ۳۰۶	۲ ۳۱۳	۱ ۳۷۲	۳ ۶۸۵	۹۱ ۵۲۲	۱۸۳ ۳۰۷	۲۷۴ ۸۲۹
۲۰۱۴	۴۱ ۹۷۵	۱۲۶ ۵۳۷	۱۶۸ ۵۱۲	۲ ۲۸۱	۱ ۲۴۳	۳ ۵۲۴	۹۳ ۴۰۶	۱۹۱ ۶۵۳	۲۸۵ ۰۵۹

آمار تصادفات ترافیکی راه‌های ترکیه، TurkStat

[1] TURKISH NATIONAL POLICE

با مشاهده آمار تصادفات به‌وضوح قابل‌درک است که تصادفات ترافیکی آثار زیان‌باری را به‌سلامت جامعه وارد می‌کند و هزینه‌های اقتصادی به افراد تحمیل می‌کند. در تحقیقات مربوط به تصادفات روش‌های مختلفی برای تحلیل تصادفات وجود دارد. امروزه روش‌های داده‌کاوی[1] بالأخص خوشه‌بندی و طبقه‌بندی برای کاهش عدم تجانس داده‌های تصادف و پیدا کردن الگوهای مبهم مورداستفاده قرار می‌گیرند. در این تحقیق از روشی ترکیبی و قابل توجیه برای خوشه‌بندی داده‌های تصادفات با عنوان الگوریتم GriDBSCAN استفاده کردیم.

۱-۲- تعریف مسئله و ضرورت انجام تحقیق

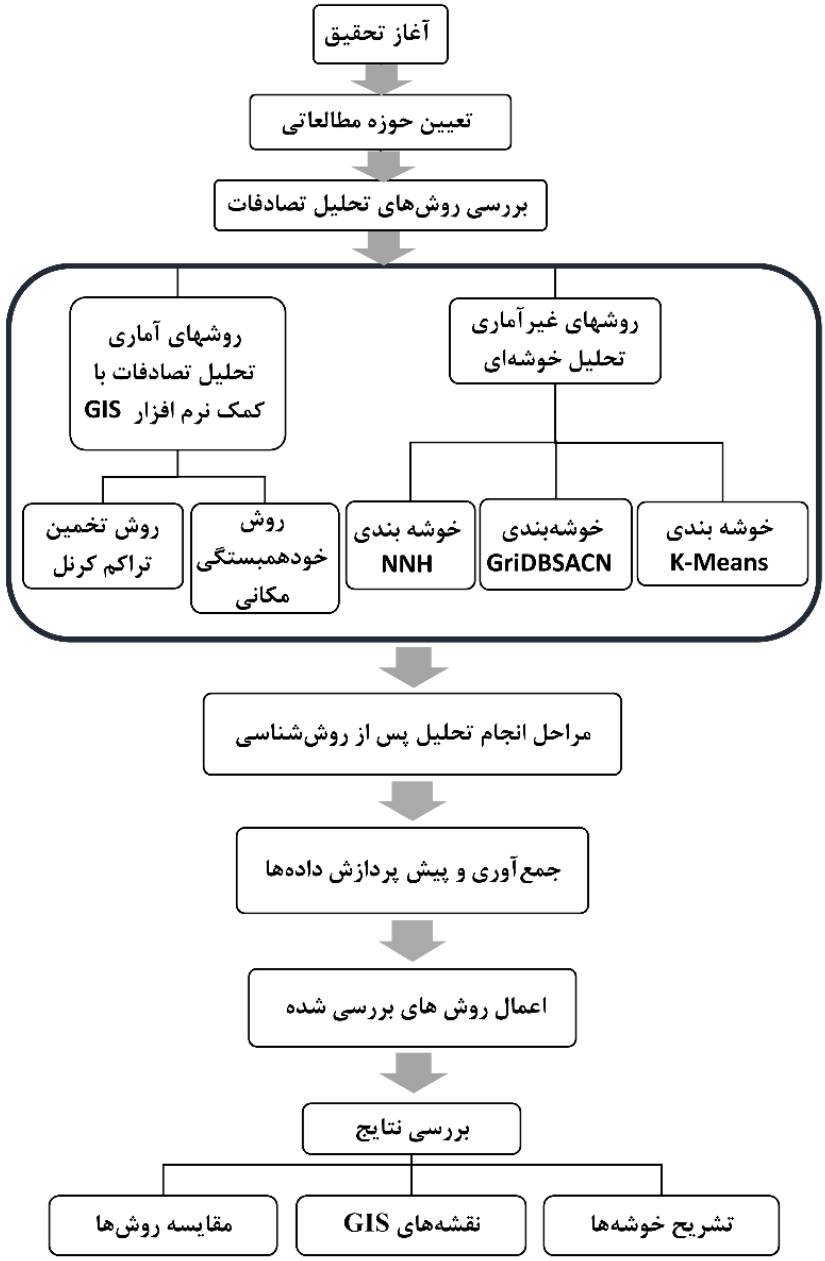
ایمنی، یک پارامتر بسیار مهم موردبررسی در بسیاری از کشورهای پیشرفته و درحال‌توسعه است که نشان‌دهنده‌ی اهمیت بالای این موضوع است. بامطالعه تصادفات می‌توان عوامل مؤثر در رخداد تصادفات ترافیکی را شناسایی کرد. تعریف و توصیف عوامل مؤثر در تصادفات گامی اساسی در جهت اقدامات عملی برای کاهش و به حداقل رساندن عوامل ایجاد تصادف است که باعث ایمنی بیشتر سیستم حمل‌ونقل و بهبود آن می‌شود. نظر به اینکه تصادف یک پدیده مکانی است، در این تحقیق رویکرد مبتنی بر مکان مورد تأکید قرارگرفته است. با استفاده از نرم‌افزارهایی مکان-مبنا همچونGIS ، مهندسان و دیگر متخصصین می‌توانند تحلیل‌های متعددی را در رابطه با منطقه جغرافیایی خاص و اتفاقات مربوط به آن منطقه مانند تصادفات را نمایش دهند.

۱-۳- روش اجرای تحقیق

در شکل ۱-۲ روند تحقیق نشان داده‌شده است. همانند دیگر تحقیقات ابتدا مطالعات در حوزه تصادفات و روش‌های تحلیل تصادف، انجام می‌شود. سپس به آماده‌سازی داده‌ها پرداختیم. سپس از

[1] Data mining

الگوریتم‌های خوشه‌بندی آماری و غیر آماری برای انجام تحلیل‌های موردنظر استفاده می‌شود. سپس به بررسی معیارهای مؤثر در خوشه‌های تصادفات پرداخته می‌شود. به عبارتی تصادفات در این مرحله بررسی می‌شوند و می‌توان نقاط پرخطر تصادفات را شناسایی و ویژگی‌های آن‌ها را به دست آورد.



شکل۱- ۲: روند کلی تحقیق

۱-۴- اهداف تحقیق

هدف اصلی این تحقیق ارائه‌ی روشی برای شناسایی نقاط حادثه‌خیز تصادفات در مناطق درون‌شهری با استفاده از روش‌های تحلیل داده‌های مـکانی با استفاده از GIS و ترکیب با روش‌های داده‌کاوی همچون خوشه‌بندی است. از دیگر اهداف این پژوهش می‌توان به موارد زیر اشاره کرد:

- تعیین تصادفات غالب در مناطق شهری
- دسته‌بندی تصادفات به گروه‌های همگن یا خوشه‌ها
- تعیین روشی مؤثر و کارآمد برای خوشه‌بندی
- تعیین خودهمبستگی مکانی عوامل مؤثر شناسایی‌شده در دوره زمانی داده‌ها
- بررسی ویژگی خوشه‌های حادثه‌خیز شناسایی‌شده در نقاط مختلف شهر

در این تحقیق به شناسایی پارامترهای تأثیرگذار در نقاط پر تصادف در مناطق مختلف شهری می-پردازیم. با تشخیص این پارامترها و سیاست‌گذاری در جهت رفع عوامل ایجاد تصادف می‌توان به میزان مطلوبی از ایمنی در راه‌ها برسیم.

۱-۵- نوآوری تحقیق

مطالعات زیادی در زمینه تحلیل تصـادفات انجام‌شده است. پژوهش‌های اخیر هرکدام بر روی پارامترهای خاص در بحث تصـادفات تـمرکز دارنـد؛ مـــانند نوع تصادف، نـــوع خودرو، نوع مـــسیر و تأثیرات آب‌وهوایی و... . داده‌های در دسترس برای این تحقیق دارای پارامترهای مختلفی همچون شرایط فیزیکی مسیر (وجود گارد ریل، شانه مسیر، شرایط روشنایی مسیر، مسیر شیب‌دار یا مسطح، مسیر قوس یا مستقیم) و برخی ویژگی‌های تصادف همچون نوع تصادف یا تعداد خودروهای درگیر تصادف هستند. ما در این تحقیق با در اختیار داشتن داده‌های جغرافیایی با مختصات جغرافیایی دقیق، با استفاده از ابزار

و روش‌های تحلیل مکانی تصادفات درون‌شهری برای دو شهر ایزمیت و گبزه (در کشور ترکیه) با ویژگی های جغرافیایی و شبکه شهری متفاوت را موردبررسی قراردادیم. درگذشته تصادفات در شهرهای دیگر ترکیه همچون افیون قره‌حصار[1] با روش‌های طبقه‌بندی مدل‌سازی شدند(پلات و دوردوران،۲۰۱۱). در این تحقیق از روش خوشه‌بندی با الگوریتم ترکیبی GriDBSCAN استفاده‌شده است. خوشه‌بندی DBSCAN به علت در نظر گرفتن تراکم[2] داده‌های مکانی و GRID با شبکه‌بندی به علت بالا بردن دقت و سرعت محاسبات در داده‌های انبوه همچون داده‌های تصادفات قابل توجیه است. همچنین از روش‌های دیگر خوشه‌بندی همچون K-Means ، NNH[3]، KDE[4] و خودهمبستگی مکانی[5] استفاده شد تا با روش GriDBSCAN مقایسه شود. به‌صورت کلی استفاده از روش‌های خوشه‌بندی در تصادفات اولاً استفاده از داده‌های دارای زیرشاخه را در خوشه‌بندی را ممکن می‌سازد. ثانیاً، به‌جای استفاده از سنجش اکتشافی[6] فاصله‌ی بین تصادفات، از تشابه مبتنی بر تراکم برای تخصیص تصادفات به خوشه‌های مختلف استفاده می‌کنند.

۱-۶- ساختار تحقیق

در فصل اول به معرفی و ضرورت انجام تحقیق و سپس به جنبه نوآوری و اهداف تحقیق پرداخته‌شده است. در فصل دوم تحقیقات و پژوهش‌های پیشین در سه بخش عوامل مؤثر بر تصادفات، روش‌های مختلف خوشه‌بندی تصادفات و استفاده از GIS در تحلیل تصادفات که درگذشته انجام‌شده، تشریح گردیده است.

[1] Afyonkarahisar
[2] Density
[3] Nearest Neighborhood Hierarchical
[4] Kernel Density Estimation
[5] Spatial Autocorrolation
[6] Heuristic measuring

در فصل سوم ضمن معرفی داده‌های مورداستفاده در این تحقیق به تشریح روش‌های متداول خوشه‌بندی نقاط تصادفات و انتخاب مناسب‌ترین و کارآمدترین آن‌ها برای این تحقیق پرداخته‌شده است.

در فصل چهارم نتایج خوشه‌بندی‌های مختلف بررسی‌شده و پارامترهای معنی‌دار در هر خوشه بررسی شده و به‌صورت آماری و تحلیل ترافیکی مورد تحلیل و بررسی قرارگرفته‌اند. درنهایت شاخص دقت پیش‌بینی برای هر نوع خوشه‌بندی محاسبه شد تا بهترین مدل مشخص شود.

در فصل پنجم خلاصه نتایج حاصل از این پژوهش ارائه‌شده است. در انتها با توجه به نتایج گرفته‌شده پیشنهادهایی جهت ادامه تحقیقات ارائه گردیده است.

فصل ۲

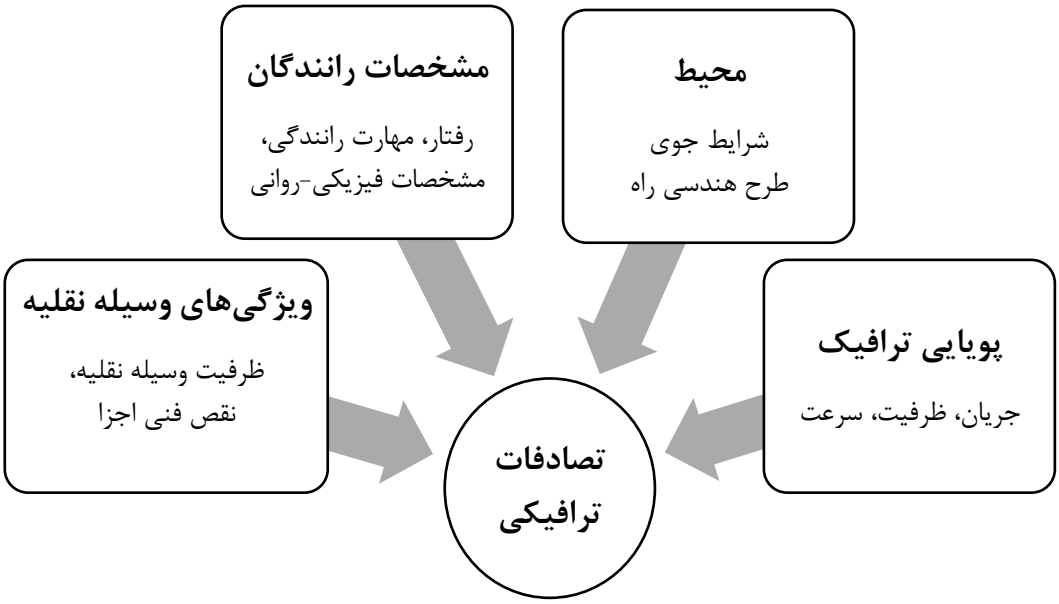
مبانی نظری و پیشینه تحقیق

۲-۱- مقدمه

در این فصل ابتدا به عوامل مؤثر بر ایجاد تصادفات می‌پردازیم و سپس به روش‌های خوشه‌بندی در تصادفات در مطالعات پیشین و در انتها به کاربرد GIS در کارهای گذشته می‌پردازیم.

۲-۲- عوامل مؤثر بر تصادفات

عوامل مؤثر بر تصادفات به چهار دسته کلی تقسیم‌بندی می‌شوند.(اوه[1] و همکاران،۲۰۰۱). با توجه به شکل ۲-۱، این عوامل با تأثیرات متقابل بر یکدیگر بر رخداد تصادفات مؤثر هستند و باوجود نقـص در یکی از این عوامل، این چرخه دچار تزلزل شده و این ناپایداری باعث وقوع تصادف می‌شود. نقاط حادثه‌خیز تصادفات از عوامل انسانی و نقص خودرو، طرح هندسی راه، کاربری‌ها و عوارض اطراف، فراوانی و نوع تصادفات تحت تأثیر قرار می‌گیرند.



شکل ۲- ۱: زنجیره عوامل مؤثر بر تصادف (اوه و همکاران،۲۰۰۱)

[1] Oh et al

۲-۲-۱- عامل محیط

طرح هندسی راه

از عوامل مهم محیطی در تصادفات می‌توان به طرح هندسی راه‌ها اشاره کرد. عوامل مؤثر بر هندسه راه را می‌توان به‌صورت زیر طبقه‌بندی کرد(معتمد، ۱۳۸۸):

- عوامل مربوط به طرح هندسی و نواقص فنی مربوط به راه
- عوامل مربوط به علائم و تجهیزات کنترلی جاده و حریم راه‌ها

در گزارشی از کمیته فنی ۳.۲ طراحی و عملکرد ایمن‌تر زیرساخت راه‌ها، انجمن جهانی راه[1] عوامل مهم در رخداد تصادفات به ویژگی‌های مسیر، طرح هندسی آن، علائم، تابلوهای هشداردهنده و کنترل ترافیک اشاره کردند. طرح هندسی که شامل ویژگی‌هایی از قبیل مسیر قوس، مسیر مستقیم، سربالایی، سرپایینی و مسیر مسطح است، از عوامل مهم در وقوع تصادفات رانندگی هستند.(Association, 2016)

چن و همکارانش اخیراً تأثیر طرح هندسی راه را در سه بزرگراه در ایالات‌متحده بررسی کردند. نتایج تحقیق نشان داد که تصادفات خسارتی در بزرگراه‌ها حساسیت بیشتری نسبت به تغییر حجم ترافیک، میانگین شیب طولی راه، عرض روفوژ میانی، عرض داخل شانه و وضعیت پیاده‌رو دارد؛ اما تصادفات فوتی و جرحی به تغییرات عرض خط، وضعیت پیاده‌رو و عرض شانه خارجی حساس است(چن و همکاران،۲۰۱۹).

[1] World Road Association

شرایط جوی

اندرسون و چپمن[1] (۲۰۱۱) در مطالعه‌ای تصادفات جاده‌ای در وست میدلند[2] انگلستان را با استفاده از سناریوهای تغییرات آب و هوایی UKCIP و شبیه‌سازی زمانی ارتباط بین دما و تصادفات شدید جاده‌ای بررسی کردند. نتایج آن‌ها نشان داد که کاهش پیش‌بینی‌شده روزهای یخبنـــدان به علت لغزندگی مســـیر، منجر به کاهـش پنجاه‌درصدی تعداد تصـادفات شد.

زائر زاده (۱۳۸۴) در بررسی نقاط حادثه‌خیز و مطالعه موردی چند نقطه حادثه‌خیز استان خراسان از دیگر عوامل مؤثر در بروز تصادفات جاده‌ای را عوامل محیطی بیان کرد. ازجمله مهم‌ترین عوامل محیطی، عوامل طبیعی و شرایط جوی است.

سجادی (۱۳۸۹) در پایان‌نامه‌ای با موضوع تحلیل تصادفات با تأکید بر خصوصیات جاده و محیط، شرایط جوی مانند بارش باران یا برف را که از عوامل تأثیرگذار بر میزان تصادفات راه‌ها است، علاوه بر تأثیر در میدان دید، سبب کاهش اصطکاک سطح جاده و به دنبال آن کاهش کارایی سیستم ترمز در وسایل نقلیه بیان کردند. نسبت تصادف در زمان بارانی نسبت به زمان غیر بارانی ۱/۷۸ برابر می‌شود.

عفتی و همکارانش (عفتی و همکاران, ۱۳۹۳) با تحلیل عوامل مکانی مؤثر بر تمرکز تصادفات در راه‌های برون‌شهری با استفاده از GIS و داده‌کاوی بیان کردند انعکاس نور خورشید نیز از عوامل محیطی است که در تسلط راننده بر محیط اطراف خود مؤثر است. در ساعات اولیه و یا پایانی روز که زاویه ارتفاعی خورشید در کمترین مقدار خود است، نور مستقیم خورشید قابلیت دید راننده را کاهش می‌دهد.

[1] Andersson & Chapman
[2] West Midland

پراسانکومار و همکارانش[1] (۲۰۱۱) با خوشه‌بندی مکانی-زمانی تصادفات جاده‌ای، اثر بادها و باران‌های موسمی را در توزیع مکانی تصادفات بررسی کردند. نتایج به‌دست‌آمده از پژوهش آن‌ها حاکی از آن است که الگوی توزیع مکانی تصادفات در هر نوع وضعیت جوی متغیر است.

۲-۲-۲- وسیله نقلیه

در این قسمت به بررسی وسیله نقلیه، به‌عنوان یکی از عوامل مؤثر بر وقوع تصادفات جاده‌ای، پرداخته می‌شود. تیان و همکارانش[2] (۲۰۱۰) همچنین مهم‌ترین عوامل مربوط به خودرو که در وقوع تصادفات مؤ-ثر هستند را به‌صورت زیر دسته‌بندی کرد:

- نقص در سیستم برقی (خرابی چراغ‌ها، زاویه اشتباه نور)
- نقص در سیستم فرمان
- نقص در سیستم ترمز (ترمز گیری با انحراف و با تأخیر)
- نقص در سیستم سفر (خرابی لاستیک‌ها، باد لاستیک‌ها)
- دیگر موارد (کیسه هوا، کمربند ایمنی و ...)

ایوان و همکارانش[3] (۲۰۰۰) نشان دادند که در تصادفاتی که تنها شامل یک خودرو می‌شود رابطه تعداد خودروها با تصادف منفی است؛ اما در مورد تصادفات شامل چند خودرو درصورتی‌که جاده تقسیم‌نشده باشد با افزایش تعداد خودرو، تعداد تصادفات نیز زیاد می‌شود، اما در جاده تقسیم‌شده (جاده‌ای که به‌وسیله نرده، چمن و غیره از هم جداشده‌اند) تصادف کاهش می‌یابد.

[1] Prasannakumar et al.
[2] Tian et al.
[3] Ivan et al.

۲-۲-۳- مشخصات رانندگان

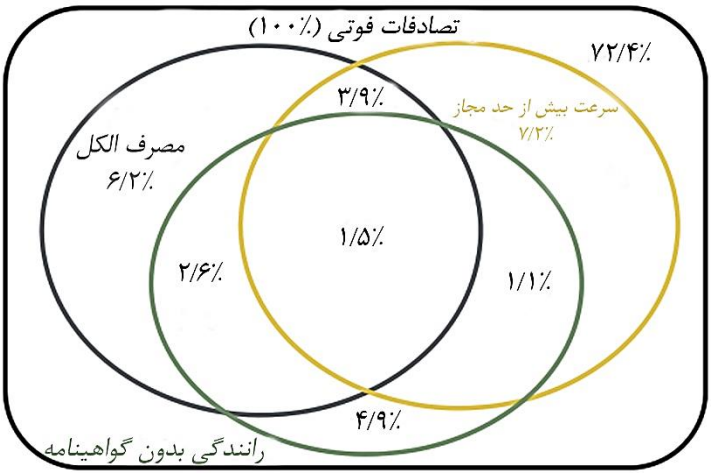
کارشناسان ایمنی بر این عقیده‌اند که تصادفات ترافیکی عمدتاً به دلیل خطاهای انسانی است. بر اساس تحقیقات سلمانی و همکارانش علل انسانی و نیز علل انسانی-محیطی () به ترتیب ۵۷٪ و ۲۶٪ از تصادفات را شامل می‌شوند. لذا، در نظر گرفتن تقابل عامل انسانی با عوامل محیطی و طرح هندسی راه در تصادفات را نمی‌توان انکار کرد(سلمانی و همکاران، ۱۳۸۷).

از عوامل انسانی مؤثر بر تصادفات می‌توان به خصوصیات فردی راننده مانند سن راننده، خستگی و خواب‌آلودگی، مصرف مواد مخدر، تخلف عمدی در هنگام رانندگی اشاره کرد. سلمانی و همکاران در سال ۱۳۸۷ مهم‌ترین عوامل انسانی تأثیرگذار بر تصادفات را به چهار گروه زیر تقسیم‌بندی کردند:

- قوه بینایی و تشخیص: عدم تمرکز ذهنی، تشخیص نادرست فاصله یا سرعت و یا ضعف دید.
- آسیب‌های جسمی: مصرف الکل، خستگی مفرط، مصرف دارو و مسائل روحی و روانی.
- شیوه رانندگی: سرعت‌های غیرمجاز و زیاد و رانندگی تهاجمی.
- عدم مهارت: تجربه ناکافی، ضعف تشخیص و واکنش نادرست رانندگان.

در تحقیق جدیدی ساگبرگ[1] (۲۰۱۸) در کشور نروژ تأثیر رانندگی افراد بدون گواهینامه در تصادفات فوتی سال ۲۰۰۴-۲۰۱۵ را بررسی کرد. این افراد ۱۰٪ کل تصادفات فوتی را شامل بودند؛ که اکثر این افراد به دو علت رایج مصرف مشروبات الکلی یا سرعت بیش‌ازحد مجاز منجر به تصادف فوتی شدند (شکل ۲-۲).

[1] Sagberg



شکل ۲- ۲: پارامترهای انسانی مؤثر بر تصادفات سال ۲۰۰۵ تا ۲۰۱۴ در کشور نروژ (ساگبرگ[1]، ۲۰۱۸)

۲-۲-۴- پویایی ترافیک

شرایط ترافیکی راه (حجم ترافیک، سرعت مجاز، ظرفیت راه و دیگر مشخصات ترافیکی) از عوامل مهم بر تصادفات است. اوه و همکارانش (اوه و همکاران، ۲۰۰۵) در مطالعه‌ای عوامل تأثیرگذار بر تصادفات در آزادراه را با استفاده از داده‌های زمان واقعی[2](داده‌های برخط) از شرایط ترافیکی به‌عنوان‌مثال کاهش ظرفیت راه و تغییرات سرعت در مسیر تونل یا عوارضی و همچنین شرایط ترافیک بلندمدت مانند میانگین ترافیک روزانه برای هرسال را در مدل‌سازی و موردبررسی قراردادند. نتایج کار آن‌ها نشان داد که کاهش تغییرات سرعت در مسیر باعث کاهش تصادفات می‌شود.

۲-۳- روش‌های خوشه‌بندی تصادفات

در این قسمت ابتدا به اهمیت داده‌کاوی[3] در مطالعات ترافیکی می‌پردازیم. سپس به بررسی خوشه‌بندی که زیرشاخه‌ی داده‌کاوی است می‌پردازیم و روش‌های استفاده‌شده در کارهای گذشته را بررسی می‌کنیم.

[1] Sagberg
[2] Real-time
[3] Data mining

۲-۳-۱-داده‌کاوی

روش داده‌کاوی به بررسی و تحلیل داده‌ها به‌منظور پیدا کردن الگوها و قوانین پنهان و معنادار درون داده‌ها اطلاق می‌گردد. انتخاب مجموعه‌ای از تکنیک‌ها برای به‌کارگیری در موقعیت خاص بستگی به سه عامل ماهیت عمل داده‌کاوی، ماهیت داده‌های موجود و مهارت‌ها و ترجیحات داده‌کاوان دارد(پرابها[۱] و شانواز[۲]، ۲۰۱۴). روش‌های داده‌کاوی به‌طور گسترده‌ای در زمینه‌های مختلف ازجمله ایمنی ترافیک مورداستفاده قرار می‌گیرند. کشف دانش[۳] در پایگاه داده موجب دستیابی آسان‌تر به اطلاعات مفید و الگوهای موجود در داده‌ها می‌شود(دونهام،۲۰۰۶).

داده‌کاوی در دو نوع هدایت‌شده[۴] و غیر هدایت‌شده[۵] ظاهر می‌شود. داده‌کاوی هدایت‌شده، دارای متغیر هدفی خاص و از پیش تعیین‌شده است که به دنبال الگویی خاص می‌گردد. درحالی‌که هدف داده‌کاوی غیر هدایت‌شده، یافتن الگوها یا تشابهات بین گروه‌هایی از اطلاعات، بدون داشتن متغیر هدفی خاص و یا مجموعه‌ای از دسته‌ها و الگوهای از پیش تعیین‌شده است. برای تبدیل مسئله موردنظر به مسئله داده‌کاوی باید آن را به یکی از شش فعالیت داده‌کاوی (دسته‌بندی[۶]، تخمین[۷]، پیش‌بینی[۸]، گروه‌بندی شباهت[۹]، خوشه‌بندی[۱۰] و توصیف و نمایه‌سازی[۱۱])، تبدیل نمود(شهرابی, ۱۳۹۲).

تعدادی از روش‌های داده‌کاوی که به‌طور مؤثر در زمینه ایمنی ترافیک استفاده‌شده‌اند، عبارت‌اند از:

---

[1] Prabha
[2] Shanavas
[3] Knowledge Discovery
[4] Supervised
[5] Unsupervised
[6] Classification
[7] Estimation
[8] Prediction
[9] Affinity Grouping
[10] Clustering
[11] Profiling

شبکه عصبی مصنوعی

در پژوهشی کنت[1] و آقایان در سال ۲۰۱۱ با استفاده از الگوریتم ژنتیک (GA)، روش جستجوی مدل و شبکه‌های عصبی مصنوعی (ANN) شدت تصادفات آزادراه تهران-قم با دوازده پارامتر مربوط به تصادفات در سال ۲۰۰۷ را مدل‌سازی کردند. این مدل‌ها با استفاده از پارامترهای ورودی رانندگی سن و جنس، استفاده از کمربند ایمنی، نوع و ایمنی وسیله نقلیه، شرایط آب و هوایی، سطح جاده، نسبت سرعت، زمان سقوط، نوع سقوط، نوع برخورد و جریان ترافیکی توسعه داده شد. نتایج نشان داد GA برای ارزیابی الگوریتم‌ها بهترین گزینه را به دست آورد. سپس روش‌های GA و PS با استفاده از بهترین معادله GA ترکیب شدند. شبکه عصبی با معماری چندلایه پیش پردازنده (MLP) استفاده‌شده است که شامل یک شبکه چندلایه‌ای خوراکی برای نورون‌های خروجی سیگموئید[2] و خروجی خطی است که می‌تواند به‌راحتی با نقشه‌های چندبعدی هم سازگار باشد.ANN در طول آموزش، تست و اعتبار سنجی اعمال شد و ۱۲ ورودی، ۲۵ نورون‌های لایه پنهان و ۳ نورون در خروجی وی داشت. مدل مناسب با توجه به R-value، RMSE، میانگین خطاهای مطلق (MAE) و مجموع خطای مربع (SSE) انتخاب شد. بالاترین R-alue برای ANN در حدود ۰,۸۷ به دست آمد که نشان می‌دهد ANN بهترین پیش‌بینی را ارائه می‌دهد. ترکیبی از روش‌های GA و PS برای رتبه‌بندی پیش‌بینی‌های مختلف از روابط خطی به معادلات پیچیده را اجازه می‌دهد. مزیت این مدل‌ها این است که با اضافه کردن داده‌های جدید می‌تواند بهبود یابد(کنت و همکاران،۲۰۱۱).

در مطالعه‌ای دلن[3] و همکارانش با استفاده از یک سری از شبکه‌های عصبی ریاضی به مدل‌سازی روابط بالقوه غیرخطی بین سطوح شدت آسیب‌دیدگی و عوامل مربوط به تصادف پرداختند. سپس آنالیز حساسیت بر روی مدل شبکه‌های عصبی آموزش‌دیده را برای شناسایی اولویت اهمیت عوامل مرتبط با

[1] Kunt
[2] Sigmoid
[3] Delen

تصادف انجام دادند. در این فرآیند، مشکل پیش‌بینی پنج کلاسه به مجموعه‌ای از مدل‌های پیش‌بینی باینری تجزیه می‌شود (با استفاده از یک نمونه آماری ملی از ۳۰۳۵۸ گزارش تصادف ثبت‌شده پلیس آمریکا) تا جزئیات اطلاعات موردنیاز برای شناسایی علت و معلول واقعی روابط بین عوامل مرتبط با تصادف و سطوح مختلف شدت جراحت را به دست آید. نتایج حاصل تغییر فاکتورهای مؤثر بر تصادفات را با تغییر شدت‌های مختلف جراحت نشان داد. (دلن و همکاران، ۲۰۰۶)

چیمبا[1] و ساندو (۲۰۰۹) با استفاده از شبکه عصبی مصنوعی از روش برگشتی[2] برای پیش‌بینی شدت جراحت در تصادفات استفاده کردند. همچنین دقت این روش را با روش پروبیت ترتیبی مقایسه کردند. نتایج نشان داد که روش شبکه عصبی مصنوعی با نورون‌ها و گره‌های بهینه‌شده بهتر از روش پروبیت ترتیبی رایج می‌باشد.

در مطالعه‌ای دیگر با استفاده از یک سری از شبکه‌های عصبی مصنوعی مدل برآورد شدت تصادف‌ها و شناسایی عوامل مربوط به تصادف در بزرگراه‌های شهری پرداخته‌شده است. اعمال شبکه‌های عصبی مصنوعی در علوم مهندسی در سال‌های اخیر ثابت‌شده است. نتایج به‌دست‌آمده نشان می‌دهد که متغیرهایی مانند عرض بزرگراه، برخورد از روبرو، نوع وسیله نقلیه در تصادف، نادیده گرفتن ترمز جانبی، فاصله پس از آن، عدم کنترل خودرو، نقض سرعت مجاز و انحراف در سمت چپ توسط رانندگان، مهم‌ترین عوامل افزایش تصادفات در بزرگراه‌های شهری است.(مقدم و همکاران،۱۳۸۹)

درخت‌های رگرسیونی و طبقه‌بندی یا درخت‌های تصمیم[3]

طبقه‌بندی و رگرسیون دو موضوع مهم در آمار و داده‌کاوی است. هر یک از این دو روش یک متغیر پاسخ را با توجه به مقادیر بردار متغیرهای پیش بینی x به دست می‌آورد. فرض کنید X دامنه x

[1] Chimba & Sando
[2] Backpropagation
[3] Decision Tree

و Y را دامنه y را مشخص کند. اگر y یک متغیر گسسته باشد، مانند مقادیر واقعی (مثلا وزن یک ماشین یا تعداد تصادفات)، مسئله رگرسیون است. در غیر این صورت، اگر Y یک مجموعه نامحدود از مقادیر غیر ارادی (مثلا نوع ماشین یا کشور مبدأ آن) باشد، مشکل طبقه‌بندی نامیده می‌شود(لو،۲۰۰۸).

درختان تصمیم گیری، یا طبقه بندی و یا رگرسیون درختی، برای سه دلیل اصلی موردتوجه است. اولا آنها نمایشی بصری دارند، در نتیجه مدل برای افراد به آسانی قابل فهم و درک است. دوم، درخت تصمیم گیری، مدل‌های غیر پارامتری است، هیچ مداخله ای از کاربر لازم نیست و به همین دلیل آنها بسیار مناسب برای کشف دانش اکتشافی هستند. سوم، الگوریتم های مقیاس پذیر، به این معنا که با توجه به افزایش اندازه داده های آموزشی[1] عملکرد ضعیف است، برای مدل‌های ساخت درخت تصمیم‌گیری وجود دارد. آخر، دقت روش درخت تصمیم‌گیری قابل مقایسه یا برتر از سایر مدل‌ها است(دوبرا،۲۰۰۲).

چانگ و چن[2] از طبقه‌بندی و رگرسیون درخت[3]، به‌عنوان یکی از گسترده‌ترین تکنیک‌های داده‌کاوی، در مدیریت کسب‌وکار، صنعت و مهندسی استفاده کردند. در این مطالعه اطلاعات مربوط به تصادفات ۲۰۰۱ – ۲۰۰۲ آزادراه ملی تایوان را جمع‌آوری کردند. مدل CART و مدل رگرسیون دوجانبه منفی برای ایجاد روابط تجربی بین حوادث ترافیکی و متغیرهای هندسی بزرگراه، خصوصیات ترافیک و عوامل محیطی تهیه شد. نتایج CART نشان داد که حجم ترافیک روزانه و متغیرهای بارش باران در محدوده زمانی معین، عوامل کلیدی برای فراوانی حوادث آزادراه بودند. با مقایسه عملکرد پیش‌بینی بین CART و مدل‌های رگرسیون منفی دوگانه منفی، این مطالعه نشان می‌دهد که CART یک روش جایگزین مناسب برای تجزیه‌وتحلیل فراوانی تصادفات آزادراه است. با مقایسه عملکرد پیش‌بینی بین

[1] Training data
[2] Chang and Chen
[3] Classification And Regression Tree (CART)

CART و مدل‌های رگرسیون دوجانبه منفی، این مطالعه نشان می‌دهد که CART یک روش مناسب برای تجزیه‌وتحلیل فراوانی تصادفات آزادراه است(چانگ و چن، ۲۰۰۵).

- **یادگیری قوانین وابستگی**[1]

گرتز[2] وهمکارانش در سال ۲۰۰۳ با استفاده از الگوریتم یادگیری قوانین وابستگی تصادفات مربوط به سال ۱۹۹۹ مربوط در بلژیک را مورد بررسی قرار دادند. این الگوها برای تحلیل و مقایسه با ویژگی های تصادفات کم‌رخداد نیز استفاده شد. از مزایای این الگوریتم، شناسایی متغیرهای مرتبط است که سهم قوی در درک بهتر شرایط حادثه و تشخیص الگوهای تصادفات توصیفی را از شرایط تصادفات بیشتر به نقاط سیاه و مناطق سیاه نشان می دهد. استفاده از این الگوریتم داده کاوی به ویژه در زمینه مجموعه داده های انبوه همچون تصادفات جاده‌ای مفید است، زیرا داده کاوی می تواند با استخراج اطلاعات از مقادیر زیادی از داده ها، توصیف شود. نتایج نشان داد که جنبه های رفتاری رانندگان در تحلیل الگوهای تصادف نمایان می‌شود. به طور کلی این عوامل نقش مهمی را در شناسایی مشکلات ایمنی ترافیک بازی می‌کنند(گرتز و همکاران،۲۰۰۳).

- **شبکه‌های کوهنن**[3]

تی سنگ و همکارانش در پژوهشی تکنیک های داده کاوی را برای پیداکردن رابطه‌ی بین عدم توجه راننده و وسایل نقلیه موتوری به کار بردند. داده‌های مورد استفاده در این تحقیق از سیستم گزارش تحلیلی فوتی ها در اداره ایمنی حمل و نقل بزرگراه‌های ملی، در سالهای ۲۰۰۰ تا ۲۰۰۳ در مریلند[4] و واشنگتن[5]، به دست آمده است. داده ها با استفاده از شبکه های کوهنون جمع شدند. سپس الگوهای و قواعد داده ها توسط درخت تصمیم گیری و مدل های شبکه عصبی بررسی شد. نتایج نشان داد، زمانی

[1] Association Rule Learning
[2] Geurts
[3] Kohonen networks
[4] Meriland
[5] Washington

که عدم توجه و شرایط فیزیکی-ذهنی راننده در یک زمان اتفاق بیافتد، راننده گرایش بالایی به تصادف با اشیاء ثابت دارد. علاوه بر این، با توجه به نوع برخورد تصادف، اهمیت نسبی تصادف با یک وسیله نقلیه متحرک نسبت به برخورد با یک جسم ثابت دو برابر بیشتر در هر تصادف است(تی‌سنگ و همکاران،۲۰۰۵).

۲-۳-۲-خوشه‌بندی

خوشه‌بندی به عمل تقسیم جمعیت ناهمگن به تعدادی از زیرمجموعه‌ها یا خوشه‌های همگن گفته می‌شود. در خوشه‌بندی داده‌ها صرفاً بر اساس تشابه گروه‌بندی می‌شوند و هیچ دسته از پیش تعیین‌شده‌ای وجود ندارد و هر گروه نیز توسط کاربر نام‌گذاری می‌شود. معمولاً از خوشه‌بندی به‌عنوان پیش‌زمینه‌ای در تحلیل‌های داده‌کاوی یا مدل‌سازی استفاده می‌شود.

در این قسمت به روش‌های خوشه‌بندی نقاط تصادف و تحلیل‌های خوشه‌ای تصادفات در کارهای گذشته می‌پردازیم. علاوه بر آن، دو روش رایج خوشه‌بندی، روش‌های غیر آماری و آماری، موردمطالعه قرارگرفته‌اند. روش غیر آماری شامل خوشه‌بندی های تفکیکی[۱]، مبتنی بر چگالی[۲] ومبتنی بر شبکه بندی[۳] و در روش خوشه‌بندی آماری نیز چگالی کرنل[۴] در دو سطح شبکه[۵] و صفحه‌ای[۶] و توابع خودهمبستگی مکانی[۷] بررسی‌شده‌اند. به‌منظور یافتن خوشه‌هایی که بهترین تناسب را با داده‌های موردنظر داشته باشند، درنهایت شاخص‌های ارزیابی ارائه می‌گردند.

در روش تحلیل خوشه‌ای که یک روش تحلیل چندمتغیری است که شامل اندازه‌های هم‌زمان روی چند متغیرند. کاهش داده‌ها یا آسان‌سازی ساختاری و دسته‌بندی کردن داده‌ها، ازجمله اهدافی

[1] Partitioning
[2] Density_based
[3] Grid-based
[4] Kernel Density
[5] Network
[6] Planar
[7] Spatial Autocorrollation

می‌باشند که در روش‌های چندمتغیری و خوشه‌بندی دنبال می‌شوند. تحلیل خوشه‌ای هم‌چنین به‌عنوان طبقه‌بندی اشیاء مشابه تعریف‌شده است به‌طوری‌که در آن، شکل تعدادی از گروه‌ها ناشناخته هستند (روسئو و کافمن[1]،۱۹۹۰). به‌طورکلی تحلیل خوشه‌ای از دو جزء اصلی تشکیل‌شده است که عبارت‌اند از اندازه‌گیری شباهت (فاصله) و الگوریتم خوشه‌بندی. هدف اصلی از تحلیل خوشه‌ای سازمان‌دهی اقلام به گروه‌هایی است که درجه شباهت برای اقلام در یک گروه حداکثر و برای آن‌هایی که بین دو گروه هستند حداقل است. علاوه بر این، روش فوق زمانی استفاده می‌شود که محققان به‌جای آزمون فرضیه-های علت و معلولی در مجموعه داده، به ویژگی‌های منحصربه‌فرد اقلام، علاقه‌مند باشند.

هدف از آنالیز خوشه‌ای تصادف‌ها، شناسایی گروه‌های همسان تصادف در مطالعاتی است که به‌منظور شناسایی راه‌کارها و اقدامات طبقه‌بندی‌شده کاهش تصادف در سطح کلان انجام می‌پذیرد. با توجه به اینکه این گروه‌ها خصوصیات مشترک بیشتری دارند، می‌توان ایده‌های کاهش تصادف را در آن‌ها به وجود آورد و راه‌کارهای رسیدن به این هدف را تدوین کرد. آنالیز خوشه‌ای عموماً روشی است که پیش از آنکه هدف‌گرا باشد، موضوع‌گرا است و می‌تواند به‌عنوان یک روش مناسب به کار رود.

۲-۳-۳-الگوریتم‌های خوشه‌بندی غیر آماری

تفاوت خوشه‌بندی آماری و خوشه‌بندی غیر آماری در این است که روش‌های خوشه‌بندی آماری خودهمبستگی بین پارامترهای مؤثر در تصادفات را مشخص می‌کند، درحالی‌که روش‌های غیر آماری بر اساس تابع فاصله هستند به عبارتی شباهت بین داده‌ها بر اساس فاصله آن‌ها تعیین می‌شود. چهار نوع طبقه‌بندی رایج در روش‌های خوشه‌بندی غیر آماری وجود دارد:(هان و همکاران،۲۰۱۱)

[1] Rousseeuw and Kaufman

- الگوریتم‌های خوشه‌بندی افرازی[1]
- خوشه بندی سلسله مراتبی[2]
- خوشه‌بندی مبتنی بر تراکم[3]
- الگوریتم‌های مبتنی بر شبکه[4]

جدول ۲-۱: طبقه بندی الگوریتم‌های خوشه‌بندی غیر آماری (هان و همکاران،۲۰۱۱)

روش خوشه‌بندی	خصوصیات عمومی
افرازی	- یافتن خوشه های منحصر به فرد از شکل کروی - مبتنی بر فاصله - استفاده از مدل Means یا Medoide (و غیره) برای نشان دادن مرکز خوشه - مؤثر برای مجموعه داده های کوچک و متوسط - تعداد خوشه‌ی از پیش‌تعیین‌شده
سلسله مراتبی	- خوشه بندی در این نوع یک تجزیه سلسله مراتبی است. - نمی توان اشتباهات ادغام شده و یا تقسیم شده را تصحیح کرد. - ممکن است تکنیک های دیگری مانند میکرو خوشه بندی را در نظر بگیریم یا ارتباطات شی را در نظر بگیریم.
مبتنی بر تراکم	- می تواند خوشه های با شکل دلخواه پیدا کند - خوشه ها، مناطق متراکم از اشیاء در فضا هستند که با مناطق کم تراکم از هم جدا می شوند - چگالی خوشه: هر نقطه باید حداقل تعداد نقاط در داخل آن همسایگی داشته باشد. - داده‌هایی که معیارهای خوشه‌بندی را ندارند داده‌های خارج از محدوده[5] به حساب می‌آیند.
مبتنی بر شبکه	- از یک ساختار شبکه چندتایی داده استفاده می‌کنند. - زمان پردازش سریع (به طور معمول مستقل از تعداد اشیاء داده، در عین حال وابسته به اندازه شبکه)

[1] Partitioning
[2] Hierarchical
[3] Density-Based
[4] Grid-Based
[5] Outlier

خوشه‌بندی افرازی

الگوریتم k-means و الگوریتم بیشینه‌سازی امید ریاضی[1] (EM)، نمایندگان اصلی برای الگوریتم‌های خوشه بندی افرازی می‌باشند. الگوریتم k-means نشان‌دهنده یک خوشه با استفاده از یک نقطه واحد است، در حالی که EM از توزیع احتمالی برای نشان دادن یک خوشه استفاده می کند. الگوریتم‌های تفکیکی نیاز به ارائه پارامتر، k، تعداد خوشه ها و انجام مکانیابی مجدد عضو بودن داده را دارند تا زمانی که عضویت دیگر تغییر نکند یا تغییر در محدوده قابل تحمل باشد. کیفیت خوشه بندی به مقدار k وابسته است و به طور کلی، الگوریتم‌های تفکیکی نمی‌توانند به صورت خوشه ای با شکل های دلخواه درآیند.

در الگوریتم افرازی، همه خوشه‌ها به‌یک‌باره مشخص می‌شوند؛ به عبارتی روش افرازی، داده‌ها را به تعداد خوشه‌های از پیش تعیین‌شده (k) از گروه‌های دوبه‌دو ناسازگار و جامع، تقسیم‌بندی می‌کند.(Han et al., 2011) در خوشه‌بندی افرازی داده‌ها به گروه‌هایی تقسیم‌بندی می‌شوند که هر گروه نشان‌دهنده یک خوشه است؛ به‌عبارت‌دیگر این روش تقسیم مجموعه‌ای از داده‌ها به زیرمجموعه غیر همپوشان[2] است به‌گونه‌ای که هر شئ داده دقیقاً در یک زیرمجموعه است؛ بنابراین تعداد خوشه‌ها از قبل تعیین‌شده است(Kaufman and Rousseeuw, 1990).

الگوریتم K-Means

این روش خوشه‌بندی یکی از مشهورترین و ساده‌ترین رویکردهای خوشه‌بندی است. مک کوئین[3] در سال ۱۹۶۷ اولین شخصی بود که از عبارت K-Means استفاده کرد(مک کوئین،۱۹۶۷). هدف، تقسیم بندی n مشاهده(اشیاء) به k خوشه است به‌طوری‌که هر مشاهده متعلق به خوشه با نزدیک‌ترین میانگین

[1] Expectation maximization
[2] Non-overlapping
[3] MacQueen

است. به‌طورکلی الگوریتم فوق تکرارشونده است و معمولاً پس از چند بار تکرار برای بهبود محلی، همگرا می‌شود .در این خوشه‌بندی، تعداد خوشه‌ها در ابتدا انتخاب می‌شوند و الگوریتم صرفاً هر شئ را به خوشه‌های از پیش تعیین‌شده، اختصاص می‌دهد. الگوریتم با اولین انتخاب تصادفی k موقعیت برای مرکز ثقل خوشه، کار می‌کند. هر مشاهده به خوشه‌ای اختصاص داده می‌شود که نزدیک‌ترین مرکز ثقل را دارد و ثقل با استفاده از مقدار میانگین مقادیر اختصاص داده‌شده، محاسبه می‌شود. سپس الگوریتم این فرآیند را تا زمانی که مرکز جرم خوشه‌ها تغییر نکند یا تغییراتی کمتر از حد آستانه داده‌شده داشته باشند، تکرار می‌کند. بدین ترتیب روش K-Means، عددی، نظارت‌نشده، غیرقطعی و بسیار تکرارشونده است. در زیر مراحل مختلف الگوریتم K-Means آورده شده است(Bradley et al., 1997)

- برای تعریف مرکز k خوشه، با k نقطه انتخابی تصادفی شروع می‌شود
- هر آیتم به نزدیک‌ترین نقطه اختصاص داده می‌شود
- مرکز (میانگین) هر خوشه محاسبه می‌شود
- از k میانگین برای تعریف مرکز k خوشه جدید استفاده می‌کنیم و هر آیتم در خوشه با نزدیک‌ترین مرکز جابه‌جا می‌شود.
- دو مرحله قبل تا زمانی که هیچ تغییری در ماهیت خوشه‌ها بین مراحل وجود نداشته باشد، تکرار می‌شود.

در الگوریتم K-Means همواره k خوشه وجود دارد و همواره حداقل یک مقدار در هر خوشه وجود دارد. خوشه‌ها افرازی هستند و همپوشانی ندارند. هر عضوی از یک خوشه به خوشه خود نزدیک‌تر است تا به خوشه دیگر، چراکه نزدیکی همواره مرکز خوشه‌ها را درگیر نمی‌کند. از مزایای این روش می‌توان به خوشه‌بندی نسبتاً ساده اشاره کرد. زمانی که مقادیر زیاد و تعداد خوشه‌های کوچکی وجود دارند، خوشه‌بندی K-Means ممکن است ازلحاظ محاسباتی سریع‌تر باشد. علاوه بر این خوشه‌بندی K-Means تمایل به تولید خوشه‌های فشرده‌تر دارد به‌خصوص اگر ازلحاظ شکل کروی باشند. با این

معنی که هر خطی که بین دو عضو یا اشیاء خوشه کشیده می‌شود درون مرزهای خوشه باقی می‌ماند. از معایب این الگوریتم نیز اینست که برای بررسی خوشه‌های غیر کروی و خوشه‌ها با اندازه‌ها و تراکم‌های مختلف، مناسب نیست. خوشه‌های کروی معمولاً مرز بسیار پیچیده دارند. این روش خوشه‌بندی اشیاء را به k کلاس انحصاری از قبل تعیین‌شده اختصاص می‌دهد. در حقیقت اشیاء نمی‌توانند به بیش از یک خوشه با درجات مختلفی از احتمال عضویت، تعلق داشته باشند. به‌علاوه، این روش در خوشه‌بندی داده‌هایی که شامل نقاط نویز است، خیلی مؤثر نیست و ازاین‌رو تشخیص نقاط نویز و حذف آن‌ها الـــزامی است.

کاربرد الگوریتم K-Means در تصادفات

(۱) اندرسون[1] در سال ۲۰۰۹ با استفاده از سیستم اطلاعات جغرافیایی[2] (GIS) و تخمین تراکم کرنل[3] به مطالعه الگوهای مکانی تصادفات جراحتی سال ۱۹۹۹ تا ۲۰۰۳ در لندن پرداخت. همچنین با استفاده از داده‌های محیطی تصادف (شرایط جاده، شرایط راننده و...) و نتایج تخمین تراکم کرنل به منظور طبقه بندی نقاط حادثه‌خیز با الگوریتم K-Means خوشه‌بندی نمود. پس از خوشه‌بندی داده‌ها به ۵ گروه و ۱۵ خوشه تقسیم بندی شدند و هر کدام از خوشه‌ها باتوجه به داده‌های محیطی توصیف شدند(اندرسون،۲۰۰۹).

(۲) در تحقیق سینگ چانگ و لی ین[4] (۲۰۱۱) با استفاده از تکنیک مدل‌سازی K-Means به کشف تأثیرات عوامل غیررفتاری که شامل ویژگی‌های هندسی بزرگراه، عوامل ترافیکی که شامل حجم ترافیک در روز و شرایط محیطی که شامل بارش سالیانه می‌باشد، پرداخته شده

[1] Anderson
[2] Geographic Information System
[3] Kernel density estimation
[4] Hsing-chung & Li-yen Chang

است. نتایج پژوهش نشان می‌دهد عوامل فوق که عوامل خطر نامیده می‌شوند، در بروز شدت تصادفات نقش مؤثری دارند.

(۳) طی تحقیقی در سال ۲۰۱۲ توسط رساک و دیوید[1] با استفاده از الگوریتمK-Means به منظور تشخیص نقش عوامل انسانی در بروز و شدت برخوردها و تصادفات جاده‌ای پرداخته شده است. نتایج پژوهش نشان می‌دهد که بستن کمربند ایمنی، داشتن گواهینامه، سن و جنس در بروز تصادفات نقش مؤثری دارند.

خوشه بندی سلسله مراتبی[2]

برعکس خوشه‌بندی افرازی که اشیاء را در گروه‌های مجزا تقسیم می‌کند، خوشه‌بندی سلسله مراتبی، در هر سطح از فاصله، نتیجه خوشه‌بندی را نشان می‌دهد. این سطوح به صورت سلسله مراتبی هستند. خوشه بندی سلسله مراتبی تکنیک یادگیری هدایت‌نشده[3] محبوبی است که به دنبال ایجاد سلسله مراتب درخوشه‌ها است. این خوشه‌بندی به طور معمول به دو دسته زنجیره‌ای[4] و منفصل[5] تقسیم می‌شود. خوشه‌بندی زنجیره‌ای به دنبال یک رویکرد پایین به بالا، یعنی هر شیء داده در خوشه خود شروع می شود و اشیاء بسته بیشتر با هم ادغام می شوند و خوشه جدید را تشکیل می‌دهند. این فرایند تکرار می شود تا خوشه‌ای برای ادغام باقی نگذاشته شود. برخلاف خوشه بندی زنجیره‌ای، خوشه بندی منفصل رویکرد بالا به پایین است که در آن تمام اشیاء داده در یک خوشه شروع می‌شوند و تقسیم در میان اشیاء داده ادامه می یابد تا هر اشیاء داده به یک خوشه واحد تعلق داشته باشند. با این حال، خوشه

[1] Rasak & david
[2] Hierarchical
[3] Unsupervised
[4] Agglomerative
[5] Divisive

بندی زنجیره‌ای از نظر محاسباتی کارآمدتر از الگوریتم خوشه بندی تقسیم شده است(کومار و توشینوال[۱]، ۲۰۱۶)(تان و همکاران[۲]، ۲۰۰۵)

خوشه‌بندی سلسله مراتبی نزدیک‌ترین همسایگی[۳] Nnh

بر اساس این روش، تفاوت بین دو خوشه حداقل فاصله‌ی بین اجسام درون هر خوشه است. فاصله در اینجا به معنی حداقل فاصله هر شیء تا شی $d(x_i, y_j)$ محاسبه می شود که x_i به خوشه اول تعلق دارد، و y_j متعلق به خوشه دوم است. به عبارت دیگر، این روش حداکثر شباهت بین دو گروه از اشیاء یا خوشه را محاسبه می کند. این روش زنجیرهای طولانی را ایجاد می کند که خوشه های ضعیف را تشکیل می دهند. این تکنیک در برخورد با اشکال غیر بیضوی مفید است، اما بسیار حساس به داده‌های نویز و ناخوشه است. فاصله ریاضی، $D(X, Y)$ بین خوشه X و Y به صورت زیر بیان می شود:

$$D(X, Y) = \min d(x, y), \quad x \in X, y \in Y \qquad (۱-۲)$$

که در آن $d(x, y)$ فاصله بین دو نقطه x و y است و X و Y دو خوشه هستند.

روش خوشه‌بندی سلسله مراتبی نزدیک‌ترین همسایگی[۴] (Nnh)، گروهی از رخدادها را که از نظر مکانی نزدیک هستند، شناسایی می کند. این خوشه ها بر مبنای یک معیار برآورده می شوند. خوشه بندی تکرار می شود تا زمانی که تمام نقاط به یک خوشه تقسیم می شوند و یا معیار خوشه ناکام می ماند. روش تحلیل خوشه ای Nnh برای تشخیص نقاط حادثه‌خیز تصادف استفاده می شود که بر شناسایی گروه هایی از داده‌ها که به صورت مکانی نزدیک هستند، متمرکز می‌شود. در خوشه بندی Nnh، فاصله اقلیدسی بین هر جفت نقطه داده بررسی و به عنوان یک معیار برای خوشه بندی استفاده

[1] Kumar & Toshniwal
۲ Tan et al.
[3] Nerast neighbourhood Hierarchical
[4] Nearest Neighborhood Hierarchical

می شود. معیار اول فاصله جستجو (d) است که پس از تعیین آن، نقاط با فاصله‌های کوچکتر در یک خوشه قرار می‌گیرند. معیار دوم حداقل تعداد نقاط n_{min} در خوشه می‌باشد. سپس نقاط منطبق با معیارهای فاصله جستجو (d) و n_{min} به عنوان یک خوشه شناسایی می‌شود. اگر تعداد نقاط تعریف شده نزدیکتر از فاصله جستجو باشد، یک خوشه جدید تولید می شود. پس از محاسبه خوشه های مرتبه اول، خوشه های مرتبه دوم و مرتبه بالا با همان شیوه تشکیل می شوند تا زمانی که فقط یک خوشه باقی بماند یا با معیارهای مذکور تطابق نداشته باشد. به همین دلیل، الگوریتم خوشه بندی Nnh خوشه ای از تمام نقاط در منطقه مورد مطالعه ایجاد نمی کند. از آنجایی که این روش به معیارهای مربوط به فاصله جستجو (d) و n_{min} بستگی دارد، تنها نقاطی که معیارهای مورد نیاز را برآورده می کنند، خوشه بندی می شوند(لوین،۲۰۱۰).

خوشه‌بندی مبتنی بر تراکم[1]

استر[2] و همکاران در سال ۱۹۹۶ برای اولین بار روش خوشه‌بندی مبتنی بر تراکم را پیشنهاد دادند. این روش‌های خوشه‌بندی بر این اصل استوارند که خوشه‌ها، ناحیه‌هایی از فضای داده با چگالی زیادی هستند که توسط نواحی با چگالی کمتر از همدیگر جدا شده‌اند. این الگوریتم‌ها هر شی را با مقدار چگالی تعیین شده توسط تعداد اجسام همسایه خود در یک شعاع داده شده مرتبط می‌کند. یک جسم که تراکم آن بیشتر از آستانه تعریف شده باشد، به عنوان یک شیء متراکم تعریف می‌شود و در ابتدا خوشه ای شکل می‌گیرد. الگوریتم‌های خوشه بندی مبتنی بر چگالی عبارت‌اند از DBSCAN، OPTICS، HOP و DENCLUE. این روش می تواند نویز (خروجی ها) را جدا کند، خوشه های شکل دلخواه را پیدا کند و هیچ فرضیه ای در مورد توزیع داده های پایه ایجاد نمی کند. با این حال، از لحاظ

[1] Density-Based clustering
[2] Ester et al.

محاسباتی بسیار سخت هستند؛ به خصوص در مراحل تولید تراکم داده ها و جستجو برای همسایگان متراکم(استر و همکاران،۱۹۹۶).

الگوریتم DBSCAN

ازجمله الگوریتم‌های مبتنی بر تراکم، DBSCAN است. این روش خوشه‌بندی بر این اصل استوار است که خوشه‌ها، ناحیه‌هایی از دو فضای داده با تراکم زیاد هستند که توسط نواحی با تراکم کمتر از همدیگر جداشده‌اند. مفهوم اصلی، تعریف خوشه جدید یا توسعه خوشه موجود بر اساس همسایگی (Eps) است. پارامترEps ، بیشترین شعاع همسایگی و پارامترMinPts ، کمترین تعداد نقاط در شعاع همسایگی نقطه هستند. با توجه به مجموعه داده‌یD ، تابع توزیع فاصله و پارامترهای Eps و MinPts، تعاریف زیر برای تعریف DBSCAN استفاده می‌شود:

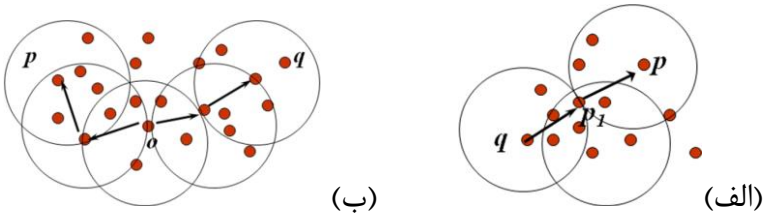
- هسته اصلی: در همسایگی شعاع ε حاوی حداقل اجزای minpts باشد.
- نقطه مرزی: همسایگی یک جسم خاص کمتر از minpts باشد
- نقاط نویزی: همسایگی یک جسم خاص شامل ۰ است.
- تراکم قابل دسترس مستقیم: اگر شی p یک شی هسته‌ای است، آنگاه شی q در همسایگی p قرار دارد، سپس q و p را تراکم قابل دسترس مستقیم می‌نامند.
- تراکم قابل دسترس: پیوند داده‌ای بین $p_1, p_2, \ldots, p_n$ وجود دارد، جایی که$p_1 = p_2$،$p_n = q$، اگر شی p_i و شی p_{i+1} ، چگالی قابل دسترس مستقیم باشد ،سپس ما q و p را تراکم قابل دسترس می نامیم (شکل ۲-۳ الف).
- اتصال تراکم: اگر جسم o در مجموعه داده وجود داشته باشد که جسم p و q را تراکم قابل‌دسترس برای o می‌سازد، سپس شیء p و q تراکم قابل‌دسترس است (شکل۲-۳ ب)

- فاصله هسته: حداقل شعاع همسایگی که یک شی را به یک هسته اصلی به نام فاصله هسته تبدیل می‌کند. اگر شی تبدیل هسته اصلی نیست، هیچ فاصله هسته‌ای نیز وجود ندارد.
- فاصله قابل دسترس: فاصله اقلیدس از شی q به هسته p به عنوان فاصله قابل دسترس نامیده می شود، اگر شی p یک هسته اصلی نیست، فاصله ای بین p و q وجود ندارد.

نقطه دلخواه $p(p \in D)$، همسایگی p به صورت رابطه ۲-۲ تعریف می‌شود:

$$N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\} \qquad (۲-۲)$$

اگر $|N_{Eps}(p)| \geq MinPts$ باشد.



شکل۲- ۳: (الف) تراکم قابل دسترس q و p (ب) اتصال تراکم p و q

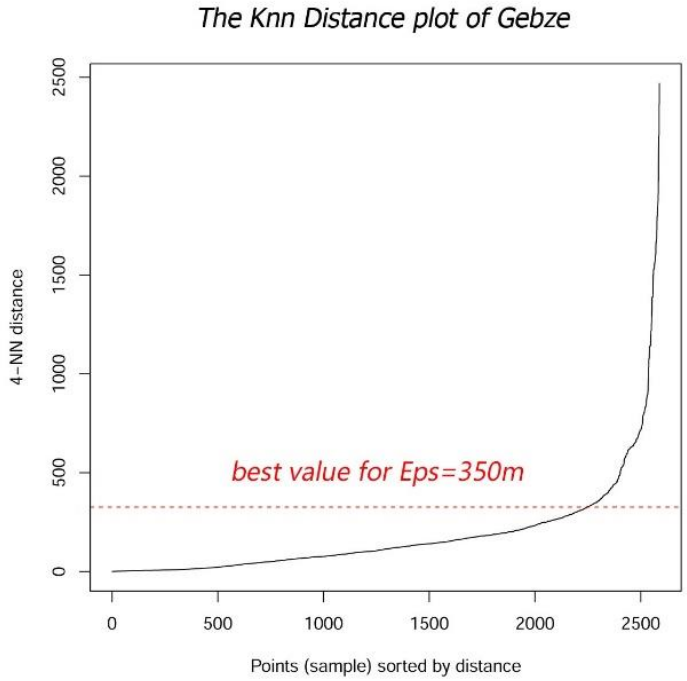
اگر p نقطه مرکزی و q همسایه p باشد؛ q متعلق به این خوشه است. همه همسایه‌های q برای اضافه شدن به خوشه بررسی می‌شوند و در غیراینصورت q با عنوان نویز برچسب دهی می‌شود. این روند برای هر نقطه در همسایگی تکرار می‌شود. اگر خوشه‌ای نتواند گسترش یابد، DBSCAN نقطه لخواه بدون برچسب دیگری را انتخاب کرده، فرایند تکرار می‌شود. به‌عبارت‌دیگر الگوریتم DBSCAN به این صورت عمل می‌کند که ابتدا نقطه دلخواه x انتخاب می‌شود سپس همه نقاطی که از سوی x در دسترس چگالی هستند جستجو می‌شوند؛ اگر x یک نقطه هسته باشد خوشه تشکیل می‌شود و اگر x یک نقطه مرزی باشد هیچ نقطه‌ای از طرف x در دسترس چگالی نخواهد بود؛ این مراحل برای نقاط بعدی تا زمانی

که همه نقاط مجموعه داده در خوشه‌ها قرار گیرند یا به‌عنوان نویز برچسب دهی شوند، ادامه می‌یابد. به‌عبارت‌دیگر روش‌های خوشه‌بندی مبتنی بر چگالی تلاش می‌کنند خوشه‌هایی را بر اساس تراکم نقاط در منطقه پیدا کنند. مناطق متراکمی که از یکدیگر قابل‌دسترس هستند برای تشکیل خوشه باهم ادغام می‌شوند. در این روش خوشه‌بندی، هر داده متعلق به یک خوشه، در دسترس چگالی سایر داده‌های متعلق به آن خوشه است و در دسترس چگالی هیچ داده دیگری قرار ندارد .روش‌های خوشه‌بندی مبتنی بر تراکم خوشه‌ها را با اشکال دلخواه پیدا می‌کنند(استر و همکاران[1]،۱۹۹۶).

شاخص البو برای تعیین مقدار شعاع *Eps*

تحقیقات استر و همکارانش[2] در ۱۹۹۶ حاکی از آن بود که روش البو برای تعیین پارامترهای بهینه خوشه‌ها در خوشه‌بندی DBSCAN استفاده می‌شود. این روش بر اساس k نزدیک‌ترین همسایگی، مقدار MinPts کمترین تعداد نقاط در شعاع همسایگی نقطه و Eps را محاسبه می‌کند. پارامتر MinPts برای تشخیص داده‌های اصلی مورد استفاده قرار می‌گیرد. داده اصلی، داده‌ای است که در شعاع همسایگی آن حداقل به تعداد MinPts داده وجود داشته باشد.

[1] Ester et al.
[2] Ester et al.



شکل۲- ۴: تعیین Eps بهینه با استفاده از نمودار Knn-distance[1]

شکل ۲-۴ روش البو را نشان می‌دهد که در آن محور افقی تعداد نمونه‌های مرتب شده بر اساس K نقطه همسایگی است و محور عمودی، فاصله K نزدیک‌ترین همسایگی است. بر اساس تحقیقات استر در سال ۱۹۹۶ میلادی برای داده‌های دوبعدی (x,y) می‌توان k را برابر با ۴ در نظر گرفت؛ بنابراین ۴ همسایگی حساب می‌شود و فاصله از تمامی نقاط دیگر به دست می‌آید. درواقع ۴ نقطه‌ای که نزدیک به نقطه مرکز هستند، بررسی می‌شوند. بیشترین فاصله تا ۴ نقطه حساب می‌شود و این کار برای همه نقاط مجموعه داده انجام می‌شود و نموداری رسم می‌شود که در شکل ۲-۴ نشان داده‌شده است. در نمودار به‌دست‌آمده نقطه‌ای که تغییراتش شدید می‌شود (نقطه عطف)، نقطه موردنظر است که با عمود کردن نقطه عطف به محور عمودی پارامتر Eps دست می‌آید. برای پارامتر Minpts نیز حداقل مقدار در این روش عدد K و بیشتر می‌باشد و به صورت سعی و خطا باید مقداری را در نظر گرفت که کمترین مقدار نویز با حفظ شاخص‌های خوشه بندی بدست آید.

[1] K nearest neighbourhood distance

مزایا و معایب روشDBSCAN

این الگوریتم نیز همانند دیگر الگوریتم‌ها دارای مزایا و معایبی می‌باشد. از مزایای این الگوریتم اینست که نیازی به مشخص بودن تعداد خوشه‌ها، به‌صورت اطلاع قبلی ندارد (برخلاف روش‌های خوشه-بندی مبتنی بر مرکز مانند مثل K-Means) می‌تواند خوشه‌های با اشکال مختلف را بیابد. همچنین می‌تواند خوشه‌ای را که کاملاً توسط خوشه‌ی دیگر احاطه‌شده است (ولی به آن متصل نیست)، شناسایی کند. مشکل تک‌یالی (خوشه‌های مختلف با مسیر نازکی از نقاط به هم متصل باشند) نیز با پارامتر MinPts کاهش‌یافته است. از مفهوم نویز پشتیبانی می‌کند. به دو پارامتر نیاز دارد Eps و MinPts و نسبت به ترتیب قرار گرفتن نقاط در پایگاه داده حساس نیست. (البته در مورد برخی از نقاط مرزی، این امر صادق نیست. نقاطی که در مرز دو خوشه‌ی متفاوت باشند، با توجه به ترتیبِ در نظر گرفتن نقاط در خوشه‌بندی، ممکن است در خوشه‌ها جابه‌جا شوند).

از معایب این الگوریتم این است که نمی‌تواند مجموعه داده‌هایی با اختلاف چگالی زیاد را خوشه-بندی کند، به دلیل اینکه مقادیر MinPts و Eps نمی‌توانند برای تمام خوشه‌ها مناسب باشند (برای رفع این مشکل، توسعه DBSCAN-DD برای آن ارائه‌شده است). برای نمونه، در نواحی متراکم، همسایگی نقاط مختلف اشتراک زیادی باهم دارند. درنتیجه نقاط تکرارِی زیادی را در محاسبات خود لحاظ می‌کند و باعث افت کارایی این الگوریتم می‌شود برای رفع این مشکل توسعه‌یIDBSCAN برای آن ارائه‌شده است.

کیفیت DBSCAN به نوع اندازه‌گیری فاصله‌ی نقاط بستگی دارد. فاصله‌ی اقلیدسی رایج‌ترین نوعی است که استفاده می‌شود. در داده‌های دارای ابعاد بالا، این نوع اندازه‌گیری فاصله بی‌فایده می‌شود. درنتیجه یافتن مقداری برای Eps دشوار می‌شود. در مقایسه با روش‌های سنتی تشخیص تراکم، روش خوشه‌بندی مبتنی بر تراکم به‌طور ذاتی می‌تواند تراکم مکانی را کشف کند. تقسیم‌بندی مجموعه داده

در آغاز این فرایند موردنیاز نیست. همچنین می‌تواند اشکال با تراکم دلخواه را از مجموعه داده پیدا کند. به‌هرحال بیشتر روش‌های خوشه‌بندی مبتنی بر تراکم، محدودیت DBSCAN اصلی را به ارث می‌برند که ویژگی‌های غیرمکانی را به‌حساب نمی‌آورند و نمی‌توانند به‌طور مستقیم برای مجموعه داده تصادف روی شبکه مکانی به کار گرفته شوند(موریرا و همکاران[1]، ۲۰۰۵).

- **کاربرد الگوریتم** DBSCAN **در تصادفات**

(۱) دوئرو و سوباشی در سال ۲۰۱۵ در مقاله‌ای چگونگی استفاده از تکنیک های داده کاوی در ITS برای تشخیص و پیش‌گیری تصادفات در بزرگراه‌ها را نشان‌دادند. در زمان ترافیک، خودروها رفتار مشابهی با خودروهای واقع در نواحی محله‌ای بسته را نشان می‌دهند. در مواردی مانند تصادفات، تغییر خط نامناسب و سرعت می‌تواند به عنوان ناهنجاری‌ها تشخیص داده شود. آن‌ها یک تصادف رانندگی شبیه‌سازی کرده و اثربخشی تکنیک‌های خوشه‌بندی متفاوت برای تشخیص تصادف رانندگی موردبررسی قرار دادند. آن‌ها نشان دادند که با دو الگوریتم DBSCAN و خوشه بندی سلسله مراتبی زنجیزه‌ای AHC[2]، اگر مقدار سرعت و موقعیت هر وسیله نقلیه داده شود، رفتار وسایل نقلیه را می توان تجزیه و تحلیل کرد و حوادث را می توان به راحتی شناسایی کرد. موفقیت الگوریتم‌های پیشنهادی با استفاده از شبیه‌سازی در یک سناریو بزرگراه نشان‌داده شده است. شبیه‌سازی نشان داد که ابزارهای داده‌کاوی به طور موفقیت‌آمیزی تصادفات را با نرخ تشخیص متوسط خطر ۱۰۰٪ و میزان اشتباه ۰٪ با استفاده از DBSCAN و خوشه بندی سلسله مراتبی نشان داد.

(۲) العتیبی[3](۲۰۱۸) با استفاده از الگوریتم‌های داده کاوی DBSCAN و فرکانس موازی[4] به بررسی پارامترهای موثر بر تصادفات پرداخت. ابتدا مناطق تصادف را به تعداد K

[1] Moreira et al.
[2] Agglomerative hierarchical clustering
[3] Alotaibi
[4] Parallel Frequent

خوشه بر اساس فرکانس تصادف آنها با الگوریتم DBSCAN تقسیم کرد. سپس الگوریتم فرکانس موازی را روی این خوشه ها اعمال کرد، تا ارتباط بین ویژگی‌های متمایز در داده‌های تصادفات رانندگی را برای تحلیل آن‌ها در جهت پیش‌بینی عوامل مختلف که بر تصادفات جاده‌ای در مکان‌های مختلف تأثیر می‌گذارد، نمایان گردد. هدف اصلی برداشت داده‌های تصادف، شناسایی مسائل کلیدی در زمینه ایمنی جاده‌ها است. داده‌های تصادف جاده‌ای با استفاده از ابزار Weka تحلیل شدند و نتایج نشان داد با استفاده از الگوریتم داده‌کاوی DBSCAN و PF نشان می دهد که این روش می تواند در داده های تصادفات جدید با ویژگی های اضافی برای شناسایی عوامل مختلف مرتبط با تصادفات جاده‌ای استفاده شود.

الگوریتم‌های مبتنی بر شبکه (*GRID*)

الگوریتم‌های خوشه بندی مبتنی بر شبکه ابتدا دامنه فضای مسئله را با شبکه مش‌بندی یکنواخت پوشش می دهد. صفات آماری برای تمام اشیاء داده شده در هر سلول شبکه مش جمع آوری شده و سپس خوشه بندی در شبکه به جای اشیاء داده ها انجام می شود. این الگوریتم‌ها به طور معمول زمان پردازش سریع‌تری نسبت به بقیه الگوریتم‌ها دارند، زیرا از یک «مجموعه داده» برای محاسبه مقادیر آماری شبکه استفاده می کنند و عملکرد خوشه بندی فقط به اندازه شبکه ها بستگی دارد که معمولا بسیار کمتر از اشیاء داده است. الگوریتم خوشه بندی مبتنی بر شبکه بر پایه STING، WaveCluster و CLIQUE است. تمام این روشها یک شبکه یکپارچه را برای پوشش دادن کل مسئله استفاده می‌کنند. برای داده‌های با توزیع بسیار نامنظم، ابعاد شبکه مش باید به اندازه کافی برای به دست آوردن کیفیت خوشه نرمتر تنظیم شود یعنی تعداد شبکه بیشتری در نظر گرفته شود. یک مش نرم‌تر[1] می تواند موجب

[1] Finer mesh

اندازه مش نزدیک یا حتی بزرگ‌تر از اندازه اشیاء داده برسد که می تواند بار محاسباتی را برای خوشه سازی افزایش دهد(مهران و مهار[1]،۲۰۰۸).

رویکردهای مبتنی بر تراکم و شبکه برای خوشه‌بندی در یک فضای چندبعدی هستند که در آن خوشه‌ها به‌عنوان مناطق متراکم‌تر از محیط اطراف آن‌ها محسوب می‌شوند. پیچیدگی محاسباتی اکثر الگوریتم‌های خوشه‌بندی حداقل به‌صورت خطی متناسب با اندازه مجموعه داده است. مزیت مهم خوشه‌بندی مبتنی بر شبکه کاهش قابل‌توجهی در پیچیدگی محاسباتی است، به‌ویژه برای خوشه‌بندی مجموعه داده‌های بسیار بزرگ است. رویکرد خوشه‌بندی مبتنی بر شبکه متفاوت از الگوریتم‌های خوشه‌بندی متعارف است که در آن نه‌تنها با نقاط داده بلکه با ارزش فضای اطراف نقاط داده را احاطه می‌کند. به‌طورکلی، یک الگوریتم خوشه‌بندی مبتنی بر شبکه معمولی شامل پنج مرحله پایه است:(گرباستز[2] و بوریسوف[3]،۲۰۰۲)

۱. ایجاد ساختار شبکه، یعنی تقسیم فضای داده به تعداد محدودی از سلول‌ها.
۲. محاسبه تراکم سلول برای هر سلول.
۳. مرتب‌سازی سلول‌ها با توجه به تراکم آن‌ها.
۴. شناسایی مراکز خوشه‌ای
۵. پیروی از سلول‌های همسایه.

- **تقسیم مجموعه داده بر اساس شبکه**

تعریف سلول شبکه در این پایان نامه به این صورت است که دو برابر شعاع همسایگی را به عنوان طول سلول شبکه انتخاب شده‌است(Mahran and Mahar, 2008). سپس مجموعه داده ها را در هر ابعاد با طول 2ε یکسان تقسیم می‌شود. بنابراین، فضای داده به سلول های شبکه های مختلف تقسیم

[1] Mahran & Mahar
[2] Grabusts
[3] Borisov

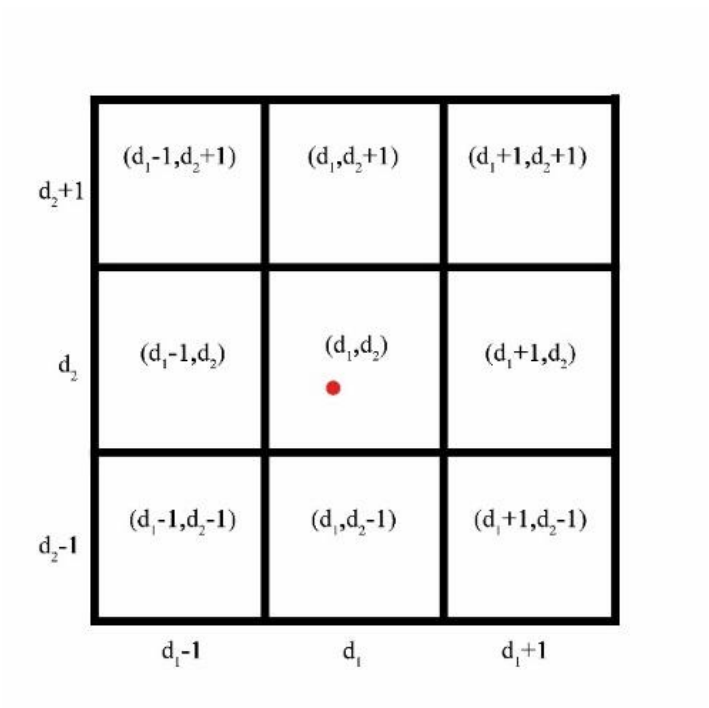
شده و با $d_1 ، d_2 \dots d_n$ مشخص می‌شود، که n ابعاد داده است، d_1 تعداد توالی سلول شبکه در ابعاد مختلف$1 \leq i \leq n$ را نشان می دهد(Yi-ling et al., 2008).

تعریف سلول شبکه p متعلق: $x_1 ، x_2 ، \dots x_n$ مختصات n از شیء p را در هر ابعاد نشان می دهد، شماره d_1 شبکه i ام اید دارای شرایط زیر باشد:

$$x_i/2\varepsilon \leq d_i \leq (x_i + 2\varepsilon)/2\varepsilon \qquad (۳-۲)$$

d_iیک عدد صحیح است.

ساختار الگوریتم شبکه فقط نیاز به عبور از ۹ سلول شبکه را دارد که در هنگام محاسبه Eps همسایگی اشیاء پس از سلول شبکه تنظیم شده است. داده های دو بعدی را به عنوان مثال در نظر بگیرید، جسم p (x ،y) متعلق به شبکه (d2 ،d1)، دامنه عبور در شکل ۲-۵ نشان داده شده است:

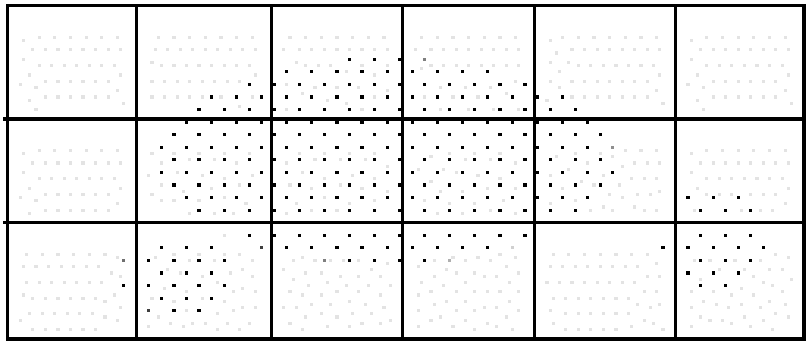


شکل۲- ۵:شبکه همسایگی در فضای دو بعدی

ساختار ترکیبی الگوریتم GriDBSCAN

الگوریتم پیشنهادی بر اساس تکنیک تقسیم کن و غلبه کن[1] است. ما یک شبکه را ایجاد می کنیم که فضای اطراف آن را پارتیشن بندی می کند و به همین ترتیب داده ها به سلول‌ها منتقل می شود. هر پارتیشن به طور جداگانه DBSCAN است و سپس خوشه های حاصل از همه پارتیشن‌ها برای ادغام خوشه بندی داده ها به یکدیگر متصل می‌شوند. این روند باید همان نتیجه‌ای را تولید کند که اگر DBSCAN در کل مجموعه داده اعمال می‌شد.

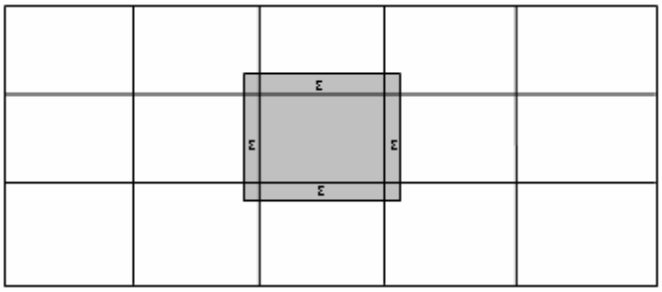
با استفاده از صفحات هم‌محور و هم‌فاصله با ابعاد "d-1" با فرض ابعاد داده‌ای "d" (شکل ۲-۶)، یک شبکه ساخته می شود. تعداد سلول‌ها به انتخاب کاربر است، اما نقطه‌ای بودن داده‌ها تأثیر زیادی بر عملکرد الگوریتم دارد. جزئیات دقیق تر شبکه بندی عملکرد آن را کاهش می دهد بنابراین ما پیشنهاد می کنیم که عرض سلول باید در هر ابعادی(ε ۲ >) باشد(مهران و مهار،۲۰۰۸).



شکل۲- ۶: پارتیشن بندی فضایی با استفاده از شبکه

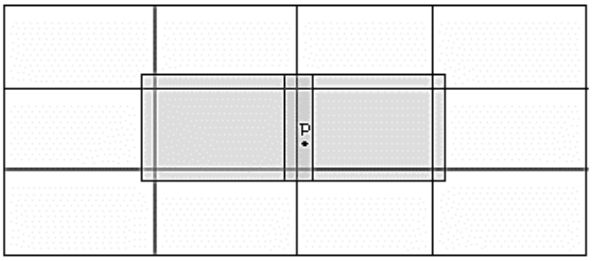
هر پارتیشن شامل تمام نقاط در داخل سلول (نقاط داخلی) و نقاط اطراف سلول در محوطه سلول (نقاط بیرونی) است (شکل ۲-۷).

[1] Divide-And-Conquer



شکل۲- ۷: پیرامون سلولε

این به این معنی است که بعضی از نقاط در پارتیشن های مختلف گنجانده می شوند، اما یک نقطه تنها در یک پارتیشن درونی خواهد بود و نقطه دیگر در پارتیشن های دیگر نیز خواهد بود (شکل ۲-۸). به این ترتیب ما مطمئن هستیم که تمام پرسش‌های محدوده برای نقاط داخلی، دقیق هستند، زیرا دایره ε-شعاع در اطراف یک نقطه داخلی، به طور کامل در پارتیشن قرار می گیرد، هرچند از معیار فاصله استفاده می‌شود، به طوری که تمام نقاط هسته به درستی برای نقاط داخلی تعیین می شود. محدوده درخواست‌های بیرونی دقیق نیست زیرا حلقه‌های ε-شعاعی اطراف آنها ممکن است قطع شوند.

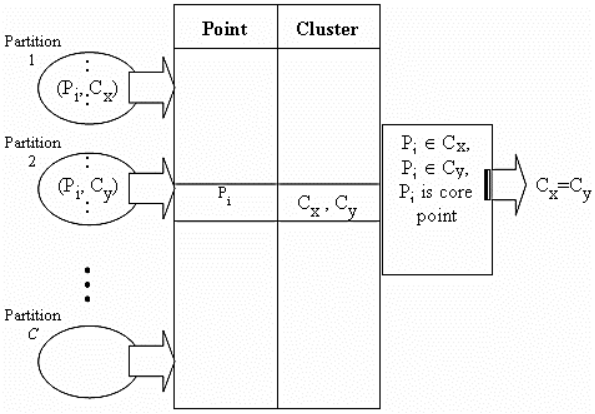


شکل۲- ۸: نقطه p درون سلول راست قرار دارد، اما در سلول چپ نیز شامل می شود

DBSCAN در هر بخش از داده ها به طور جداگانه با توجه به تمایز نقاط اصلی و نقاط مرزی اعمال می شود و مجموعه ای متمایز از خوشه-ID برای هر پارتیشن را رزرو می کند.

خوشه بندی حاصل از همه پارتیشن ها به صورت پیوسته برای ساخت یک مجموعه شامل تمام نقاط و شناسه خوشه خود استفاده می شود. از آنجا که یک نقطه می تواند به چندین پارتیشن تعلق

دارد، ممکن است به خوشه های چندگانه اختصاص داده شود، در این صورت تمام این خوشه ها باید تحت شرایطی که هر نقطه نقطه هسته است (شکل ۲-۹) ادغام شود. به عبارت دیگر یک نقطه اصلی برای جمع کردن تمام خوشه هایی که به آن اختصاص داده شده کافی است.



شکل۲- ۹: روند ادغام دو روش Grid و DBSCAN

۲-۳-۴- شاخص‌های ارزیابی خوشه‌بندی[1]

نتایج حاصل از اعمال الگوریتم‌های خوشه‌بندی روی یک مجموعه داده با توجه به انتخاب‌های پارامترهای الگوریتم‌ها می‌توانند بسیار متفاوت از یکدیگر باشند. به عبارتی تخمین تعداد بهینه خوشه‌ها در دسته‌بندی بدون‌نظارت[2] داده‌ها، از زمینه‌های چالش‌برانگیز برای محققان در سال‌های اخیر بوده که منجر به ارائه شاخص‌های اعتبار خوشه‌بندی متعددی شده است. در این تحقیق شاخص‌های اعتبارسنجی رایج برای محاسبه تعداد بهینه خوشه‌ها برای ورود به الگوریتم‌های خوشه‌بندی بدون نظارت بررسی می‌شوند. هدف از اعتبارسنجی خوشه‌ها یافتن خوشه‌هایی است؛ که بهترین تناسب را با داده‌های موردنظر داشته باشند(کواچز و همکاران[3]،۲۰۰۵).

[1] Cluster Evaluation
[2] Unsupervised
[3] Kovács et al.

روش‌های ارزیابی خوشه‌های حاصل از خوشه‌بندی به سه دسته کلی تقسیم می‌شوند: معیارهای نسبی، معیارهای درونی، معیارهای خارجی. هم معیارهای خارجی و هم معیارهای درونی بر مبنای روش‌های آماری عمل می‌کنند و پیچیدگی محاسباتی بالایی را نیز دارا هستند. معیارهای خارجی عمل ارزیابی خوشه‌ها را با استفاده از بینش خاص کاربران انجام می‌دهند. معیارهای درونی عمل ارزیابی خوشه‌ها را با استفاده از مقادیری که از خوشه‌ها و نمای آن‌ها محاسبه می‌شود، انجام می‌دهند .پایه معیارهای نسبی، مقایسه بین خوشه‌بندی (الگوریتم به‌علاوه پارامترهای آن) مختلف است. یک و یا چندین روش مختلف خوشه‌بندی چندین بار با پارامترهای مختلف روی یک مجموعه داده اجرا می‌شوند و بهترین خوشه‌بندی از بین تمام خوشه‌ها انتخاب می‌شود. در این روش مبنای مقایسه، شاخص‌های اعتبارسنجی هستند؛ که این تحقیق در دسته معیارهای نسبی قرار دارد(کواچز و همکاران،۲۰۰۵).

شاخص‌های اعتبارسنجی برای سنجش میزان صحت نتایج خوشه‌بندی به‌منظور مقایسه بین روش‌های خوشه‌بندی مختلف یا مقایسه نتایج حاصل از یک روش با پارامترهای مختلف مورداستفاده قرار می‌گیرند. شاخص‌های ارزیابی نسبی بسیار متنوعی پیشنهاد شده‌اند؛ ازجمله، شاخص F-measure، شاخص precision و recall، شاخص Rand و AIR، شاخص MF که در این بخش به تشریح این شاخص ها می‌پردازیم.

شاخص رَند [1]

برای نشان دادن میزان شباهت بین دو شیوه برچسب‌گذاری می‌توان از شاخص رَند استفاده کرد. این شاخص توسط دانشمند آمار ویلیام رند (William Rand) در سال ۱۹۷۱ در مقاله‌ای با عنوان معیارهای هدف برای ارزیابی روش‌های خوشه‌بندی[2] معرفی شد(رند،۱۹۷۱). برای محاسبه آن باید دو پارامتر را اندازه‌گیری کنیم. A تعداد زوج‌هایی که هم در خوشه‌ها و هم در دسته‌ها در کنار هم هستند.

[1] Rand Index
[2] Objective criteria for the evaluation of clustering methods

به بیان دیگر هم در خوشه‌ها دارای برچسب یکسانی هستند و هم برچسب دسته‌ها برای آن‌ها یکسان است. B تعداد زوج‌هایی که هم در خوشه‌ها و هم در دسته‌ها از یکدیگر جدا هستند. یعنی برچسب خوشه‌هایشان متفاوت است و البته برچسب دسته‌های متفاوتی نیز دارند.حال برای محاسبه شاخص رند کافی است که حاصل جمع A و B را به تعداد کل زوج‌ها تقسیم کنیم.

$$\mathrm{Rand(S,C)} = \frac{\mathrm{A+B}}{\binom{N}{2}} = \frac{\mathrm{A+B}}{\frac{\mathrm{N(N-1)}}{2}} \qquad (۴-۲)$$

اگر خوشه‌ها مطابق با دسته‌ها ایجاد شده باشند، شاخص رند برابر با ۱ خواهد بود. ولی اگر خوشه‌بندی به صورت تصادفی ایجاد شده باشد، دلیلی ندارد که مقدار این شاخص برابر با صفر باشد. برای رفع این مشکل از شاخص رند اصلاح شده[1] استفاده می‌شود. این شاخص را به صورت ARI نشان داده و به شکل زیر محاسبه می‌شود:

$$\mathrm{ARI(S,C)} = \frac{\mathrm{Rand(S,C)-E(Rand(S,C))}}{\mathrm{max(Rand(S,C))-E(Rand(S,C))}} \qquad (۵-۲)$$

منظور از E نیز امید-ریاضی شاخص رند است.

خصوصیات شاخص رند اصلاح شده

حدود مقدارها ARI :مقداری بین ۱ و ۱- خواهد بود. در حالتی که ARI=1 باشد، مطابقت کامل بین برچسب‌های واقعی و خوشه‌ای وجود دارد و در مقابل اگر مقدار این شاخص برابر با ۱- باشد نشانگر برچسب‌گذاری تصادفی در حین خوشه‌بندی است.

[1] Adjusted Rand Index

- **شاخص کارا برای مقایسه چندین روش** :از آنجایی که این شاخص به توافق بین برچسب‌ها تکیه دارد، می‌توان از آن برای مقایسه دو روش خوشه‌بندی نیز استفاده کرد. برای مثال می‌توان مطابقت بین شیوه خوشه‌بندی و برچسب‌گذاری در الگوریتم K-Means را با روش فازی-C Meansبررسی کرد.
- **بدون وابستگی به تعداد خوشه‌ها** :با توجه به شیوه محاسبه این شاخص، تفاوت بین تعداد خوشه‌ها و دسته‌ها نیز در آن لحاظ شده است.
- **تقارن در ARI**: اگر جای دسته‌ها و خوشه‌ها عوض شود، شاخص ARI تعییری نمی‌کند. به این معنی که $ARI(S,C) = ARI(C,S)$.
- **عدم حساسیت به تغییر برچسب‌ها** :اگر برچسب‌های خوشه‌ها تغییر کند در نتیجه ARI تغییری بوجود نمی‌آید. برای مثال اگر همه برچسب‌های ۱ در خوشه‌بندی به برچسب ۲ تبدیل شوند، حاصل ARI تغییری نخواهد کرد.

فرض کنید در اینجا n_{ij} تعداد عناصر مشترک در دسته i و خوشه j باشد. همچنین a_i نیز مجموع مقدارهای n_{ij} برای سطر i ام و b_j نیز جمع مقدارهای n_{ij} در ستون j ام باشد. جدول متقاطع زیر تعداد عناصر مشترک در بین دسته‌ها و خوشه‌ها را نشان می‌دهد.

$C \backslash S$	S_1	S_2	...	S_l	Sums
C_1	n_{11}	n_{12}	...	n_1	a_1
C_2	n_{21}	n_{22}	...	n_2	a_2
⋮	⋮	⋮	⋱	⋮	⋮
C_k	n_{k1}	n_{k2}	...	n_{kl}	a_k
Sums	b_1	b_2	...	b_l	

حال براساس محاسبه زیر می‌توان شاخص رند اصلاح شده را بدست آورد:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}} \qquad (۶-۲)$$

شاخص فولکز- مالوز[1]

برای استفاده از این شاخص باید در مورد چند اصطلاح، آگاهی داشته باشید:

- **مثبت صحیح** :اگر زوجی از مشاهدات که در یک دسته هستند، در یک خوشه نیز قرار بگیرند، نتیجه خوشه‌بندی را برای این زوج مثبت صحیح[2] می‌نامیم. مثبت صحیح را با TP نیز نشان می‌دهند.

- **منفی صحیح** :اگر زوجی از مشاهدات که در دو دسته مجزا قرار دارند، در دو خوشه مجزا نیز جای گیرند، نتیجه خوشه‌بندی را برای این زوج منفی صحیح[3] می‌نامیم. منفی صحیح را باTN نیز نشان می‌دهند.

- **مثبت کاذب** :اگر زوجی از مشاهدات که در دو دسته مجزا قرار دارند، در یک خوشه جای گیرند، نتیجه خوشه‌بندی برای این زوج مثبت کاذب[4] است. مثبت کاذب را با FP نیز نشان می‌دهند.

- **منفی کاذب** :اگر زوجی از مشاهدات که در یک دسته قرار دارند به اشتباه در دو خوشه قرار گیرند، نتیجه خوشه‌بندی برای این زوج منفی کاذب[5] است. منفی کاذب را با FN نیز نشان می‌دهند.

[1] Fowlkes–Mallows Index
[2] True Positive
[3] True Negative
[4] False Positive
[5] False Negative

جدول زیر که وظیفه مقایسه بین تعداد نتایج خوشه‌بندی توسط الگوریتم و دسته‌بندی اصلی را به عهده دارد، به ماتریس درهم‌ریختگی[1] شهرت دارد که توجه به آن به درک و شیوه محاسبه شاخص فولکز- مالوز و اصطلاحات گفته شده، کمک می‌کند.

جدول۲- ۲: تعاریف زوج های مثبت و منفی

نتیجه خوشه‌بندی برای زوج (x,y)(x،y)		دسته‌بندی واقعی	
		در یک دسته	در دو دسته مجزا
خوشه‌بندی	در یک خوشه	TP	FP
	در دو خوشه مجزا	FN	TN

حال اگر تعداد زوج‌های مثبت صحیح را با|TP|، تعداد مثبت کاذب را با |FP| و تعداد منفی کاذب را با |FN| نشان بدهیم، شاخص فولکز- مالوز که با FMI نشان داده می‌شود به صورت زیر محاسبه می‌شود:

$$\mathrm{FM} = \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}} \qquad (۷-۲)$$

شاخص فولکز- مالوز یک روش ارزیابی خارجی است که برای تعیین شباهت بین دو خوشه بندی (خوشه های به دست آمده پس از الگوریتم خوشه بندی) استفاده می شود. این اندازه گیری شباهت می تواند بین دو خوشه بندی سلسله مراتبی یا دسته بندی و طبقه بندی معیار باشد. مقادیر بیشتری برای شاخص فولکز- مالوز نشان می دهد که شباهت بیشتری بین خوشه ها و طبقه بندی های معیار وجود

[1] Confusion Matrix

دارد. شاخص فولکز- مالوز، هنگامی که نتایج حاصل از دو الگوریتم خوشه بندی برای ارزیابی نتایج استفاده می شود، به صورت زیر تعریف می شود:(Fowlkes and Mallows, 1983)

که TP تعداد مثبت واقعی است، FP تعداد مثبت کاذب است و FN تعداد منفی کاذب است. از آنجا که این شاخص مستقیما با تعداد مثبت های واقعی متناسب است، یک شاخص بالاتر، به معنی همبستگی بیشتر بین دو خوشه بندی مورد استفاده برای تعیین شاخص استفاده شده است. یکی از روش های اصلی برای بررسی اعتبار این شاخص، مقایسه دو دسته بندی است که به یکدیگر وابسته نیستند. فولکز و مالوز نشان دادند که برای دو خوشه بندی غیر مرتبط، ارزش این شاخص زمانی که تعدادی داده های مورد نظر برای خوشه بندی افزایش می‌یابد به صفر می رسد؛ در حالی که ارزش شاخص رند برای داده های مشابه به سرعت به ۱ تبدیل می شود؛ بنابراین فولکز- مالوز شاخص بسیار دقیق‌تری برای داده‌های غیر مرتبط است. این شاخص همچنین زمانی که نویز به یک مجموعه داده اضافه شود به خوبی عمل می کند. فولکز و مالوز نشان داد که مقدار شاخص با افزایش مولفه نویز کاهش می یابد. این شاخص همچنین حتی زمانی که مجموعه داده های نویز دارای تعداد خوشه ای متفاوت از خوشه های مجموعه داده اصلی بودند، شباهت را نشان داد؛ بنابراین این شاخص را یک ابزار قابل اعتماد برای اندازه گیری شباهت بین دو خوشه می‌سازد.

شاخص اندازه F [1]

همانطور که دیده شد، استفاده از مفاهیم مثبت کاذب، مثبت صحیح، منفی کاذب و منفی صحیح در محاسبه شاخص ارزیابی نتایج خوشه‌بندی اهمیت زیادی دارد. با توجه به این مفاهیم می‌توان شاخص رند را به صورت زیر نیز محاسبه کرد:

[1] F-measure Index

$$\mathrm{Rand}(S,C)=\frac{|TP|+|TN|}{|TP|+|FP|+|TN|+|FN|} \qquad (۸-۲)$$

در این حالت وزن یا اهمیت مربوط به منفی کاذب (FN) یا مثبت کاذب (FP) در محاسبه شاخص یکسان است. ولی در اندازهF می‌توان وزن آن‌ها را با پارامتر β تغییر داد. اگر P و R را به صورت زیر تعریف کنیم:

$$P=\frac{|TP|}{|TP|+|FP|}, R=\frac{|TP|}{|TP|+|FN|} \qquad (۹-۲)$$

آنگاه شاخص F را به ازاء $\beta \neq 0$به شکل زیر محاسبه می‌شود:

$$F_\beta=\frac{(\beta^2+1)PR}{\beta^2P+R} \qquad (۱۰-۲)$$

در تجزیه و تحلیل آماری طبقه بندی دوتایی، نمرهF_1 (همچنین F-score یا F-measure) یک نوع اندازه گیری از دقت آزمون است(ساسکی[1]،۲۰۰۷).

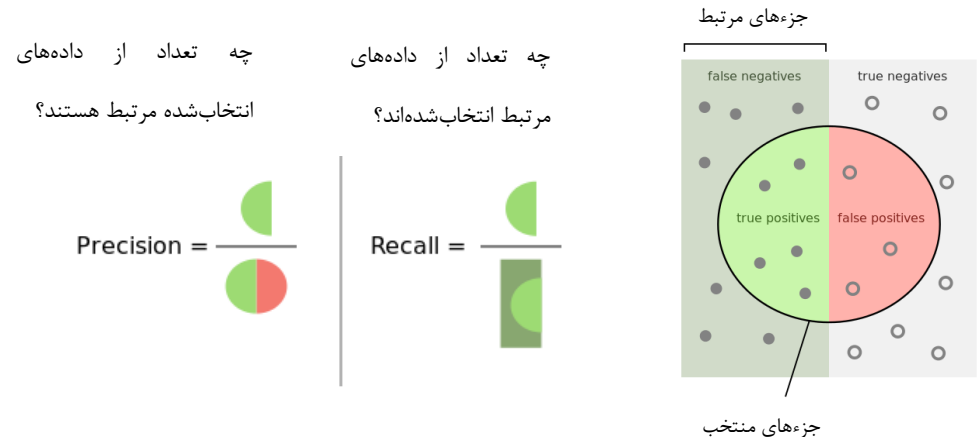
از هر دو آزمون تشخیص[2] p و حساسیت[3] r برای محاسبه نمرهF_1 استفاده می‌شود. P تعداد نتایج مثبت صحیح تقسیم بر تعداد تمام نتایج مثبت توسط طبقه بندی شده است و r تعداد نتایج منفی صحیح تقسیم بر تعداد تمام نمونه های مربوطه (تمام نمونه هایی که باید مثبت باشند). نمره F_1 میانگین هارمونیک دقت و حساسیت است، جایی که نمره F_1 به بهترین وجه باشد به ۱ (precision و recall کامل) و در بدترین حالت به ۰ می رسد. اندازه گیری رایج F یا F-score متعادل (نمره F_1) میانگین هارمونی precision و recall است:

[1] Saski
[2] Precision
[3] Recall

$$F_1 = \left(\frac{recall^{-1}+precision^{-1}}{2}\right)^{-1} = 2.\frac{\text{precision.recall}}{\text{precision+reca}} \qquad (۱۱-۲)$$

شاخص دقت و حساسیت

در تشخیص الگو، بازیابی اطلاعات و طبقه‌بندی دوتایی[1]، precision (ارزش پیش‌بینی مثبت) کسری از نمونه های مربوطه در میان نمونه‌های بازیابی شده است، در حالی که recall (همچنین به عنوان حساسیت شناخته می شود) کسری از نمونه های مربوط است که در کل مقدار نمونه های مربوطه بازیابی شده است؛ بنابراین precision و recall بر درک و اندازه گیری ارتباط استوار است. به عبارت ساده، precision بیشتر به این معنی است که یک الگوریتم نتایج مرتبط تری نسبت به نتایج نامناسب ارائه می دهد، در حالی که recall بیشتر به این معنی است که از یک الگوریتم، نتایج کاملا مرتبط بدست می‌آید.



شکل ۲- ۱۰: نمای شماتیک تعریف precision وrecall

[1] Binary

شاخص جاکارد[1]

شاخص جاکارد، همچنین به عنوان اشتراک در اتحاد یک مجموعه و ضریب تشابه جاکارد (ضریب اصل ضوابط de communauté توسط پل جاکارد)، آماری برای مقایسه شباهت و تنوع مجموعه های نمونه مورد استفاده قرار می گیرد. ضریب تشابه جاکارد اندازه گیری شباهت بین مجموعه نمونه های محدود است و اندازه اشتراک تقسیم بر اندازه اتحاد از مجموعه های نمونه تعریف شده است.(جاکارد،۱۹۰۱) از رابطه ۲-۵ این شاخص محاسبه می‌شود:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|} \qquad (۱۲-۲)$$

اگر A و B هر دو تهی باشند، $J(A,B) = 1$ تعریف می کنیم و $0 \leq J(A,B) \leq 1$.

شاخص NMI[2]

نرمال اطلاعات متقابل (NMI) یعنی امتیاز اطلاعات متقابل (MI[3]) نرمال شده به مقیاس نتایج خوشه، بین ۰ (نبود اطلاعات متقابل) و ۱ (همبستگی کامل) است. اطلاعات متقابل یعنی با کاهش انتروپی برچسب های کلاس بندی که ما دریافت می کنیم اگر برچسب خوشه را بدانیم نشان می‌دهد. از رابطه ۲-۷ این مقدار بدست می‌آید.

$$NMI(Y,C) = \frac{2 \times I(Y;C)}{[H(Y)+H(C)]} \qquad (۱۳-۲)$$

که در آن

$Y =$ برچسب کلاس

[1] Jaccard
[2] Normalized Mutual Information
[3] Mutal information

C = برچسب های خوشه ای

$H(.)$ = آنتروپی

$I(Y; C)$ = اطلاعات متقابل Y و C / w b

NMI اندازه گیری خوبی برای تعیین کیفیت خوشه بندی است. یک اندازه گیری خارجی است، زیرا ما به برچسب کلاس نمونه ها نیاز داریم تا NMI را تعیین کنیم. از آنجا که نرمال است، ما می توانیم NMI را در بین خوشه بندی های مختلف با تعداد مختلف خوشه اندازه گیری و مقایسه کنیم.

شاخص جینی[1]

در اقتصاد، ضریب جینی که گاهی اوقات شاخص جینی یا نسبت جینی نامیده می‌شود، اندازه گیری پراکندگی آماری است که برای نشان دادن توزیع درآمد یا ثروت، توزیع ساکنان یک ملت و مهمترین آن، اندازه گیری نابرابری استفاده می شود. این شاخص توسط جامعه‌شناس و محقق ایتالیایی Corrado Gini و در سال ۱۹۱۲ ارائه شد. ضریب جینی ضریب نابرابری را در بین مقادیر توزیع فراوانی (به عنوان مثال، سطوح درآمد) اندازه گیری می کند. ضریب جینی صفر برابری مطلق را نشان می دهد، در حالیکه تمام مقادیر یکسان هستند (به عنوان مثال که هر کس درآمد مشابهی دارد). ضریب جینی ۱ (یا ۱۰۰٪) بیانگر نابرابری حداکثر بین ارزش ها می‌باشد (به عنوان مثال برای تعداد زیادی از مردم که تنها یک نفر تمام درآمد و مصرف و همه دیگران را دارد، ضریب جینی تقریبا نزدیک به یک است) با این وجود، اگر یک نفر به سهم منفی در مجموع (به عنوان مثال، داشتن درآمد منفی یا ثروت)، یک مقدار بیشتر از یک ممکن است رخ دهد. برای گروه های بزرگتر، مقادیر نزدیک به ۱ یا بالاتر از ۱ در عمل بسیار بعید است.

[1] Gini index

شاخص خلوص

یکی از ساده‌ترین شاخص‌های ارزیابی بیرونی در خوشه‌بندی، شاخص خلوص[1] است که درصد مطابقت بین برچسب‌های خوشه‌بندی و برچسب‌های واقعی را می‌سنجد. در این حالت برچسب هر خوشه با برچسب واقعی دسته‌ای که بیشترین اشتراک را دارد مطابقت پیدا کرده و تعداد نقاطی از خوشه که در دسته صحیح طبقه‌بندی شده‌اند شمارش می‌شوند. نسبت این تعداد به تعداد کل نقاط شاخص خلوص را می‌سازد و شکل محاسباتی آن به صورت زیر است:

$$Purity(S,C) = \frac{\sum_m maxn|Sm \cap Cn|}{N} \qquad (۱۴-۲)$$

باید توجه داشت که منظور از $|Sm \cap Cn|$ تعداد نقاط مشترک از خوشه Cnبا دسته N و Sm نیز تعداد کل نقاط است. با توجه به شیوه محاسبه شاخص خلوص، مشخص است که حداکثر مقدار برای آن ۱ خواهد بود و این در زمانی اتفاق می‌افتد که برچسب‌های حاصل از خوشه‌بندی کاملا با برچسب‌های واقعی مطابقت داشته باشند. همینطور اگر هیچ برچسب خوشه‌ای با برچسب واقعی مطابقت نداشته باشد، این شاخص صفر می‌شود.

خصوصیات شاخص خلوص

سادگی در محاسبات :با توجه به شیوه محاسبه این شاخص، محاسبات طولانی برای بدست آوردن میزان مطابقت لازم نیست.

مستقل از تعداد خوشه‌ها :شاخص خلوص به تعداد خوشه‌ها توجه ندارد. در نتیجه نمی‌توان این شاخص را به عنوان معیاری برای سنجش مطابقت تعداد خوشه‌ها نیز در نظر گرفت.

[1] Purity Index

کاهش کارایی با افزایش تعداد خوشه‌ها :اگر تعداد خوشه‌ها زیاد باشد و هیچ هماهنگی نیز بین برچسب‌های واقعی و خوشه‌ای وجود نداشته باشد ممکن است شاخص خلوص به ۱ نزدیک شود که یک عیب برای چنین شاخصی است.

۲-۳-۵-انواع الگوریتم‌های خوشه‌بندی آماری

در سال ۱۹۹۰ با تمرکز بر شناسایی نقاط پرحادثه، خوشه‌بندی آماری، حرکتی از آماره‌های سراسری به وجود آمد. همچنین روش‌هایی مانند برآورد تراکم کرنل و خودهمبستگی مکانی برای شناسایی موقعیت خوشه‌ها توسعه یافتند. دو روش اصلی که ما در این تحقیق مورد بررسی قرار دادیم عبارتند از روش تراکم کرنل در سطح صفحه‌ای و آماره خودهمبستگی مکانی که با استفاده از نرم افزار GIS امکان‌پذیر است.

روش تخمین تراکم کرنل صفحه‌ای KDE[1]

این روش در سال ۱۹۵۶ توسط محققی به نام رزنبالت در جهت تخمین ناپارامتری تراکم معرفی شد. این ابزار برای تخمین تراکم عوارض در همسایگی یک نقطه را محاسبه می‌کند. تخمین تراکم کرنل روشی رایج برای تحلیل داده‌های یک‌بعدی و دوبعدی و همچنین شناسایی نقاط متمرکز یا حادثه‌خیز است و از طرفی اطلاعات مفیدی را در مورد داده‌ها ارائه می‌دهد (دنگ[2]، ۲۰۰۷). این روش می‌تواند به‌سرعت نقاط حادثه‌خیز را از پایگاه‌های داده‌ای بزرگ مانند تصادفات را شناسایی کرده و درنتیجه خروجی آماری قابل قبولی را ارائه دهد.

از مزایای این روش می‌توان به نمایش نقاط متمرکز تصادفات در سطح (منطقه‌ای) و نقطه و شبکه راه‌ها به‌خصوص در تصادفات اشاره کرد. این روش برخلاف روش‌های خودهمبستگی مکانی که

[1] Planar Kernel Density Estimation
[2] Duong

نقاط متمرکز را شناسایی می‌کنند می‌توانند مناطق حادثه‌خیز را شناسایی کنند. مزیت اصلی استفاده از تخمین تراکم کرنل، تشخیص محدوه ی گسترش خطر تصادفات است درحالی‌که در خوشه‌بندی، صرفا دسته بندی بر اساس تراکم داده ها خوشه بندی انجام می‌شود. درخوشه‌بندی، اطراف هر خوشه صرفاً در نهایت دارای درجه‌ای خطر تصادف برای آن‌هایی است که وارد آن خوشه می‌شوند(سابل و همکاران[1]،۲۰۰۵).

هر نقطه داده x_i و $i = 1,2, \dots, n$ نمونه مستقل و با توزیع یکسانی با چگالی نامعلوم f هستند. تابع برآورد تراکم کرنل ($\widehat{f_h}(x)$) چگالی نامعلوم $f(x)$ را در نقطه x برآورد می‌نماید. تابع برآورد تراکم کرنل $\widehat{f_h}(x)$ به‌صورت رابطه شماره ۲-۸ محاسبه می‌شود:

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right) \qquad (۱۵-۲)$$

که در آن $\widehat{f_h}(x)$تابع برآورد تراکم کرنل، $k\left(\frac{x-x_i}{h}\right)$ تابع کرنل مثبت متقارن با انتگرال واحد است و h دلالت بر پهنای باند یا شعاع همسایگی دارد. به ازای کاهش پهنای باند(h)، منحنی حاصل از برآورد کرنل ناهموارتر شده و جزئیات جعلی بیشتری را از تراکم واقعی به نمایش می‌گذارد و به ازای افزایش پهنای باند، منحنی هموار و باعث محو شدن جزئیات واقعی تابع تراکم می‌گردد. تخمین کرنل مجموع «برآمدگی»هایی است که مشاهدات در مرکز آن‌ها واقع‌شده‌اند. تابع کرنلk، شکل برآمدگی‌ها و پهنای باند (h)، طول آن‌ها را مشخص می‌کند. از تابع کرنلی که به‌طورمعمول استفاده می‌شود می‌توان به توزیع نرمال اشاره کرد.

[1] Sabel et al.

تحلیل خودهمبستگی مکانی[1]

زمانی که مقادیر یک متغیر که از نظر جغرافیایی به هم نزدیک هستند باهم مرتبط می‌باشند که به این تعریف خودهمبستگی مکانی می‌گویند. ارزیابی روش‌های خودهمبستگی مکانی عمدتاً به‌وسیله تحقیقات انسلین[2] ۱۹۹۵ و موران[3] ۱۹۴۸ شروع شد. هنگامی‌که حوادث وابسته هستند گفته می‌شود که همبستگی مکانی مثبت است. اگر عوارض و یا مقادیر متغیرهای مربوط به آن‌ها به‌طور تصادفی در فضا توزیع‌شده باشند ظاهراً نباید بین آن‌ها ارتباطی وجود داشته باشد (گیریفیس[4]، ۲۰۰۸). چندین روش برای آزمایش همبستگی مکانی وجود دارد نظیر Moran's I، * Gi Ord-Getis و Anselin Local Moran's Iکه در ادامه به بررسی آن‌ها پرداخته می‌شود(Moran 1948).

روش خودهمبستگی مکانی Moran's I

Moran's I که مقدار خروجی آن بین ۱ و ۱- تعریف‌شده است، یکی از روش های خودهمبستگی مکانی است که در زمینه تحلیل الگوهای توزیع عوارض و پدیده‌ها در مکان استفاده می‌شود. این ابزار یکی از کاربردی‌ترین و مهم‌ترین ابزارهای تحلیلی برای تحقیق در مورد داده‌های مکانی است. به‌طور مثال در روش‌هایی که نیاز به داده‌های مستقل است با کمک ابزار خودهمبستگی مکانی می‌توان دریافت که آیا داده‌ها از نظر مکانی مستقل از یکدیگر هستند و یا خیر .درصورتی‌که نتایج تحلیل خودهمبستگی Moran's I نشان‌دهنده وجود خودهمبستگی مکانی بین داده‌ها باشند، در آن صورت نمی‌توان از روش‌هایی که نیازمند داده‌های مستقل هستند استفاده نمود. ابزار تحلیل خودهمبستگی مکانی Moran's I به بررسی خودهمبستگی مکانی بر اساس مکان دو مقدار ویژگی موردنظر عوارض جغرافیایی می‌پردازد. اگر تعدادی عوارض جغرافیایی با ویژگی مشخصی موجود باشند

[1] Spatial Autocorrolation
[2] Anselin
[3] Moran
[4] Griffith

این ابزار نشـــان می‌دهد که الگوی پراکنش این عوارض با در نظر گرفتن مقادیر ویژگی موردمطالعه از الگوی خوشـه‌ای یا پراکنده برخوردار اسـت. نتایج حاصـل از این تحلیل نشـان می‌دهد که آیا عوارض به‌صورت تصادفی، پراکنده یا خوشه‌ای در مکان توزیع‌شده‌اند. این ابزار در حقیقت آمـــاره و یا شاخص Moranرا محاسبه می‌کند و با استفاده از امتیاز استانداردZ(Z-Score) و P-Value به ارزیابی و معنادار بودن شاخص محاسبه‌شده می‌پردازد رابطه، نشان‌دهنده شاخص Moran است.

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{(\sum_{i\neq j}^{n}\sum_{j=1}^{n} w_{ij})(\sum_{i=1}(x_i-\bar{x})^2)} \qquad (۱۶-۲)$$

در این رابطه n نشان‌دهنده تعداد رخدادها و$\bar{x}$ مقدار متوسط کل رخدادها و x_j تعداد رخدادها در موقعیت i ام وj ام است. $(x_i - \bar{x})$ نیز تفاضـــل بین مقدار ویژگی عـــــارضه i با میانگین آن است. w_{ij} ماتریس وزن مکانی تعریف شده برای تعیین درجه تأثیرات محلی رویدادها است. فرمول ریاضی Z-Score ،اهمیت نتایج به دست آمده را به وسیله رابطه ۲-۱۰ آزمایش می‌کند:

$$z = \frac{I_O - I_E}{SD_{IE}} \qquad (۱۷-۲)$$

در رابطه ۲-۱۰ ،I_O، I_Eو SD_{IE} به ترتیب نشان‌دهنده مشاهدات، انحراف مورد انتظار(میانگین) و انحراف استاندارد هستند.

Z-Score بزرگ و P-Value کوچک برای هر نقطه نشان‌دهنده خوشه‌بندی مکانی مقادیر بالا است. Z-Score پائین و منفی وP-Value کوچک نشان‌دهنده خوشه‌بندی مکانی مقادیر پایین است. برای ابزارهای تحلیل الگو، مقدارP-Value احتمال تولید الگوی مکانی مشاهده‌شده به وسیله فرآیندهای تصادفی است. P-Value خیلی کوچک (احتمال پایین) به این معنا است که الگوی مکانی مشاهده‌شده نتیجه فرایندهای تصادفی بوده است، بنابراین فرض صفر رد می‌شود. به طور کلی اگر مقدار شاخص Moranنزدیک به عدد ۱+ باشد داده‌ها دارای خودهمبستگی مکانی و دارای الگوی خوشه‌ای بوده و اگر

مقدار شاخص Moran نزدیک به عدد ۱- باشد، آنگاه داده‌ها از هم گسسته و پراکنده هستند. در مورد این ابزار فرضیه صفر آن است که هیچ نوع خوشه‌بندی مکانی بین مقادیر ویژگی مرتبط با عوارض مکانی مورد نظر وجود ندارد. حال زمانی که مقدارP-Value بسیار کوچک و مقدارZ-Score محاسبه‌شده (قدر مطلق آن) بسیار بزرگ باشد (خارج از محدوده اطمینان قرار گیرد)، آنگاه می‌توان فرضیه صفر را رد کرد. اگر مقدار شاخص Moran بزرگتر از صفر باشد، داده‌ها نوعی خوشه‌بندی مکانی را نشان می‌دهند. اگر مقدار شاخص کمتر از صفر باشد عوارض مورد مطالعه دارای الگوی پراکنده هستند.

۲-۳-۶- استفاده از GIS در مطالعات پیشین

رایج‌ترین کاربرد نرم‌افزار GIS در سطح ماکروسکوپیک تعیین نقطه پرحادثه می‌باشد. نقاط حادثه‌خیز مناطقی هستند که در آن‌ها یک واقعه با تمرکز بالا رخ می‌دهد. در پژوهش حاضر منظور از نقاط حادثه‌خیز مناطقی می‌باشد که دارای تمرکز بالای تصادفات منجر به جراحت می‌باشند. نتایج تحلیل نقاط حادثه‌خیز می‌توانند به‌صورت تصویری در غالب نقشه‌هایی که تحت عنوان «سیستم اطلاعات جغرافیایی» GIS نامیده می‌شوند به نمایش در آیند. این قابلیت باعث افزایش امکان شناخت وضع موجود و مقایسه نقاط حادثه‌خیز با شرایط کلی منطقه می گردد. این نتایج تصویری می‌توانند توسط مدیران جامعه و طراحان شهری موردبررسی قرار گیرند تا دلایل احتمالی شکل گیری چنین الگویی هایی در وقوع حوادث خاص مشخص شده و راهکارهای مؤثر در جهت حل مشکلات موجود اتخاذ گردد .درحالی‌که متمرکز بودن وقایعی نظیر جرایم، حوادث طبیعی و همینطور تصادفات رانندگی منطقاً دور از انتظار نیست اما به دلیل دخیل بودن عوامل مختلف در تحلیل نقاط حادثه‌خیز، لازم است حداکثر دقت در انتخاب روش تحلیل و پارامتر های مؤثر در هر روش صورت گیرد تا بهترین و قابل اعتماد ترین نتایج حاصل گردند. با اینکه امروزه روند شناسایی نقاط حادثه‌خیز با استفاده از برنامه های

رایانه ای خیلی پیچیده نمی‌باشد در هر حال جهت کسب بهترین نتایج، آگاهی از زمینه های تئوری و منطق عملکردی این ابزارها ضروری می‌باشد.

یکی از مزایای تحلیل‌های GIS محور امـکان روی هم گذاری نتایج این تحلیل ها با لایه های مختلف مشخصات شبکه های حمل و نقل می‌باشد که قابلیت بررسی ارتباط بین تراکم تصادفات با هر یک از مشخصات را ممکن می‌سازد.

نقاط حادثه‌خیز توسط روش‌هایی که بمنظور شناسایی تمرکز وقایع طراحی شده اند تعیین می گردند .روش‌های بسیاری وجود دارند که برای تعیین نقاط حادثه‌خیز مورداستفاده قرار می گیرند و در برخی مواقع نتایج متفاوتی خواهند داشت. درحالی‌که اکثر مطالعات GIS نه‌تنها به‌منظور نمایش تصویری نتایج بلکه به‌عنوان ابزاری جهت جمع بندی اطلاعات از منابع مختلف نیز به کار گرفته شده هنوز تعدادی از ابزارهای تحلیل آمار مکانی آن طور که باید و شاید از طرف ارگان های مربوطه مورداستفاده قرار نمی گیرد. همانطور که قبلاً نیز اشاره شد باوجود پیشرفت های سریع GIS در سال های اخیر همچنان بسیاری از نهاد های مرتبط به دلیل نداشتن پایگاه داده GISمحور، امکان استفاده از GIS و ابزار های سودمند و در عین حال ساده وصریح این نرم افزار ها را ندارند، یکی دیگر از مضرات عدم استفاده از GIS در تعیین نقاط حادثه‌خیز تصادفات، احتمال نادیده گرفته شدن برخی اطلاعات مربوط به موقعیت مکانی می‌باشد که ممکن است تأثیر بسزایی نیز در وقوع تصادفات داشته باشند. تحلیل نقاط حادثه‌خیز تصادفات نیازمند درک هم جانبه روند وقوع تصادف، شدت صدمات بوجود آمده و محیط اطراف محل حادثه می‌باشد. محیط GIS به دلیل ارائه سیستم کارآمد مرتبط نمودن پایگاه های داده متعدد مناسب استفاده در این گونه موارد می‌باشد. این پلت فرم یک سیستم مرجع مکانی برای گزارش گیری در سطوح مختلف ارائه می‌کند که ضمن قابلیت وارد کردن داده‌های آماری و موقعیتی تصادفات امکان برآورد خطر و شدت صدمات وارده را فراهم می نماید .طبقه بندی نقاط حادثه‌خیز تصادفات معمولاً بر اساس اطلاعات موجود درباره همان تصادفات مانند نوع تصادف، زمان وقوع، نوع مصدوم، نوع

وسیله نقلیه و... صورت می گیرد. همین موضوع اغلب باعث محدود شدن جوانب تحلیل مسئله می گردد و منجر به از قلم افتادن بر خی از عوامل محیطی، اجتماعی و اقتصادی به دلیل نبود داده‌های اولیه می گردد. با این همه، نکته مثبت، رویکرد پژوهشگران عرصه حمل و نقل به سوی استفاده روزافزون از این ابزار ها و مطالعه و بررسی در بهبود و گسترش کاربرد بیش از پیش این روش ها در تحلیل مشکلات موجود می‌باشد.

پاراسانا کومار و همکارانش (پاراسانا کومار و همکاران، ۲۰۱۱)، تحلیل خوشه ای را با توجه به نوع حادثه و زمان وقوع با استفاده از Index Moran روش خودهمبستگی مکانی مطالعه کردند. تحلیل خوشه ای با آمار KDE و Gtis-Ord GI انجام شد. در نتیجه تجزیه و تحلیل ها، لکه های گرم و سرد نقاط تصادف تعیین شد. موریرا و همکاران (۲۰۱۲) سه روش مختلف برای شناسایی مکان راه‌های خطرناک (HRL) در شهر ویارئال پرتغال را مورد بررسی قرار داند : الگوریتم خوشه بندی Nnh، KDE و تراکم نقطه ای. سرعت بالا به عنوان یک عامل شاخص تصادفات تک خودرویی(تصادف خروج از جاده) پیدا شد. اکی و کرتی (۲۰۱۲) خوشه ای از تصادفات در شبکه را با استفاده از نرم افزار Sanet جستجو کردند. خوشه مکانی- زمانی با روش ناکس انجام شد.

تسا اندرسون (۲۰۰۹) با ارائه دو روش به مطالعه تصادفات جاده‌ای پرداختند. در روش اول آن‌ها با استفاده از سیستم اطلاعات جغرافیایی و تخمین تراکم کرنل، الگوهای مکانی مرتبط با تصادفات جراحتی را در لندن موردمطالعه قراردادند. هم‌چنین آن‌ها از روش خوشه‌ای، با بهره‌گیری از داده‌های محیطی و نیز نتایج حاصل از روش تراکم کرنل، جهت دسته‌بندی مناطق حادثه‌خیز جاده‌ای استفاده کردند. پردازش خوشه‌ای در مطالعه آن‌ها ۵ گروه و ۱۵ خوشه ایجاد کرد و تفاوت اصلی بین گروه‌ها و خوشه‌ها، واریانس متغیرهای ساختاری می‌باشد. نتایج آن‌ها نشان داد که واریانس متغیرها در هر گروه به علت ساختار سلسله مراتبی به‌طور قابل‌ملاحظه‌ای کاهش می‌یابد. مزیت اصلی روش تخمین تراکم

کرنل، تشخیص توزیع ریسک تصادف در محدوده خوشه‌ای است که به دلیل استقلال مکانی، احتمال رخداد تصادف در آن بیشتر می‌باشد.

از مزایای استفاده از GIS می‌توان به امکان مدیریت و پردازش داده‌های مکانی و اطلاعات مرتبط با این داده‌ها اشاره کرد(Longley et al., 2005). تشخیص نقاط حادثه‌خیز با استفاده از GIS امکان‌پذیر شده است و ازآنجایی‌که در سال‌های اخیر ثبت داده‌ها به‌صورت دقیق و توسط GPS انجام می‌شود، تشخیص مناطقی که از جنبه‌های مختلف به هم مشابهت دارند، می‌تواند درک دقیق‌تر و جامع‌تری را از ویژگی تصادفات در اختیار قرار دهد. به این منظور، در کنار استفاده از GIS و روش خوشه‌بندی می‌توان به این هدف دست‌یافت.

باران اولاک و همکارانش در سال ۲۰۱۷ تصادفات را با در نظر گرفتن سن افراد موردبررسی قراردادند. سه مرحله برای تحلیل مکانی انجام شد. ابتدا خوشه‌بندی بر اساس تخمین تراکم کرنل در شبکه (با استفاده از افزونه SANET[1] در GIS) استفاده شد. سپس ضریب تفاوت نرخ تراکم (DRD[2]) برای دو گروه سنی برای مقایسه تراکم بیشینه و نرمال شده به دست آمد. در مرحله سوم ضریب جمعیت(PF[3]) برای بررسی تأثیر وابستگی مکانی محاسبه شد. با استفاده از این ضرایب می‌توان مناطق مهم برای سیاست‌گذاری در جهت تصمیم‌گیری برای ایمنی با محوریت سن افراد را مشخص نمود(Ulak, Ozguven et al.۲۰۱۷).

سلطانی و عسکری در سال ۲۰۱۷ برای خیابان‌های شریانی درجه اول شیراز توسط GIS خوشه‌بندی مکانی را با روش‌هایی همچون انجام دادند. این تحقیق به‌منظور یافتن الگوهای زمانی و مکانی تصادفات در محیط شهری انجام شد. نقشه‌های موازی برای مقایسه سه فاکتور زمان رخداد، شدت و موقعیت مکانی تصادف استفاده شد. تحلیل فراوانی سالانه تصادفات در ۱۵۶منطقه تصادفات

[1] Spatial analysis along network
[2] Density ratio difference
[3] population factor

ترافیکی در شیراز از سال ۲۰۱۰ تا ۲۰۱۴ نشان داد که دو روش شاخص I موران[۱] و گتیس اورد *Gi [۲] خوشه‌بندی معناداری از الگوهای تصادف ایجاد می‌کند. این تحقیق نشان داد که تصادفات دارای الگوی خوشه‌ای است اما در تحلیل مکانی-زمانی الگوی تصادفی از خود نشان داد.

شفابخش و همکاران[۳] در سال (۲۰۱۷) برای تصادفات در نقاط درون‌شهری مشهد از GIS برای تحلیل مکانی از روش تخمین چگالی کرنل و از افزونه SANET برای تخمین چگالی کرنل در شبکه راه‌ها استفاده کردند. سپس با کمک نتایج به‌دست‌آمده، سطح موردمطالعه را در پنج منطقه با اولویت‌های متفاوت جهت اقدامات ایمنی طبقه‌بندی کردند.

[1] Moran’s I index
[2] Getis ord Gi*
[3] Shafabakhsh et al.

فصل ۳

روش تحقیق

۳-۱- مقدمه

در این بخش ابتدا داده‌های مورد استفاده و محدوده مطالعه را بررسی کرده و سپس به شرح روش‌های ذکر شده در فصل دوم می‌پردازیم و روند دقیق انجام شده در این پایان‌نامه آورده شده است. در تحلیل نقاط حادثه‌خیز تصادفات شهری، همانند نقاط حادثه‌خیز بزرگراه‌ها برای شناسایی نقاط حادثه‌خیز تعریف یا چارچوب معینی در کارهای گذشته وجود ندارد. این عمدتا به این دلیل است که پیچیدگی و ترکیب بودن مدهای مختلف حمل و نقل در شبکه شهری می‌باشد و همچنین مبداءها و مقصدهای فروانی برای مناطق درون‌شهری وجود دارد. با توجه به اینکه حجم ترافیک در مناطق شهری به‌صورت مکانی تصادفی نیستند، تصادفات نیز ازلحاظ مکانی تصادفی نیستند. علاوه بر این وجود جابجایی عابرین پیاده در مناطق شهری بر پیچیدگی تحلیل شبکه شهری می‌افزاید که بیشتر به کابری اطراف مسیر مرتبط می‌باشد. به همین دلیل در مناطق شهری نمی‌توان مناطق را به بخش های یکنواخت و نرمال تقسیم کرد.

پس از آشنایی و تحلیل پارامترهای مختلف تأثیرگذار بر وقوع تصادفات، همچنین بررسی روش‌های خوشه‌بندی مکانی که در فصل دوم ارائه شد، شرایط لازم جهت ارائه روشی برای خوشه‌بندی بر مبنای اطلاعات مکانی به‌منظور تحلیل مکانی تصادفات مهیا است. در این فصل ابتدا به تشریح منطقه موردمطالعه، جمع‌آوری و آماده‌سازی داده‌ها پرداخته می‌شود. روش‌های خوشه‌بندی غیر آماری GriDBSCAN و Nnh و K-Means به‌طور جداگانه برای دو شهر ایزمیت و گبزه به‌منظور تعیین خوشه‌ها با بیشترین تعداد تصادفات انجام می‌شود. بعد از آن روش‌های آماری همچون توابع همبستگی Gi ord-Getis وAnselin Local Moran’sI ، تابع برآورد تراکم کرنل بر روی داده‌ها اعمال شد. سپس مقایسه‌ای بین خروجی حاصل از روش‌های خوشه‌بندی غیر آماری و آماری صورت پذیرفت تا مشخص شود که کدام روش خوشه‌بندی غیر آماری مناطق حادثه‌خیز را بهتر مدل می‌کنند. سپس از بین روش‌های مذکور بر اساس داده‌های اولیه، روش خوشه‌بندی غیرآماری که مناطق

حادثه‌خیز را بهتر مدل می‌کند و همچنین بهترین روش از بین روش‌های آماری انتخاب می‌شوند. بعد از آن، روش‌های برگزیده روی داده‌های واقعی پیاده‌سازی می‌شوند و درنهایت نقاط حادثـه‌خیز با داده-های واقعی شناسایی می‌شوند.

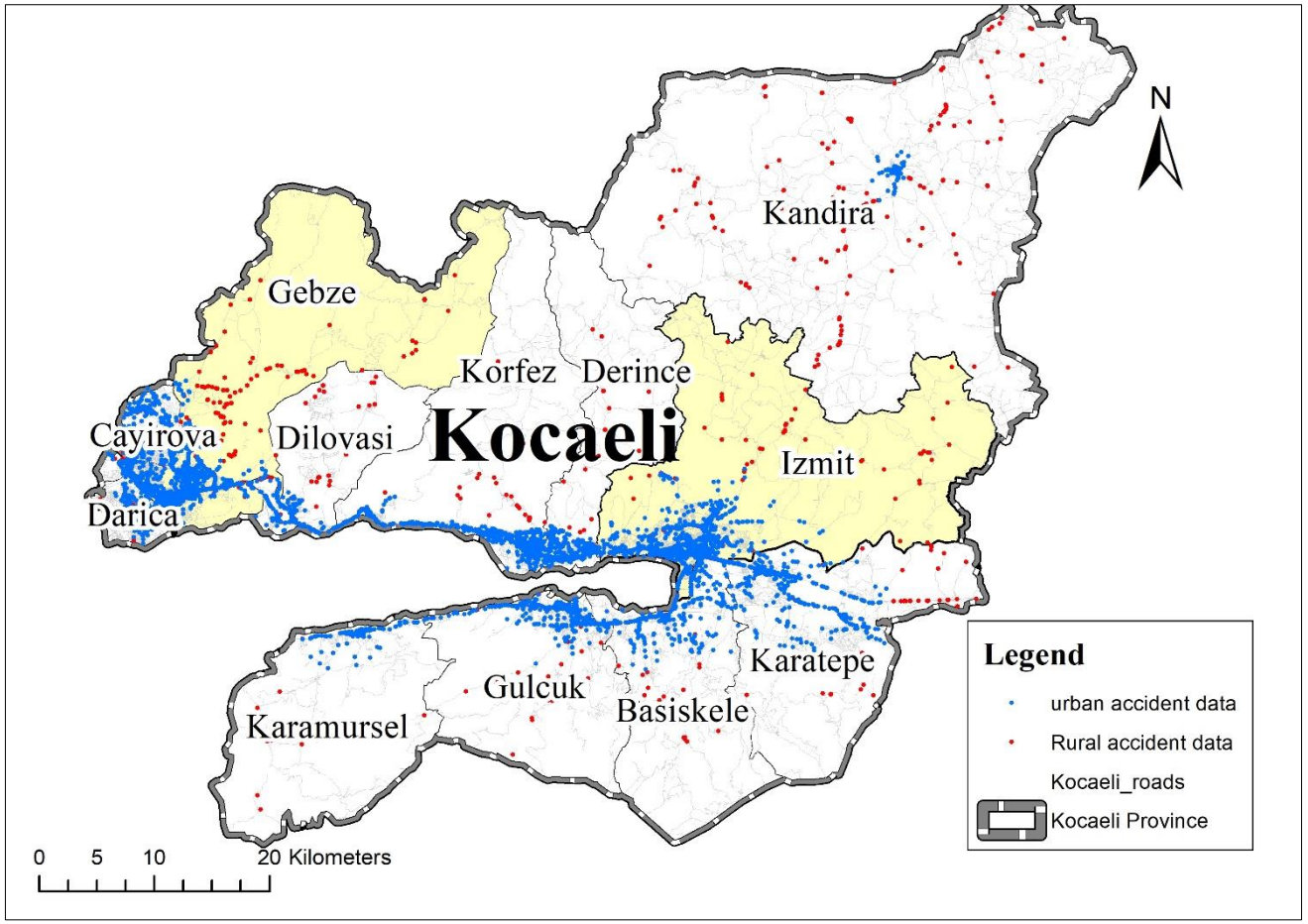
۳-۲- محدوده موردمطالعه

در این پایان‌نامه داده‌های مربوط به استان کوجائلی[1] استفاده‌شده است. دادههای مورد استفاده در این تحقیق شامل موارد زیر است :نقشه راه‌های استان کوجائلی در فرمت .shp و در قالب یک لایه خطی و فاقد اطلاعات توصیفی راه، موقعیت نقاط تصادفات موجود در فرمت .xlsx با مشخصات مکانی، شرایط جوی، شرایط هندسی مسیر، و... که به تشریح در ادامه بررسی شده است. با توجه به دقت بالای ثبت داده‌های تصادف در این استان که با GPS به صورت دقیق مختصات مکانی نقاط تصادف ثبت شده است لذا در تحقیقات دیگر نیز با داده‌های مکانی با مختصات دقیق می‌توان مطالعاتی از این دست انجام داد. ما ۲ شهر ایزمیت[2]، گِبزه[3] از این استان را به تفکیک موردبررسی قراردادیم. این شهرها در مسیر راه‌های شریانی درجه یک ترکیه و در کنار دریای مرمره می‌باشند. بر اساس آمار نفوس رسمی ترکیه سال ۲۰۱۸ جمعیت هرکدام از شهرها برای سال‌های ۲۰۱۳-۲۰۱۴ در جدول شماره ۳-۱ آمده است. درشکل۳-۲ موقعیت جغرافیایی استان و شهرهای انتخابی قابل مشاهده است.

جدول۳-۱ : جمعیت شهرهای استان کوجائلی(سال۲۰۱۳-۲۰۱۴) (Turkstat, ۲۰۱۸)

منطقه	۲۰۱۳	۲۰۱۴
کوجائلی	۱,۶۷۶,۲۰۲	۱,۷۲۲,۷۹۵
ایزمیت	۳۳۲,۷۵۴	۳۳۸,۷۱۰
گِبزه	۳۲۹,۱۹۵	۳۳۸,۴۱۲

[1] Kocaeli
[2] İzmit
[3] Gebze

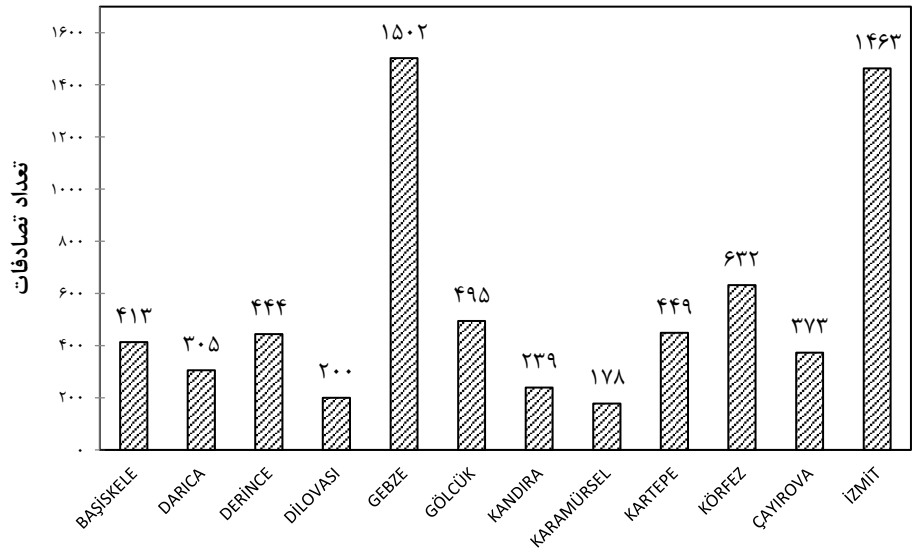


شکل ۳- ۱: استان کوجائلی ترکیه و توزیع نقاط تصادف درون‌شهری و برون‌شهری برای سال ۲۰۱۳-۲۰۱۴

شکل۳- ۲: عکس ماهواره‌ای شهرهای اطراف دریای مرمره

در شکل ۳-۱ توزیع نقاط در کل استان قابل مشاهده است و نقاط آبی نشان‌دهنده نقاط درون-شهری است که تعداد تصادفات درون‌شهری به مراتب بیشتر است. در شکل ۳-۲ موقعیت جغرافیایی محدوده مورد مطالعه قابل مشاهده است.

با توجه به آمار گزارش شده از سازمان پلیس ملی ترکیه شکل ۳-۳ تعداد کل تصادفات در دو شهر گبزه و ایزمیت بیشترین تعداد تصادفات را در بر داشته است که به ترتیب ۱۵۰۲ و ۱۴۶۳ تصادف از ۶۶۸۹ تصادف این استان مربوط به این دو شهر است.



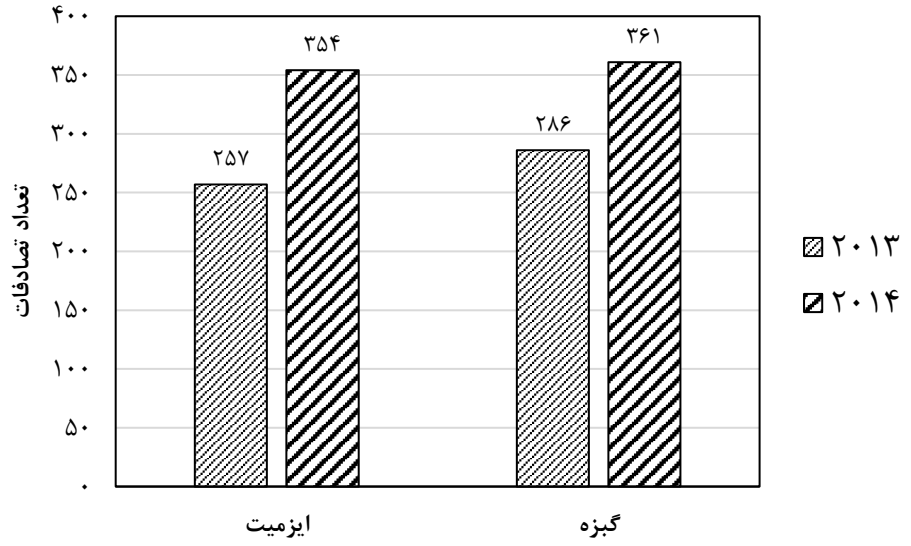
شکل۳- ۳: تعداد تصادفات کل به تفکیک شهرهای استان کوجائلی در سال ۲۰۱۳-۲۰۱۴

ابتدا شهر گبزه که واقع در ۶۵ کیلومتری (۳۰ متری) جنوب شرقی استانبول، در خلیج ایزمیت، بازوی شرقی دریای مرمره واقع شده است، بررسی کردیم. گبزه دومین منطقه بزرگ در کوجائلی پس از ایزمیت ، مرکز استانی از نظر جمعیت است. گبزه در سال های اخیر رشد سریع جمعیتی را تجربه کرده است، از ۱۵۹۱۱۶ نفر در سال ۱۹۹۰ به ۲۹۹۰۴۷ نفر در سال ۲۰۱۱. منطقه گبزه در قسمت غربی استان کوجائلی واقع شده است، با همسایگان توزلا در غرب؛ پندیک، شمال غربی؛ شیل، شمال؛ و کورفز، به شرق. این ناحیه شامل ۶ شهرداری (گبزه، داریچا، چاییروا، دیلوواسی، شکرپینار و توسانچیل) و ۲۲ روستا است. انتهای شمالی پل عثمان Gazi در این منطقه قرار دارد؛ این پل با طول ۴ کیلومتر دریای مرمره را از کاباوورون به دیلبرونو را وصل می‌کند.

سپس شهر ایزمیت واقع در قسمت شرقی این استان که در مسیر محور اصلی D100 (جاده متصل کننده شرق به غرب) و در کنار دریای مرمره قرار دارد، موردبررسی قرار دادیم. ساختار توپوگرافی ایزمیت از دو بخش تشکیل‌شده است. ابتدا در دشت‌های مسطح ساخته شده بود که مرکز شهر می‌باشد که شبکه‌های راه آهن و بزرگراه از این منطقه عبور می‌کنند که نزدیک مرز دریایی است. بخش دوم بر روی تپه ها ساخته شد که بسیاری از خانه های تاریخی از دوران عثمانی در این بخش ساخته شده است.

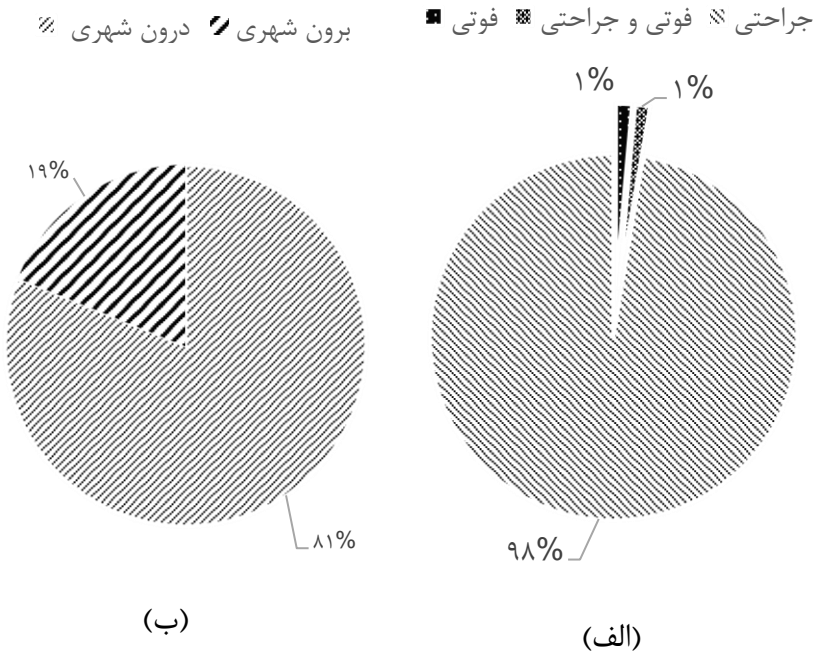
۳-۲- معرفی داده‌ها

تعداد کل تصادفات در این استان برای سال ۲۰۱۳-۲۰۱۴ به تعداد ۶۶۸۹ تصادف بود. آمار تصادفات برای مناطق درون‌شهری ۵۶۸۹ تصادف و برون‌شهری ۱۰۰۰ تصادف بود. پاکسازی داده‌ها در دو جهت سطری و ستونی(اصلاح سطری یعنی اصلاح تک به تک رکوردها از لحاظ کامل بودن پارامترها برای هر رکورد تصادف و اصلاح ستونی یعنی اصلاح زیر شاخه‌های هر پارامتر در هر ستون) انجام شد. برای مثال تعدادی از ستونها به دلیل ثبت ناقص یا تعداد کم رکوردهای تصادفات حذف شد.پس از پاکسازی داده‌ها ۶۴۷ و۶۱۱ تصادف درون‌شهری به ترتیب برای شهر گبزه و ایزمیت برای تحلیل و اعمال روشهای خوشه‌بندی مورد استفاده قرار گرفت.در شکل ۳-۴ تعداد تصادفات درون‌شهری(پس از پاکسازی) به تفکیک برای سال ۲۰۱۳- ۲۰۱۴ قابل مشاهده است.

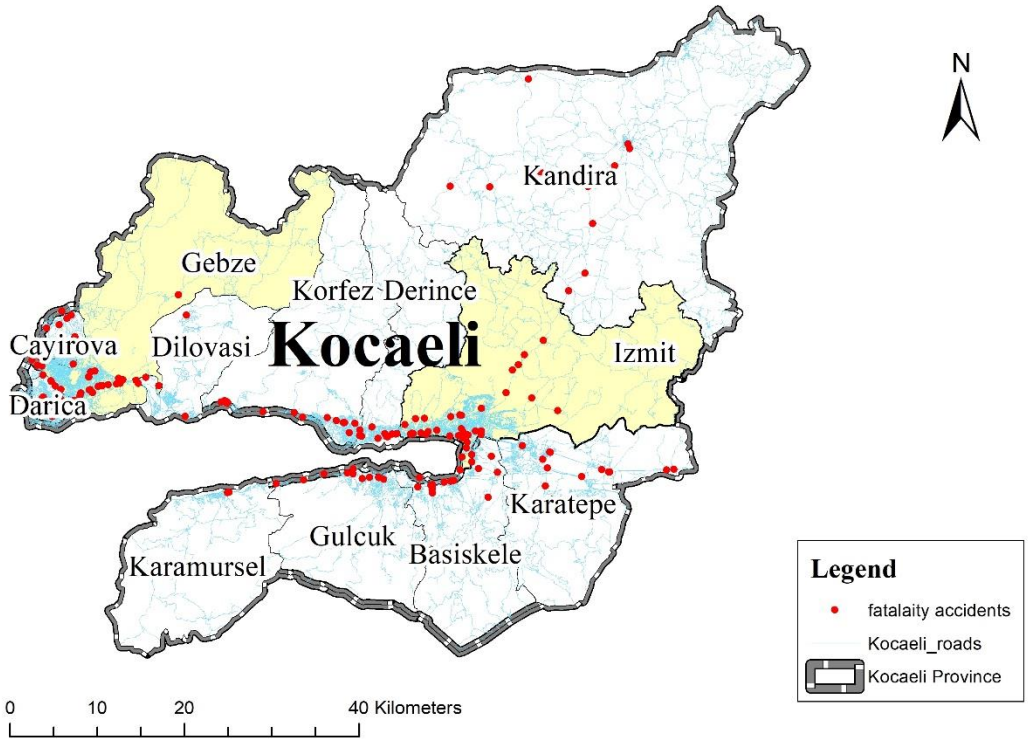


شکل۳- ۴- تعداد تصادفات درون‌شهری به تفکیک شهرها در سال ۲۰۱۳ و ۲۰۱۴ پس از پاکسازی داده‌ها

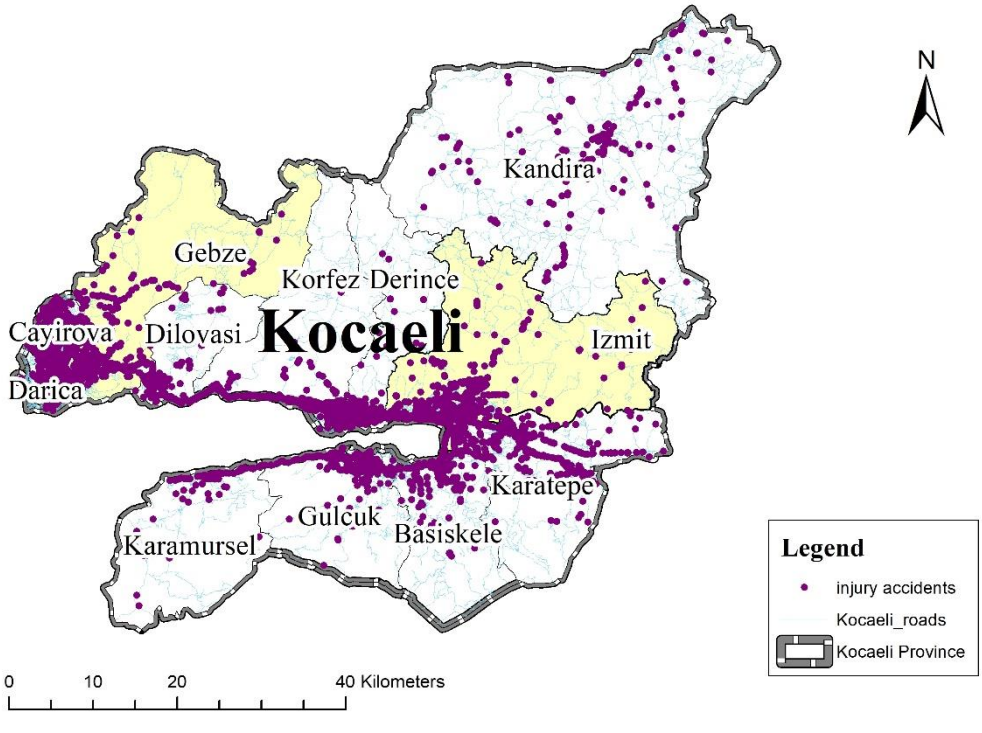
با توجه به فراوانی داده‌های درون‌شهری شکل ۳-۵ صرفا از داده‌های درون‌شهری استفاده شد و همینطور به دلیل ناچیز بودن تعداد تصادفات فوتی گزارش شده شکل ۳-۵ در داده ها صرفا از داده های جراحتی استفاده شد. در شکل۳-۶ نیز بر روی نقشه استان کوجائلی قابل تشخیص می‌باشد.



شکل۳- ۵: درصد تصادفات (الف) فوتی و جراحتی در داده های (ب) درون‌شهری و برون‌شهری در کل استان



(الف)



(ب)

شکل۳- ۶: نقاط تصادف فوتی(الف) و جراحتی (ب) در کل استان کوجائلی

۳-۴-چارچوب تعیین نقاط پرتراکم با روش‌های خوشه‌بندی

۳-۴-۱- آماده سازی داده ها

برای انجام صحیح تحلیل های GIS و همینطور خوشه‌بندی مناسب به داده‌هایی با کیفیت بالا نیاز است. بنابراین اطلاعات ورودی باید آماده‌سازی گردد. در این تحقیق نیز با توجه به دقت داده‌های تصادف و نیز انتخاب روش‌های مکانی برای تحلیل تصادفات در مناطق و شبکه درون‌شهری، جهت آماده‌سازی داده‌ها انجام شد. پس از آن، داده‌های تصادف را هم در خوشه بندی هم در تحلیل های GIS مورد استفاده قرار گرفت. آماده سازی داد ها بدین ترتیب بود که ابتدا نقاط کل تصادف برای این استان را که شامل همه ی تصادفات رخ داده در سال ۲۰۱۳-۲۰۱۴ بود را از لحاظ جامعه آماری برای نقاط تصادفات درون‌شهری و برون‌شهری مقایسه کردیم و با توجه به تعداد قابل قبول تصادفات در نقاط درون‌شهری این نقاط برای مطالعه انتخاب گردید. سپس تک تک رکوردهای تصادف را از لحاظ کامل بودن داده ها در تمام پارامترهای مرتبط با تصادف بررسی و پاکسازی انجام شد.

در این نوع تحلیل ها هر چقدر منطقه مورد مطالعه مخصوصا برای هر شهر مقیاس کوچکتری باشد، تحلیل های دقیق تری انجام خواهد شد. به همین علت، در انتها با توجه به بررسی مطالعات پیشین و توصیه های محققین، داده های هر شهر را تفکیک نموده و جداگانه برای خوشه بندی و تحلیل های GIS مورد استفاده قرار دادیم.

عوامل ثبت‌شده در داده‌های موجود به تعداد ۴۰ پارامتر بود اما به دلیل کامل نبودن عوامل تنها ۲۰ پارامتر از توصیفی و عددی از کل گزارش سازمان پلیس ترکیه مورد استفاده قرار گرفت. در جدول ۳-۳ سه داده عددی (سرعت مجاز و تعداد خودروهای درگیر تصادف و تعداد افراد مجروح) و جدول ۳-۴ داده‌های توصیفی را به همراه فراوانی داده‌های تصادف، نشان می‌دهد که شامل ۱۷ داده کیفی می‌باشد.

جدول ۳- ۲- خلاصه داده‌های عددی به تفکیک شهر برای سال ۲۰۱۳-۲۰۱۴

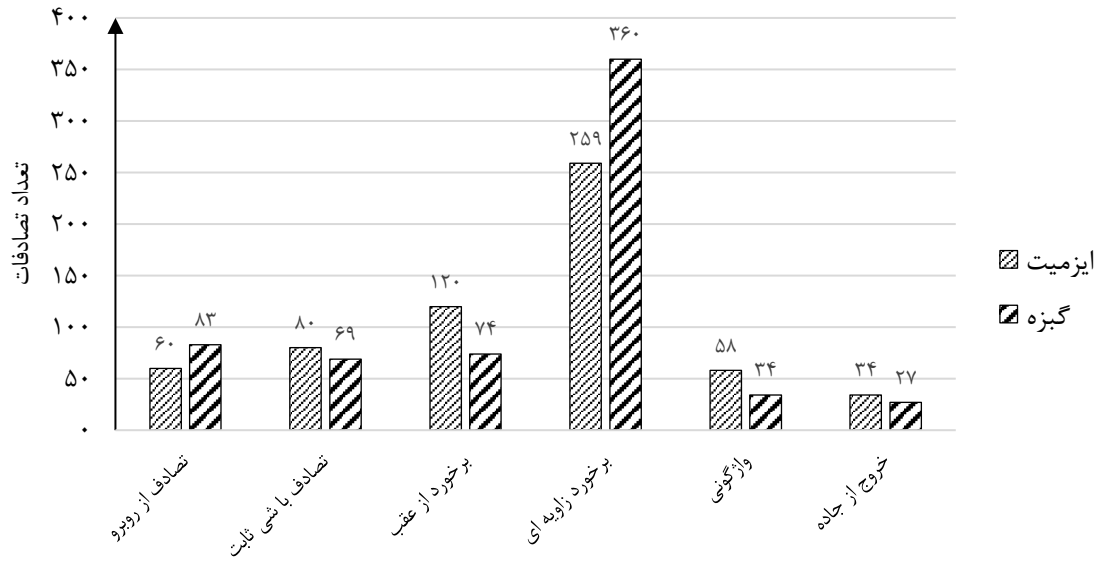
متغیر های عددی		کمینه	بیشینه	میانگین	انحراف معیار
تعداد مجروحین در هر تصادف	گبزه	۱	۱۷	۱/۷۵۸	۱/۴۱۱
	ایزمیت	۱	۹	۱/۶۹۹	۱/۱۷۷
تعداد خودروی درگیر در هر تصادف	گبزه	۱	۴	۱/۹۴۶	۰/۵۴۲
	ایزمیت	۱	۶	۱/۹۵۳	۰/۷۴۲

جدول ۳- ۳: داده‌های توصیفی تصادفات جرحی درون‌شهری شهر گبزه و ایزمیت برای سال ۲۰۱۳-۲۰۱۴

			گبزه	ایزمیت
ردیف	متغیرهای توصیفی	زیرشاخه	فراوانی نسبی ٪(مطلق)	فراوانی نسبی ٪(مطلق)
۱	نوع راه	جداشده	٪۴۸/۷ (۳۱۵)	٪۵۵/۱ (۳۳۷)
		یک‌طرفه	٪۶/۵ (۴۲)	٪۹/۳ (۵۷)
		دوطرفه جدا نشده	٪۴۴/۸ (۲۹۰)	٪۳۵/۶ (۲۱۷)
۲	طبقه راه	بزرگراه	٪۷/۵ (۴۹)	٪۶۹/۶ (۴۲۵)
		خیابان	٪۹۲/۵ (۵۹۸)	٪۳۰/۴ (۱۸۶)
۳	وجود گارد ریل	خیر	٪۸۳/۱ (۵۳۸)	٪۷۱/۷ (۴۳۸)
		بله	٪۱۶/۹ (۱۰۹)	٪۲۸/۳ (۱۸۳)
۴	وجود شانه مسیر	خیر	٪۸۵/۵ (۵۵۳)	٪۷۵/۱ (۴۵۲)
		بله	٪۱۴/۵ (۹۴)	٪۲۴/۹ (۱۵۹)
۵	وجود خط‌کشی	خیر	٪۳۳/۴ (۲۱۶)	٪۲۵ (۱۵۳)
		بله	٪۶۶/۶ (۴۳۱)	٪۷۵ (۴۵۸)
۶	علائم راهنمایی رانندگی	خیر	٪۵۵,۷۳ (۳۲۹)	٪۴۸/۵ (۲۹۶)
		بله	٪۴۴,۲۶ (۳۱۸)	٪۵۱/۵ (۳۱۵)
۷	علائم چراغ‌دار	خیر	٪۹۲/۵ (۵۹۹)	٪۸۵ (۵۲۵)
		بله	٪۷/۵ (۴۸)	٪۱۵ (۸۶)
۸	تقاطع	چهارراه	٪۲۵/۶ (۱۶۶)	٪۱۸/۲ (۱۱۱)
		دور از تقاطع	٪۳۵/۱ (۲۲۷)	٪۵۷/۸ (۳۵۳)
		میدان	٪۸/۳ (۵۴)	٪۳ (۱۸)
		تقاطع T شکل	٪۲۷/۲ (۱۷۶)	٪۱۵/۸ (۹۷)
		تقاطع Y شکل	٪۴,۸۸ (۲۴)	٪۵/۲ (۳۲)
۹	محل تصادف	میانه راه	٪۲/۹ (۱۹)	٪۶/۲ (۳۸)

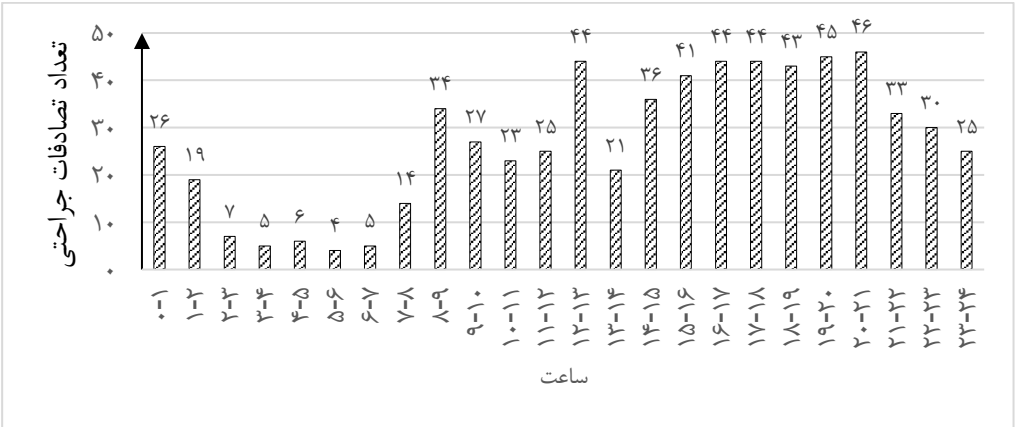
		روی مسیر	٪۸۷/۹ (۵۶۹)	٪۸۱/۸ (۵۰۰)
		در شانه راه	٪۱/۶ (۱۰)	٪۲/۴ (۱۵)
		در کناره راه	٪۴/۳ (۲۸)	٪۵/۵ (۳۳)
		در پیاده‌رو	٪۳/۳ (۲۱)	٪۴/۱ (۲۵)
۱۰	روز و شب	روز	٪۴۸/۷ (۳۱۵)	٪۵۵/۳ (۳۳۸)
		شب	٪۵۱/۳ (۳۳۲)	٪۴۴/۷ (۲۷۳)
۱۱	آب‌وهوا	هوای صاف	٪۸۶/۲ (۵۵۸)	٪۸۳/۷ (۵۱۱)
		بارانی/ برفی	٪۱۳/۸ (۸۹)	٪۱۶/۳ (۱۰۰)
۱۲	وضعیت سطح جاده	خشک	٪۸۱/۶ (۵۲۸)	٪۷۸/۵ (۴۸۰)
		خیس	٪۱۸/۴ (۱۱۹)	٪۲۱/۵ (۱۳۱)
۱۳	هندسه افقی مسیر	قوس	٪۱۴/۷ (۹۵)	٪۱۶/۲ (۹۹)
		مسیر مستقیم	٪۸۵/۳ (۵۵۲)	٪۸۳/۸ (۵۱۲)
۱۴	هندسه قائم مسیر	شیب‌دار	٪۳۵/۵ (۲۳۰)	٪۲۵/۳ (۱۵۵)
		شیب صفر	٪۶۴/۵ (۴۱۷)	٪۷۴/۷ (۴۵۶)
۱۵	روشنایی مسیر	خیر	٪۴۳/۴ (۲۸۱)	٪۲۰/۶ (۱۲۶)
		بله	٪۵۶/۶ (۳۶۶)	٪۷۹/۴ (۴۸۵)
۱۶	نوع تصادف	تصادف از روبرو	٪۱۲/۸ (۸۳)	٪۹/۸ (۶۰)
		برخورد با شیء ثابت	٪۱۰/۷ (۶۹)	٪۱۳/۱ (۸۰)
		برخورد از عقب	٪۱۱/۴ (۷۴)	٪۱۹/۹ (۱۲۰)
		برخورد زاویه‌ای	٪۵۵/۶ (۳۶۰)	٪۴۲/۴ (۲۵۹)
		واژگونی	٪۵/۳(۳۴)	٪۹/۵ (۵۸)
		خروج از جاده	٪۴/۲(۲۷)	٪۵/۵ (۳۴)
۱۷	تعداد خودرو	چند خودرویی	٪۸۲/۷ (۵۳۵)	٪۷۶ (۴۶۴)
		تک خودرویی	٪۱۷/۳ (۱۱۲)	٪۲۴ (۱۴۷)

انواع تصادف در داده های موجود را می‌توان در شکل ۳-۷ مشاهده کرد. همانطور که قابل ملاحظه است تصادف از نوع برخورد زاویه‌ای از بیشترین تصادفات است چرا که در تقاطعات بیشترین نوع برخورد، برخورد زاویه‌ای است. برای در نظر داشتن معناداری داده ها آزمون خی دو بر روی هر کدام از پارامترها انجام شد.که در جدول ۳-۴ آمده است.

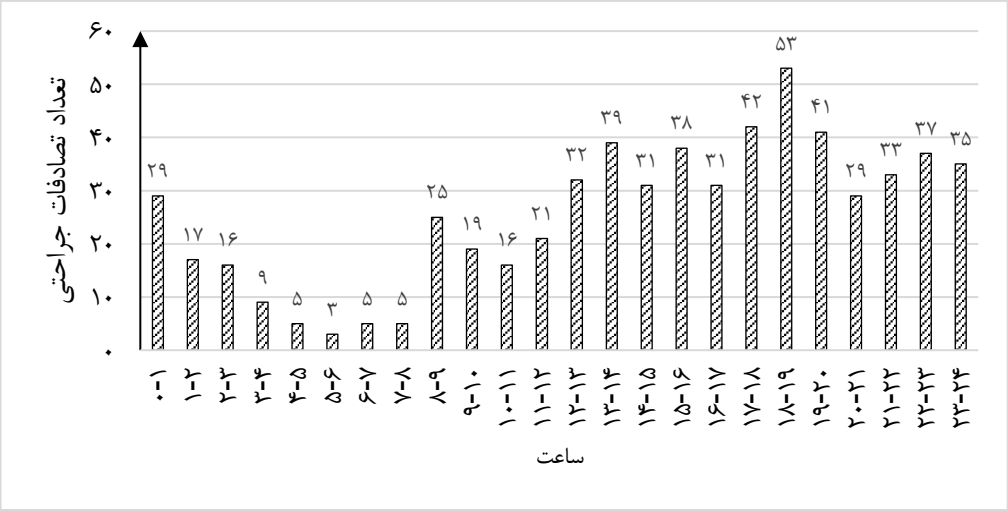


شکل۳- ۷:انواع تصادف در مناطق درون‌شهری به تفکیک شهر ایزمیت و گبزه در سالهای ۲۰۱۳-۲۰۱۴

علاوه بر نوع تصادف داده ها، زمان وقوع تصادف نیز مورد توجه قرار گرفته است. در شکل ۳-۸ و ۳-۹ توزیع ساعتی اطلاعات مربوط به تصادفات جراحتی دو شهر نشان داده شده است. طبق این گزارش، تعداد حوادث به خصوص در ساعات شب بسیار زیاد است. از آنجایی که مطالعه در حوادث ساعت ها دشوار است، تصادفات محاسبه می شود و با توجه به ساعتهای پیک و غیر اوج، دوره های زمانی محاسبه می شوند.

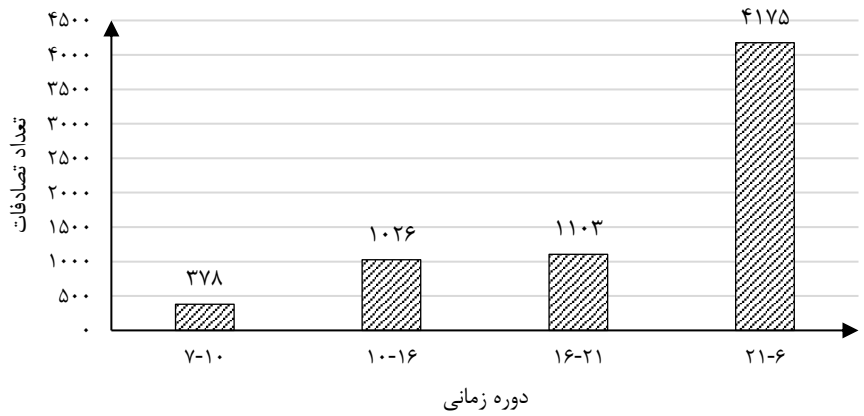


شکل۳- ۸: شکل۳- ۹: توزیع زمانی تصادفات جراحتی شهر گبزه(۶۴۷) در سال۲۰۱۳-۲۰۱۴



شکل۳- ۱۰: توزیع زمانی تصادفات جراحتی شهر ایزمیت(۶۱۱) در سال ۲۰۱۳-۲۰۱۴

با توجه به زمان وقوع تصادفات جراحتی در دوره های زمانی مختلف در شکل ۳-۱۰، از تعداد کل تصادفات ۶۶۹۸ ؛ بین ساعت ۷ تا ۹ صبح ۸۶۱(۱۲/۸ درصد) تصادف مشاهده شد که کمترین مقدار تصادفات درساعت ترافیک اوج صبح بوده است. بین ۱۰ صبح تا ۳ بعد از ظهر ۲۰۹۶ (۳۱/۱ درصد)، بین ساعت ۴ بعد از ظهر تا ۸ شب ۲۲۶۱ (۳۳/۵ درصد) و بین ساعت ۹ شب تا ۶ صبح ۱۵۳۲(۲۲/۷ درصد) دیده شد. تعداد تصادفات مربوط به هر کدام از دوره‌های زمانی از یکدیگر متفاوت هستند. این تفاوت با کمک آزمون Chi-Square مورد آزمایش قرار گرفت. براساس نتایج، توزیع حوادث مربوط به دوره‌های زمانی از لحاظ آماری معنی دار هستند.



شکل۳- ۱۱: تعداد تصادفات به تفکیک دوره های زمانی در

در طول دو سال ۲۰۱۳-۲۰۱۴ از ۶۶۹۸ تصادفات در این استان تعداد ۱۱۴۷۴ نفر مجروح شده اند. از این تعداد مجروح ۱۲۷۳ نفر تصادف با عابر پیاده بود که به علت کم بودن جامعه آماری نسبت به تعداد کل مقدار قابل توجهی برای بررسی تصادفات عابرین پیاده نبود.

شناسایی نقاط حادثه‌خیز می‌تواند هم با کل داده‌های تصادف، هم با داده‌های تفکیک شده بر حسب نوع تصادف یا تعداد افراد کشته یا مجروح انجام شود. نقاط حادثه‌خیز شناسایی‌شده به‌صورت نقشه موضوعی با تعداد افراد کشته یا مجروح که خروجی خوشه‌بندی های مختلف به همراه نقطه حادثه‌خیز می‌باشد. ازآنجایی‌که روش‌های تحلیل موجود در GIS از روش‌های پرکاربرد مورداستفاده در پژوهش‌های اخیر مربوط به شناسایی نقاط حادثه‌خیز می‌باشد نتایج دیگر خوشه‌بندی‌ها باید با این روش مقایسه شود. روش‌های مورد استفاده از نرم افزار ها برای بررسی تراکم و خوشه‌بندی تصادفات به ترتیب زیر است:

- خوشه‌بندی Nnh و Kmeans در نرم افزار CrimeStat
- خوشه بندی GriDBSACN در نرم‌افزار Elki
- روش تراکم کرنل صفحه‌ای و شبکه ای، در نرم افزار GIS

بعد از به دست آمدن خوشه‌ها، برای ارزیابی عوامل مؤثر بر تصادفات می‌توان با آزمون‌های آماری همچون ANOVA معنی‌داری داده‌های خوشه بندی را بررسی نمود.

۳-۵-خوشه‌بندی Nnh و K-Means در نرم افزار CrimeStat

همانطور که در فصل دوم ذکر شده است، تکنیک های بسیار زیادی وجود دارد که در آن ایده اصلی محاسبه تعدادی نقاط یا داده مانند تصادفات در یک منطقه محدود که توسط محقق تعیین شده است، می‌باشد(لوین،۲۰۱۰). در این مطالعه، روش تحلیل خوشه ای Nnh برای تشخیص نقاط حادثه‌خیز

تصادف استفاده می شود که بر شناسایی گروه هایی از داده‌ها که به صورت مکانی نزدیک هستند، متمرکز می‌شود. در خوشه بندی Nnh، فاصله اقلیدسی بین هر جفت نقطه داده بررسی و به عنوان یک معیار برای خوشه بندی استفاده می شود. معیار اول فاصله جستجو (d) است که پس از تعیین آن، نقاط با فاصله‌های کوچکتر در یک خوشه قرار می‌گیرند. معیار دوم حداقل تعداد نقاط n_{min} در خوشه می-باشد. سپس نقاط منطبق با معیارهای فاصله جستجو (d) و n_{min} به عنوان یک خوشه شناسایی می‌شود. اگر تعداد نقاط تعریف شده نزدیکتر از فاصله جستجو باشد، یک خوشه جدید تولید می شود. پس از محاسبه خوشه های مرتبه اول، خوشه های مرتبه دوم و مرتبه بالا با همان شیوه تشکیل می شوند تا زمانی که فقط یک خوشه باقی بماند یا با معیارهای مذکور تطابق نداشته باشد. به همین دلیل، الگوریتم خوشه بندی Nnh خوشه ای از تمام نقاط در منطقه مورد مطالعه ایجاد نمی کند. از آنجایی که این روش به معیارهای مربوط به فاصله جستجو (d) و n_{min} بستگی دارد، تنها نقاطی که معیارهای مورد نیاز را برآورده می کنند، خوشه بندی می شوند.

همانطور که در بالا ذکر شد، چندین روش تحلیل مکانی وجود دارد که می تواند مورد استفاده قرار گیرد. یکی دیگر از روش‌های شناخته‌شده تحلیل خوشه‌ای نقاط ، خوشه بندی K-Means است که داده‌ها را به گروه های K که توسط کاربر تعریف شده است، تقسیم می‌کند. این روش K تعداد خوشه از تمام داده ها ایجاد می‌کند. تمام نقاط به یک گروه اختصاص داده می شوند و تمام نقاط در این روش خوشه بندی می‌شوند و در یک خوشه قرار می‌گیرند. این روش زمانی می تواند مفید باشد که می‌خواهیم گروه بندی تمام داده‌ها را کنترل کنیم. برای مثال، اگر قرار است ۵ منطقه از بهترین مناطق ایستگاه آمبولانس را پیدا کنیم، روش خوشه‌بندی K-means با تعریف ۵ خوشه یک ایده در مورد متراکم ترین مناطق می دهد. برخلاف روش خوشه بندی K-Means، الگوریتم خوشه بندی Nnh امکان کنترل اندازه گروه بندی با تعیین فاصله آستانه و n_{min} در خوشه را فراهم می کند. خوشه بندی Nnh فرصتی را

برای شناسایی محدوده جغرافیایی متراکم داده‌ها فراهم می آورد. این مزیت الگوریتم Nnh، شناسایی نقاط حادثه‌خیز را برای هدف این مطالعه فراهم می کند.

۳-۵-۱-نحوه کار در نرم‌افزار CrimeStat

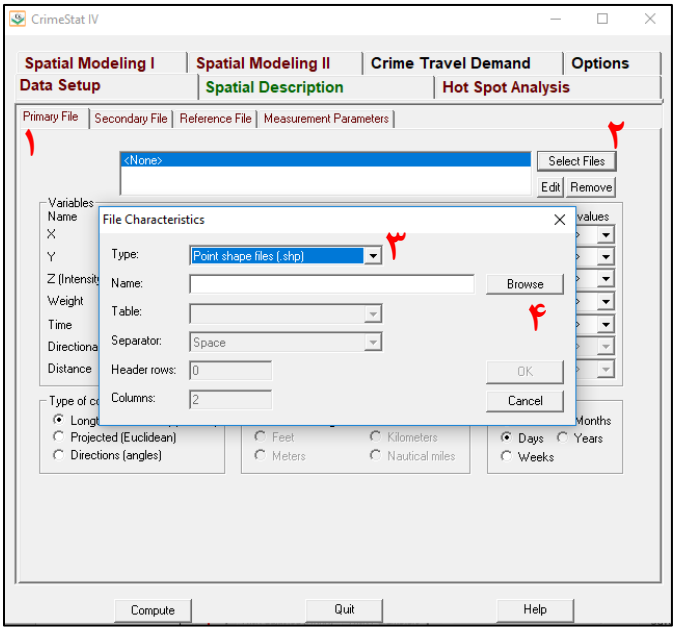
اجرای الگوریتم Nnh در نرم افزاری به نام CrimeStat IV قابل اجرا است. با توجـــه به شـکل۳-۱۲ مراحل اجرای نرم افزار بدین ترتیب است:

۱-کلیک برروی سربرگ Primary File

۲- سپس بر روی Select Files کلیک می‌کنیم.

۳-سپس در پنجره باز شده File Characteristics در قسمت Type نوع Point shape files (.shp) را انتخاب می‌کنیم

۴-سپس از Browse فایل داده ها به فرمت .shp وارد نرم افزار می‌شود.



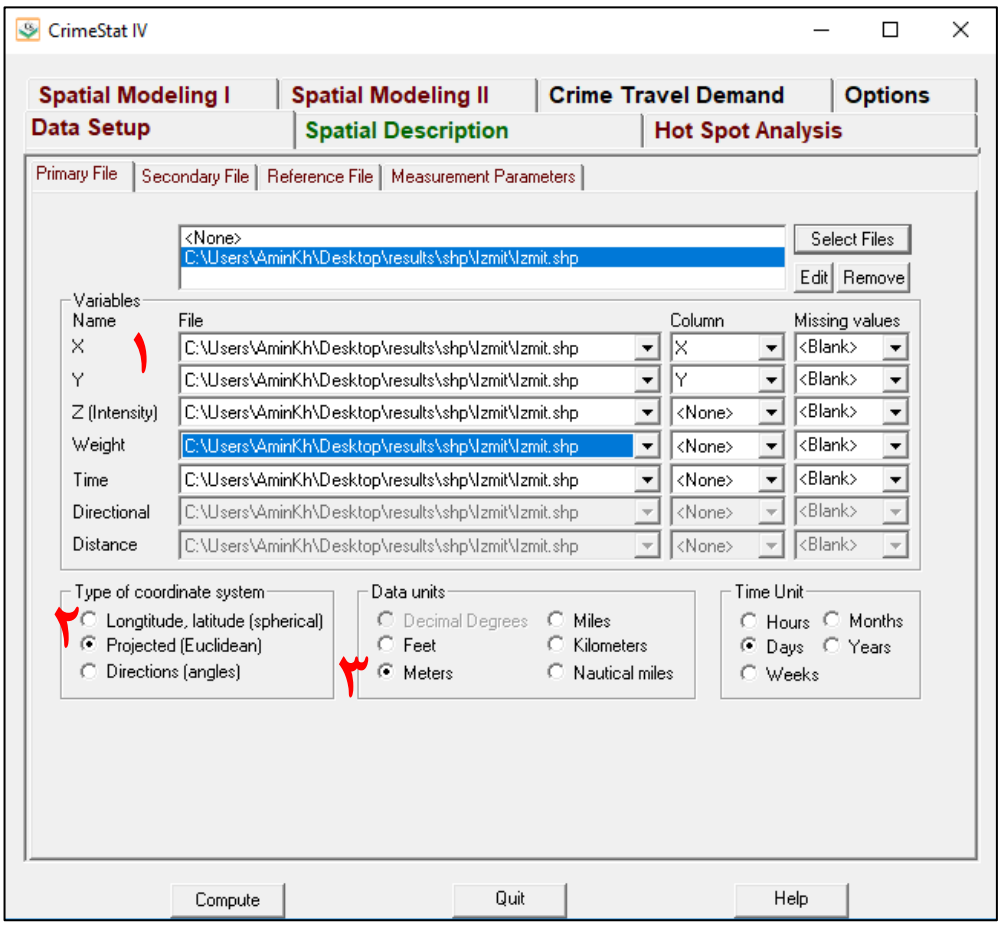
شکل۳- ۱۲: مرحله اولیه وارد کردن فایل shp داده‌های تصادف

در ادامه باتوجه به شکل ۳-۱۳ به ترتیب:

۱- مشخصات جغرافیایی X و Y را در قسمت Column انتخاب می‌کنیم.

۲- سپس در قسمت Type of coordinate system نوع سیستم مختصات projected(Euclidean) یا همان مختصات اقلیدسی که به از نوع UTM می‌باشد را انتخاب می‌کنیم.

۳- در قسمت Data units واحد طول متر را انتخاب می‌کنیم.

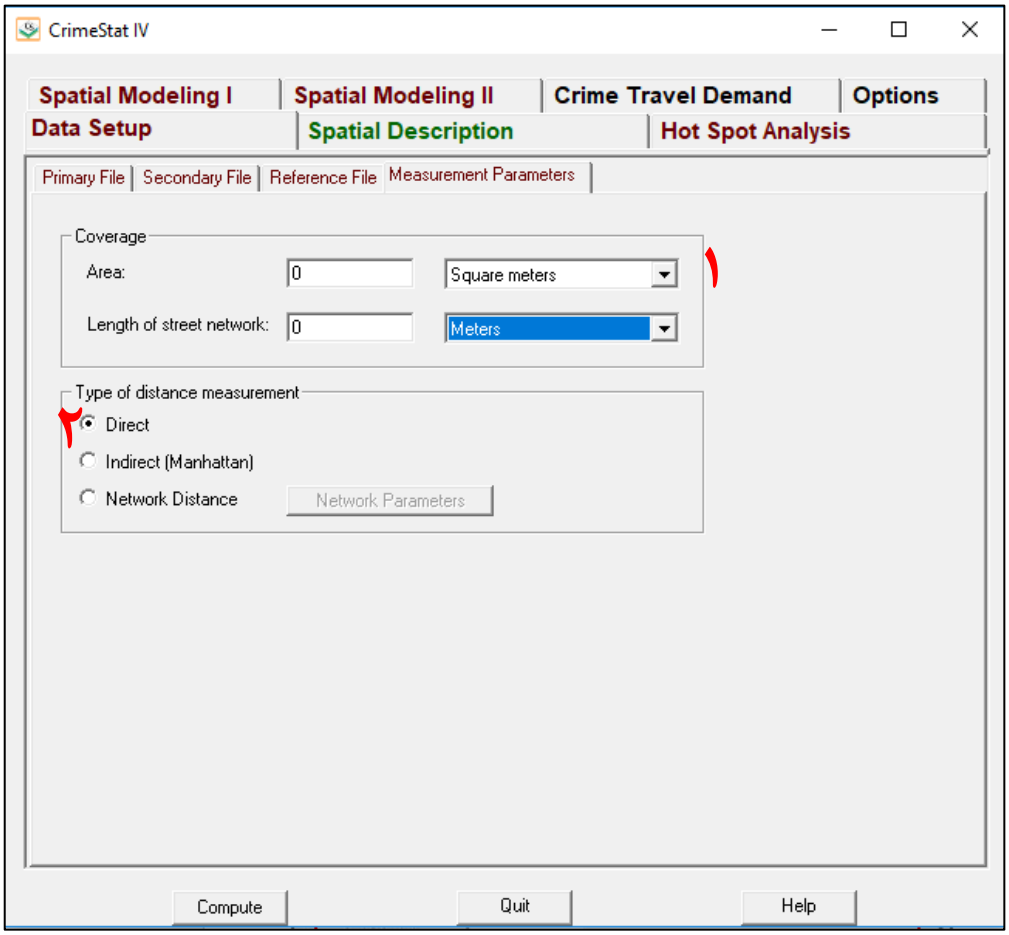


شکل۳- ۱۳: وارد کردن مشخصات جغرافیایی

با توج به شکل ۳-۱۴ در سربرگ Measurement Parameters داریم:

۱- واحد مساحت را متر مبرع انتخاب می‌کنیم.

۲- نوع اندازه گیری فاصله را Direct که همان نوع اقلیدسی است اتخاب می‌کنیم.



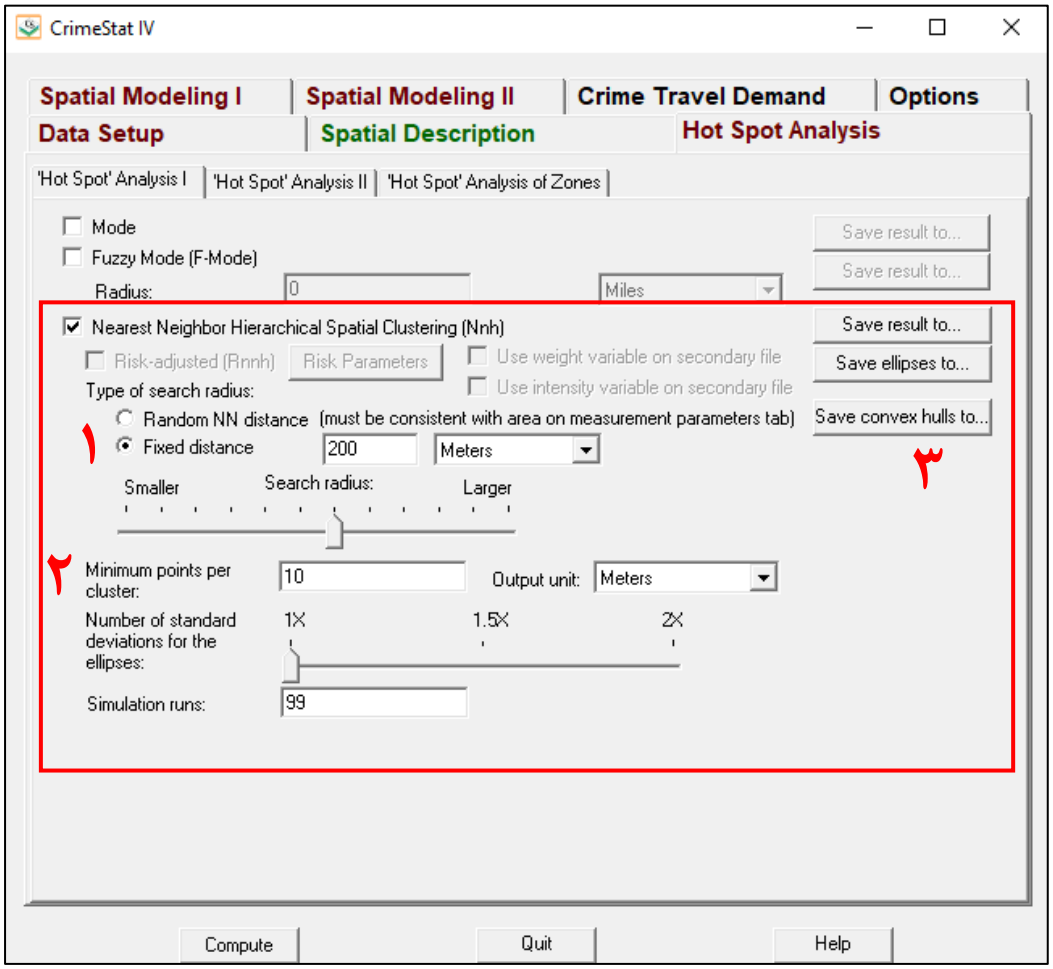
شکل۳- ۱۴: پارامترهای اندازه‌گیری

در سریرگ Hot Spot Analysis در سربرگ اول Hot Spot Analysis I خوشه‌بندی Nnh را فعال وا به ترتیب پارامترها را وارد می‌کنیم. باتوجه به شکل ۳-۱۵ داریم:

۱- انتخاب Fixed distance که همان پارامتر d یا همان شعاع جستجو خوشه است و در قسمت کناری واحد متر را برای شعاع در نظر می‌گیریم.

۲- در قسمت Minimum points per cluster که همان n_{min} می‌باشد را وارد می‌کنیم.

۳- در سمت راست سه نوع ذخیره خروجی نتایج وجود دارد. گزینه اول به صورت جدولی است مختصات نقاط مرکز خوشه بدست آمده از الگوریتم را می‌توان در محیط GIS مشاهده نمود. گزینه دوم Save Ellipse to… برای خروجی خوشه‌ها به صورت بیضی شکل و گزینه سوم Save convex hull to… به صورت چندضلعی حاصل از وصل نقاط مرزی هر خوشه است.



شکل۳-۱۵: محیط نرم‌افزار CrimeStat IV برای محاسبه خوشه‌بندی NNH

۳-۶-خوشه بندی GriDBSCAN

همانطور که در فصل دوم اشاره شد،در این مرحله ما باید مقادیر مختلف Eps را برای هر سطح چگالی پیدا کنیم. سپس ما از این مقادیر Eps برای خوشه بندی برای پیدا کردن تعداد صحیح خوشه

ها استفاده خواهیم کرد. اولا، برای تعیین پارامترهای Eps و MinPts، باید رفتار فاصله را از نقطه به نزدیکترین همسایه k-k، یعنی k-dist ببینیم. هر کدام از k-dist برای تمام نقاط داده به ازای برخی از مقادیر (k) محاسبه می شود. پس از آن انتظار می رود که تغییر ناگهانی در گراف به دست آید. این تغییر شدید در مقدار k-dist با مقدار مناسب Eps برای هر سطح چگالی مجموعه داده مطابقت دارد.که خط افقی عمود بر محور عمودی مقدار Eps را به ما می‌دهد. نمودار K-dist برای انتخاب نه تنها پارامترهای Eps، بلکه تجزیه و تحلیل سطح چگالی مجموعه داده نیز کشیده می شود. توجه داشته باشید که مقدار Eps که در این روش تعیین می شود بستگی به k دارد، اما با تغییر k مقدار Eps به طور چشمگیری تغییر نمی‌کند(گائونکار و ساوانت،۲۰۱۳).

اگر مجموعه داده ها با تراکم متنوعی باشد، بستگی به تراکم خوشه و توزیع تصادفی نقاط، نقاط دارای سطح چگالی یکسان ، محدوده‌ی تنوع بزرگی نخواهد داشت، در حالی که تغییرات شدیدی انتظار می رود بین دو سطح چگالی دیده‌شود. به این ترتیب، منحنی های صاف و متفاوتی وجود خواهد داشت.

۳-۶-۱-حذف نویزها:

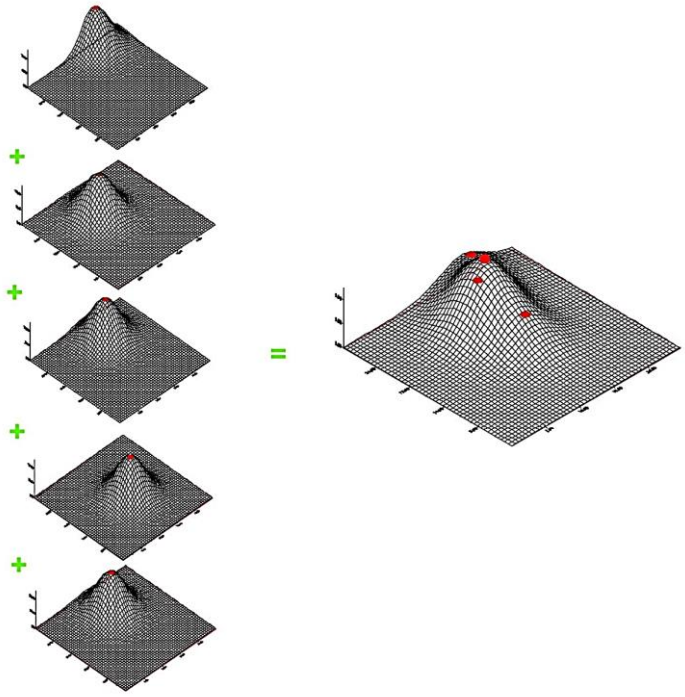
هر مجموعه داده ای تقریبا همیشه حاوی نویز است. این داده‌ها به هیچ یک از خوشه ها تعلق ندارند. به این معنی که محدوده های ناخوشه‌ها به طور کلی نسبت به نقاط خوشه ها پراکنده هستند، و فاصله یک ناخوشه به نزدیکترین خوشه فاصله زیادی نسبت به فاصله نقاط خوشه ها در هر خوشه است. هر روش خوشه بندی نیاز به مکانیسمی برای از بین بردن خوشه‌ها دارد.

در الگوریتم DBSCAN ۸۸، ناخوشه‌ها با توجه به فاصله‌ی بیشترشان از نقاط دیگر، کمتر تمایل دارند با نقاط دیگر ادغام شوند و رشد کمتری نسبت به خوشه های واقعی داشته باشند. بنابراین خوشه هایی که به آرامی در حال رشد هستند شناسایی و به عنوان ناخوشه از خوشه‌بندی حذف می‌شوند. همچنین، از آنجایی که تعداد نقاط در مجموعه ای از نویزها معمولا کمتر از تعداد در یک

خوشه است و این نویزها خوشه های بسیار کوچکی هستند، ما می توانیم به راحتی چنین گروه های کوچکی را شناسایی کرده و آن‌ها را حذف کنیم. به این ترتیب، مرحله نهایی، از بین رفتن خروجی، گام لازم برای خوشه بندی خوب است. در الگوریتم DBSCAN این اتفاق می‌افتد و نقاط نویزبه راحتی مشخص می‌شوند.

۳-۷-روش تراکم کرنل صفحه‌ای و شبکه ای، در نرم افزار GIS

KDE یک تکنیک درونیابی است که در شناسایی نقاط حادثه‌خیز مورد استفاده قرار می گیرد. این روش براساس یک تابع چگالی نقطه‌ای است که برای تمام سلول‌های یک شبکه (محدوده مطالعه) محاسبه می شود. محاسبه KDE در GIS بدین ترتیب است که ابتدا محدوده مورد مطالعه باید تعریف شود. سپس تفکیک فضایی سلول شبکه ای (یا با فاصله سلول یا تعداد ستون ها) تعریف می شود. تابع کرنل دقیقا بر روی هر نقطه تصادف قرار گرفته و سطح متقارن را تشکیل می‌دهد. سطح تابع چگالی با توجه به مقدار شعاع جستجو توسز کاربرتعریف می‌شود. برای هر سلول در شبکه مشخص شده (محدوده مورد مطالعه)، این روش ارزش تمام سطوح را برای این سلول مرجع، جمع می‌کند (لوین،۲۰۱۰). در شکل ۳-۱۲، خلاصه ای از روش محاسبه KDE را می توان مشاهده کرد.



شکل۳- ۱۶: جمع‌بندی سطوح کرنل نرمال برای ۵ نقطه(لوین،۲۰۱۰)

۳-۸-مقایسه روش های خوشه بندی

۳-۸-۱-شاخص دقت پیش‌بینی (PAI)

همانطور که در پیشینه تحقیق توضیح داده شد، تکنیک های بسیاری برای شناسایی ناحیه های حادثه‌خیز تصادف وجود دارد. با این وجود، وقتی تصمیم به استفاده از این تکنیک ها می‌شود، انتخاب روش بهتر مشکل می شود. تکنیک های مختلف شناسایی نقاط حادثه‌خیز نتیجه ی مختلفی از لحاظ اندازه، شکل و مکان را ارائه می‌دهند. به همین دلیل نتایج تکنیک‌های خوشه‌بندی داده‌های نقطه‌ای باید با یکدیگر مقایسه شوند. با این حال، هیچ دانش دقیق و یا تعریف دقیق در مورد تکنیک شناسایی نقطه حادثه‌خیز برای تشخیص حوادث با اهمیت بالا وجود ندارد. با توجه به این موضوع، توسط چی نی و همکارانش[1] (۲۰۰۸) برای تعیین اینکه آیا تفاوتی برای لحاظ کردن یا پیش‌بینی نقاط حادثه‌خیز وجود

[1] Chainey et al.

دارد یا نه PAI[1] را مطرح کرد. این روش در نقشه برداری شناسایی نقاط جرم برای مقایسه دو تکنیک درنظر گرفتن و پیش‌بینی این نقاط استفاده شده است. روش شناسی بستگی به نسبت درصد امتیازات در نقاط حادثه‌خیز در تمام نقاط در کل منطقه مورد مطالعه به درصد از نقاط داغ مشخص شده در تمام منطقه مورد مطالعه است. درصد امتیاز در نقاط داغ در تمام نقاط منطقه مورد مطالعه به عنوان نرخ ضربه[2](HR) نامیده می شود. به طور مثال، پاترن و همکارانش (۲۰۰۹) PAI را به عنوان یک نرخ درصدی، یعنی درصد ضربه به درصد منطقه برای ۲۰ نقطه حادثه‌خیز شناخته شده توصیف کردند.

توابع ریاضی درصد ضربه (معادله) و PAI در زیر آمده است :

$$\mathrm{HR} = \left(\frac{n}{N}\right) * 100 \qquad (۱-۳)$$

$$\mathrm{PAI} = \left(\frac{(\frac{n}{N})*100}{(\frac{a}{A})*100}\right) \qquad (۲-۳)$$

که در آن n تعداد حوادث در نقطه داغ و N تعداد حوادث ناحیه مطالعه است، A مساحت نقطه داغ است و A منطقه ناحیه مطالعه است. PAI اندازه گیری شده براساس داده های در طول زمان است و معیاری است که توسط آن تکنیک های شناسایی نقاط حادثه‌خیز می توانند با دیگر مقایسه شوند.

۳-۸-۲-روش ANOVA

تجزیه و تحلیل واریانس یک طرفه (ANOVA) یک روش آماری پارامتری است که یک تحلیل یکنواخت واریانس را برای یک متغیر وابسته توسط یک عامل واحد (مستقل) تولید می کند و به نسبت F بستگی دارد. به طور عمده برای مقایسه اختلافات مهم بین دو یا چند گروه وابسته انتخاب شده استفاده می شود. بر خلاف فرض صفر که کل جمعیت به معنی برابر است، می توان ادعا کرد که حداقل

[1] Prediction Accuracy Index
[2] Hit Rate

میانگین یک گروه متفاوت است. اگر این مقدار به طور قابل توجهی متفاوت باشد، این گروه ها می توانند از طریق تست دامنه post-hoc و مقایسه چندگانه عددی مرتب شوند. در نرم افزار SPSS، یک ابزار آماری که معمولا مورد استفاده قرار می گیرد، تست های دامنه ها، زیرمجموعه های همگن را که از یکدیگر متفاوت نیست، تشخیص می دهند و در سطح آلفا ۰,۰۵ به طور معناداری متفاوت هستند.

۳-۸-۳-ضریب همبستگی اتا

هنگامی که یک متغیر دارای مقیاس اسمی و رتبه ای (مثل جنس، نژاد، میزان رضایت و ...) و متغیر دیگر مقیاس فاصله ای یا رتبه ای داشته باشد مثل درآمد، معدل، اندازه، تعداد و ... باید شاخصی انتخاب شود که بتوان از روی یک متغیر، متغیر دیگر را پیش بینی نمود. از جمله شاخص ها در این وضعیت شاخص "نسبت همبستگی" است که با نماد η^2 (مجذور اتا) نشان داده می شود. برای فهم بهتر این ضریب، فرضیه زیر را در نظر داشته‌باشید:

"بین تعداد افراد مجروح و نوع تقاطع ارتباط وجود دارد."

در این فرضیه، نوع تقاطع یک متغیر کیفی با مقیاس اسمی و تعداد افراد مجروح متغیری کمی با مقیاس فاصله ای(نسبی) است . برای توضیح این شاخص باید گفت که این شاخص بر دو مفهوم کلی بنیان گذارده شده است، نخست آن که با فرض ندانستن نوع تقاطع ممکن است چه میزان خطا در پیش بینی تعداد افراد مجروح رخ دهد (e_1) و دوم آن که با فرض دانستن نوع تقاطع، به چه میزان خطای مرحله قبل (e_1) کاهش خواهد یافت. چنان چه ارتباطی بین این دو متغیر وجود داشته باشد می بایست تعداد خطاهای مرحله اول (e_1) کاهش یافته باشد. میزان این ضریب از رابطه زیر به دست می آید:

$$\eta^2 = \frac{e_1 - e_2}{e_1} \qquad (۳-۳)$$

برخلاف متغیرهای اسمی، که در پیش بینی آن ها درست و غلط وجود دارد، در پیش بینی متغیرهای کمی (تعداد افراد مجروح) درست و غلط مطرح نمی شود. چنان چه تعداد واقعی افراد مجروح ۳ نفر باشد، تعداد ۱ و ۴ هر دو اشتباه می باشند ولی پیش بینی تعداد ۲ نزدیک به واقعتر می باشد، به بیان دیگر هر چه $|x_i - \bar{x}|$ کوچک تر باشد میزان خطای کمتری رخ داده است. در آمار به جای $|x_i - \bar{x}|$ معمولا از مجذور خطاها استفاده می شود $((x_i - \bar{x})^2)$، پس می توان گفت که اگر از متغیر نوع تصادفات در مثال فوق هیچ اطلاعی وجود نداشته باشد مجموع مجذور خطاها برابر است با:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \hat{x})^2 + \sum_{i=1}^{n}(\hat{x} - \bar{x})^2 \qquad (۴-۳)$$

$\bar{x}$ مقادیر پیش بینی شده این متغیر و $\hat{x}$ مقدار واقعی متغیر تعداد افراد مجروح، x_iدر این فرمول میانگین آن است. پس می توان گفت که کل مجذور خطا برابر است با:

(۵-۳) میزان تغییر خطا + میزان تغییر پیش بینی شده = کل مجذور خطا

فصل ۴

نتایج

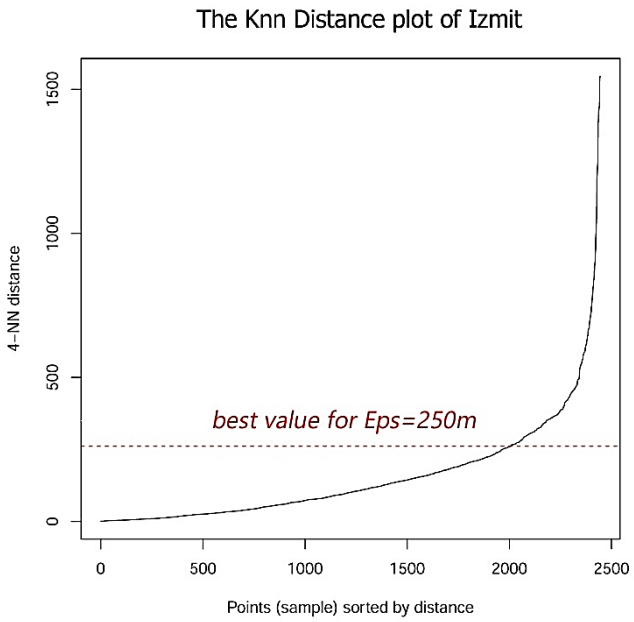
۴-۱- مقدمه

در این فصل به بررسی و تشریح نتایج حاصل از اعمال روش‌های ذکر شده در فصل سوم پرداخته شده است. نتایج برای دو شهر گبزه و ایزمیت به صورت جداگانه مورد تحلیل قرار می‌گیرند. روند تحلیل بدین گونه است که ابتدا نتایج انواع روش‌های ذکر شده برای خوشه بندی را بررسی می‌کنیم و سپس با استفاده از شاخص PAI[1] یا شاخص دقت پیش‌بینی و خوشه‌ها را از لحاظ پارامترهای موجود در آن خوشه بررسی می‌کنیم. هر خوشه را به صورت جداگانه از لحاظ آماری بررسی کرده و داده های معنا دار برای هر خوشه را مورد بررسی و تحلیل قرار می‌دهیم.

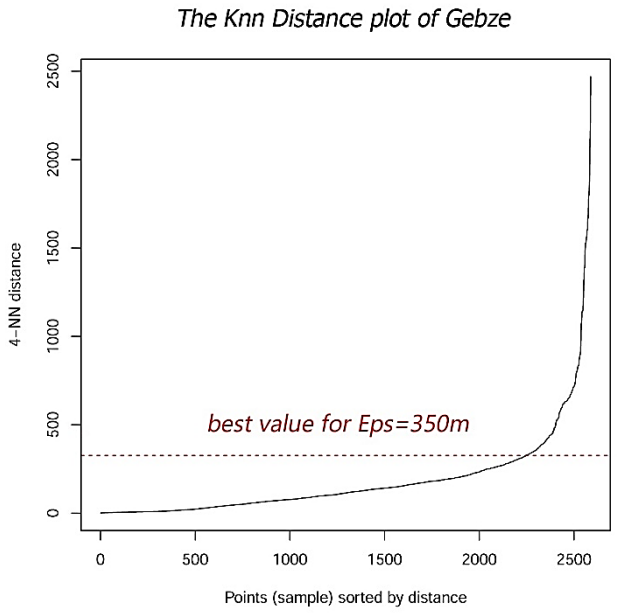
۴-۲- نتایج حاصل از خوشه بندی GriDBSCAN

با توجه به روش ذکرشده در فصل سوم برای هر کدام از شهرها با توجه به نمودار Knndistance در نرم‌افزار Rstudio 1.0.136، برای هر شهر مقادیر بهینه شعاع حداکثر جستجو Eps برای خوشه بندی در نظر گرفته شد(شکل۴-۱). بر اساس تحقیقات استر در سال ۱۹۹۶ میلادی برای داده‌های دوبعدی (x,y) می‌توان k را برابر با ۴ در نظر گرفت چرا که استر ثابت کرد برای داده‌های دو بعدی مقادیر بیشتر از ۴ نمودار تغییری نخواهد کرد. مقدار $d > 2\text{Eps}$ نیز برای شبکه‌بندی با توجه به مطالعه مهران و مهار همانطور که در فصل پیشین به تشریح ذکر شد، در نظر گرفته می‌شود(مهران و مهار[2]،۲۰۰۸).

[1] Pridction Accuracy Index
[2] Mahran & Mahar



(الف)

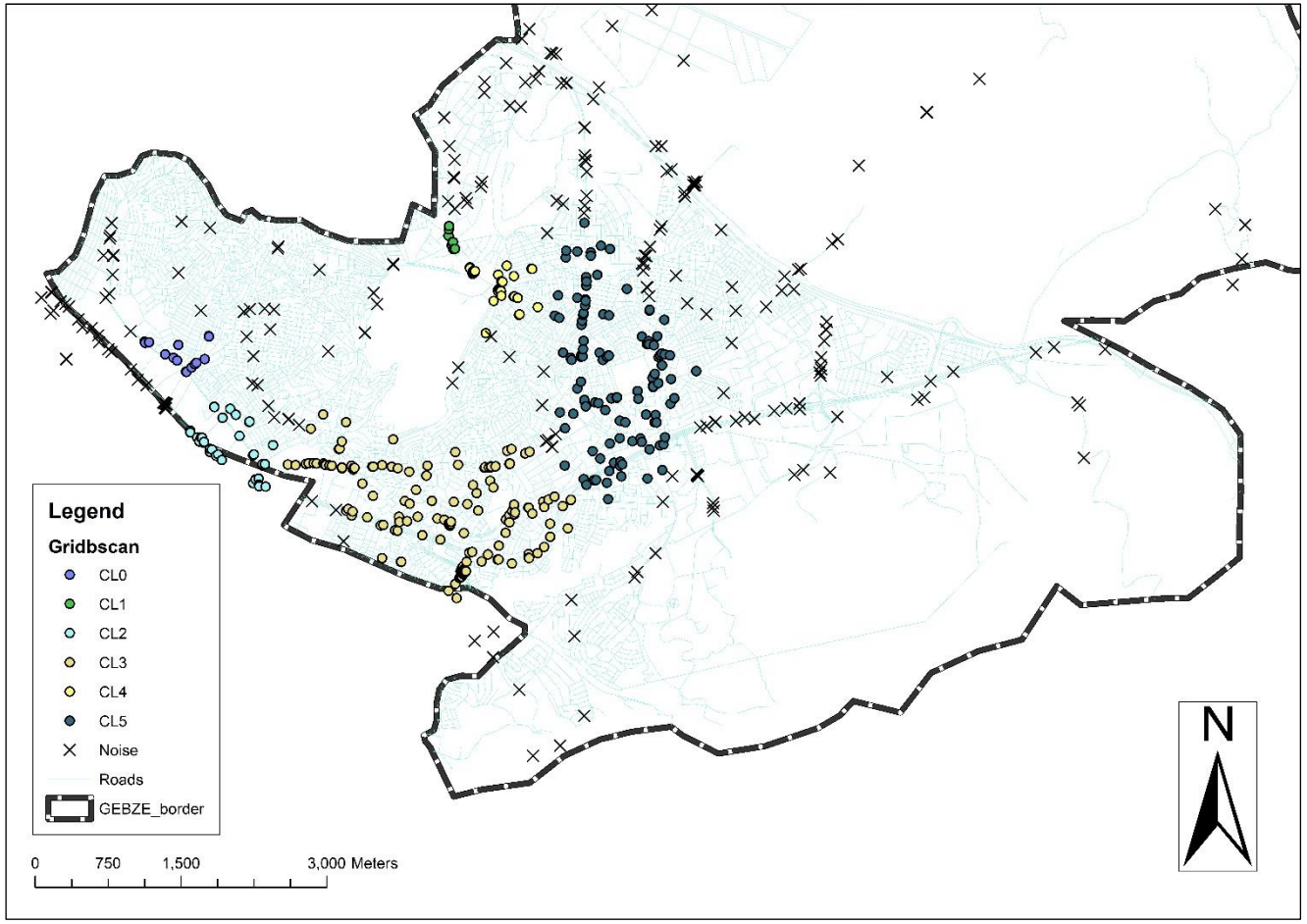


(ب)

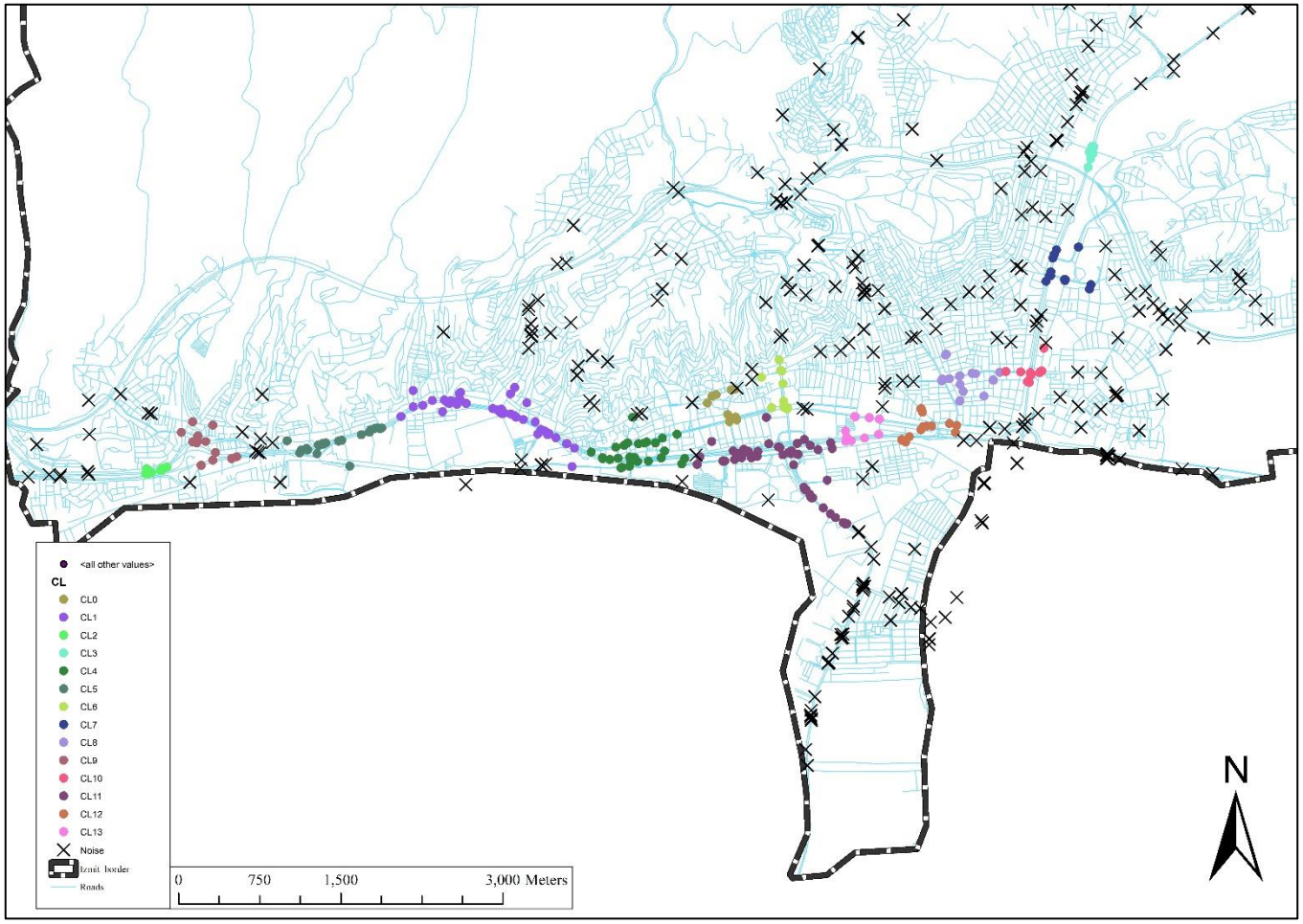
شکل۴- ۱: نمودار Knn با مقدار K=4 و مقدار بهینه Eps(m) برای شهر (الف) ایزمیت (ب) گبزه

همانطور که در شکل۴-۲ و ۴-۳ قابل مشاهده است الگوریتم GriDBSCAN مکان‌هایی که نقاط تصادف در آن تراکم بیشتری نسبت به مکان‌های دیگر دارد را خوشه‌بندی می‌کند و نقاطی که پراکنده هستند به عنوان نقاط نویز یا ناخوشه می‌شناسد.

در ادامه خوشه بندی شهر ایزمیت و گبزه به روش GriDBSCAN در شکل های ۴-۳ و ۴-۴ قابل مشاهده است.



شکل۴- ۲: خوشه بندی GriDBSCAN با مقدار Eps=350m و minPts=15 وd=700m برای شهر گبز

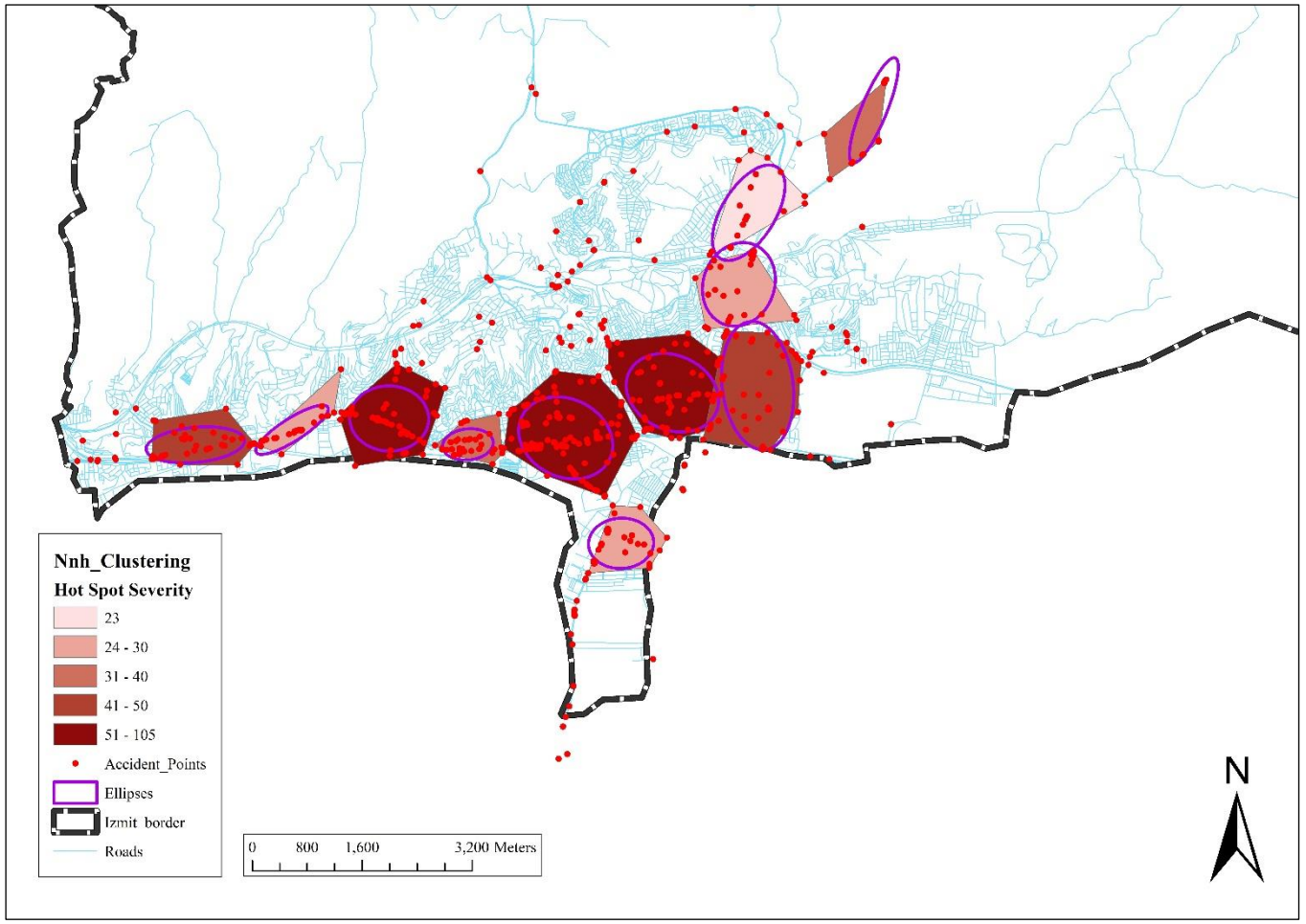


شکل۴- ۳: خوشه بندی GriDBSCAN با مقدار $Eps = 250m$ و $minPts = 10$ و $d = 500m$ برای شهر ایزمیت

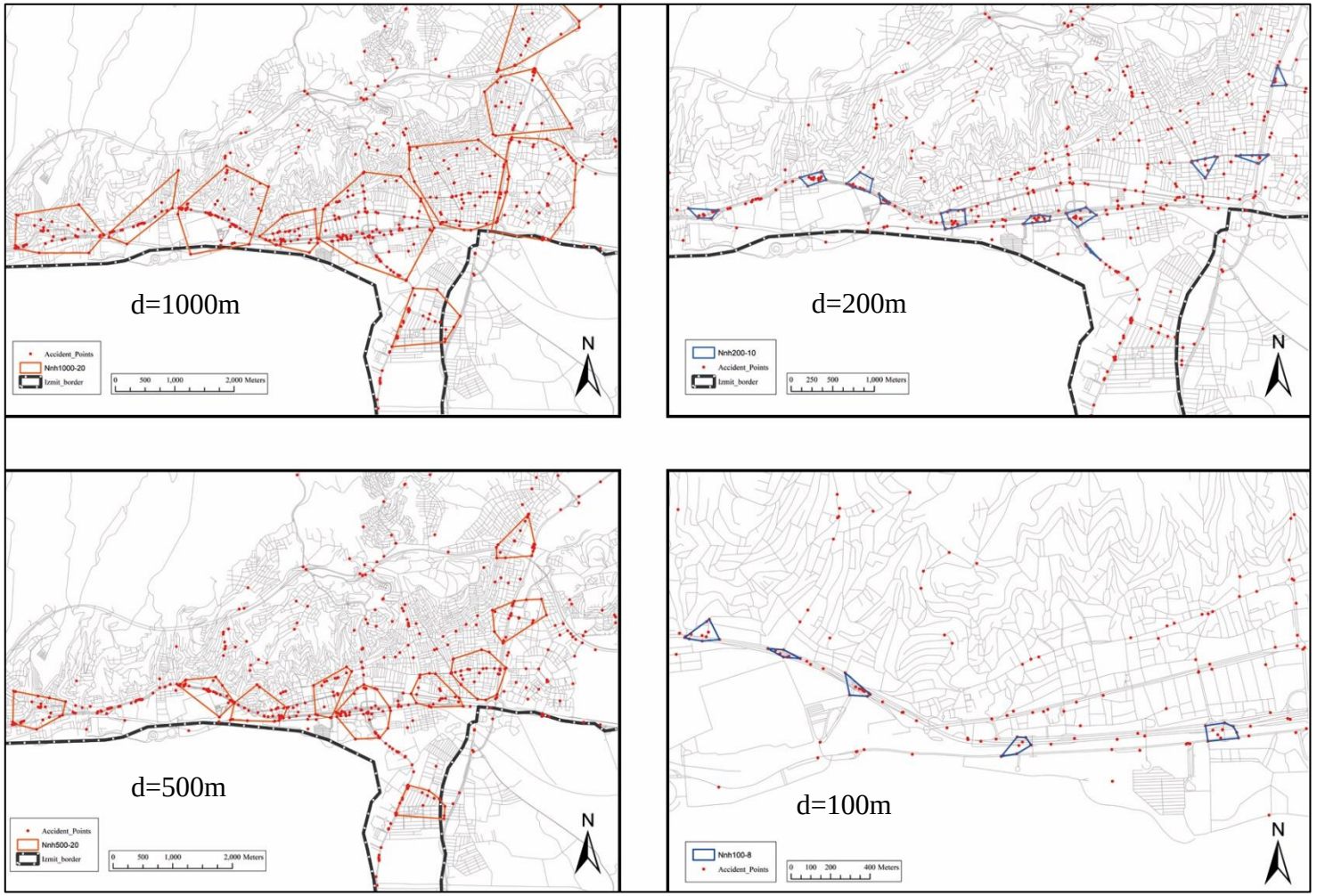
۴-۳-روش خوشه بندی Nnh

همانطور که در روش تحقیق ذکر شد، در این روش تعیین شعاع جستجو و تعداد حداقل نقاط به عهده کاربر می‌باشد. تشخیص نقاط حادثه خیز در این روش با فاصله آغازین ۱ کیلومتر آغاز شده است که در روش های تحلیل نقاط حادثه‌خیز تصادفات فوتی نیز استفاده می شود. برای اینکه تعداد زیادی از تصادفات جراحتی را در نظر بگیریم، ابتدا مقدار n_{min} عدد ۲۰ انتخاب شد. در هنگام ایجاد نقشه برای تصویر سازی نتیجه روش خوشه بندی، از روش نقشه موضوعی با مقادیر شدت نقاط حادثه خیز استفاده می شود. همانطور که در شکل ۴-۴ دیده می شود، روش خوشه بندی Nnh با معیارهای داده شده، خوشه های بسیار بزرگی ایجاد شده است که بیش از یک جاده و تقاطع را پوشش می دهد. به منظور درک بهتر ارتباط تصادفات با اجزای محیط شهری، شناسایی مناطق کوچکتر مناسب‌تر و معنی دار است.در شکل ۴-۴ هر دو نوع خروجی بیضی و چندضلعی قابل مشاهده است که تفاوتشان صرفا در محصور کردن هر خوشه است.

در حین بررسی تصادفات درون‌شهری برای تشخیص نقاط حادثه‌خیز، فاصله جستجو نقاط حادثه-خیز امری مهم به شمار می‌آید. از آنجا که هدف اصلی این پژوهش این است که ویژگی های نقاط تصادف منجر به جراحت در ارتباط با محیط شهری را درک کنیم، فاصله جستجو نقاط حادثه‌خیز بسیار مهم است. همانطور که در شکل ۴-۴ قابل مشاهده است، خوشه های به دست آمده برای نقاط حادثه‌خیز تصادفات با محدوده ۱۰۰۰ متر و مقدار $n_{min} = 20$ بیش از یک جاده را پوشش می دهند. به همین دلیل، فاصله را تا ۵۰۰ متر کاهش دادیم. نقاط حادثه‌خیز شناسایی شده توسط فاصله آستانه ۵۰۰ متر نیز منطقه‌ی بزرگی را پوشش داد. بر این اساس، بار دیگر با ۲۰۰ متر محاسبه کردیم که نتایج بهتری را نسبت به مقادیر قبلی نشان داد.

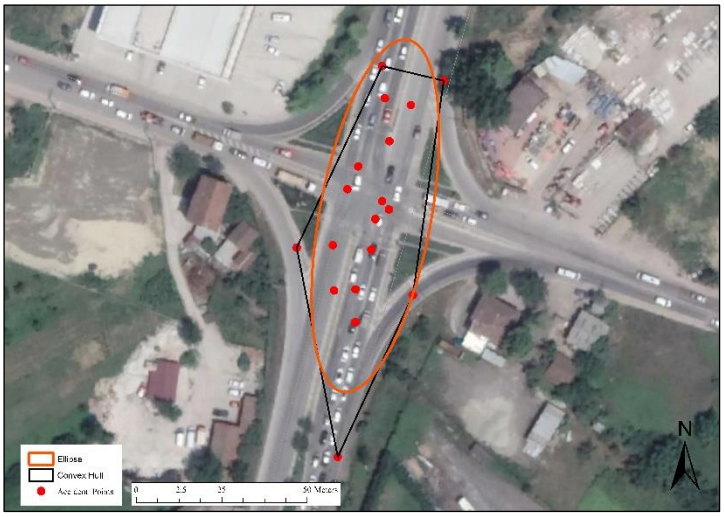


شکل۴- ۴:نقشه موضوعی تصادفات جراحتی با شدت نقاط حادثه خیز به روش Nnh
با مقادیر $n_{min} = 20$ و فاصله جستجو ۱۰۰۰ متر



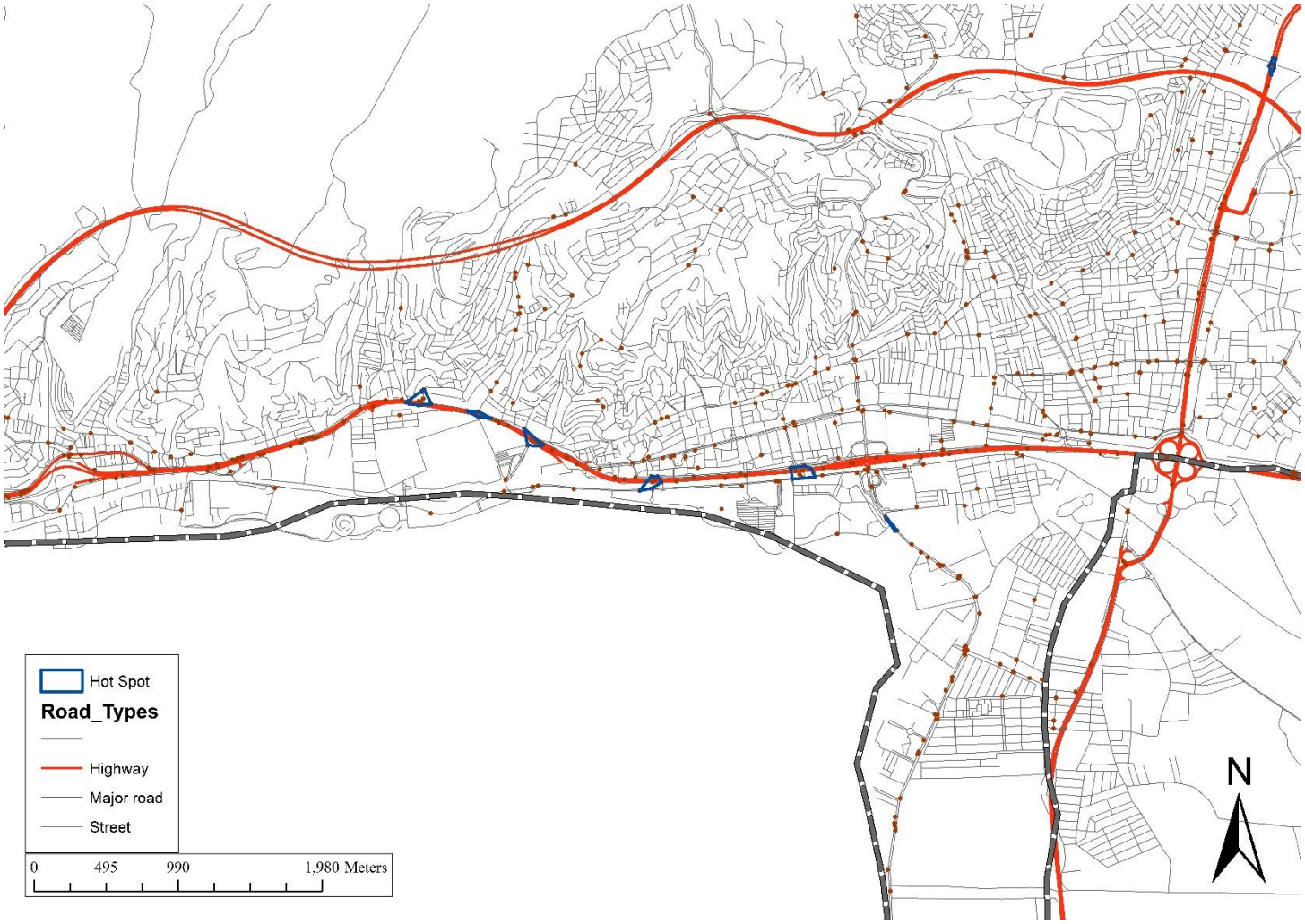
شکل۴- ۵: فواصل جستجو (d) مختلف با $n_{min} = 10$ برای محاسبه خوشه‌بندی Nnh برای تصادفات شهر ایزمیت

با این‌حال، با مقدار ۲۰۰ متر نیز در بعضی از مناطق شهری، بیش از یک جاده را پوشش می‌دهند و بزرگتر از تقاطع ها هستند. در نتیجه، پس از یک دوره تکرار برای پیدا کردن مقدار آستانه مناسب برای خوشه بندی(شکل ۴-۵)، فاصله ۱۰۰ متر به اندازه کافی برای نشان دادن نقاط حادثه‌خیز ترافیکی در مناطق شهری به دست آمد. انتخاب فاصله در روش Nnh امری مهم است، چرا که در اصل، دامنه تحلیل خوشه‌ای را تعریف می کند. برای بررسی صرف بزرگراه ها با محدوده‌ی سرعت مجاز بالا، ممکن است فاصله ۱ کیلومتر قابل قبول باشد، اما برای مکان های شهری؛ محدوده ۱۰۰ متر ، به طور کلی، برای دستیابی به نقاط حادثه‌خیز واقع در تقاطعات اصلی شهری مناسب است. در نظر گرفتن فاصله های کوچکتر از قبیل ۵۰ متر یا ۲۰ متر برای تشخیص نقاط حادثه‌خیز در نزدیکی تقاطعات، بر اساس معیار n_{min} قابل انجام است اما امکان نبودن خوشه، هنگام نبودن نقاط کافی برای جستجو وجود دارد.

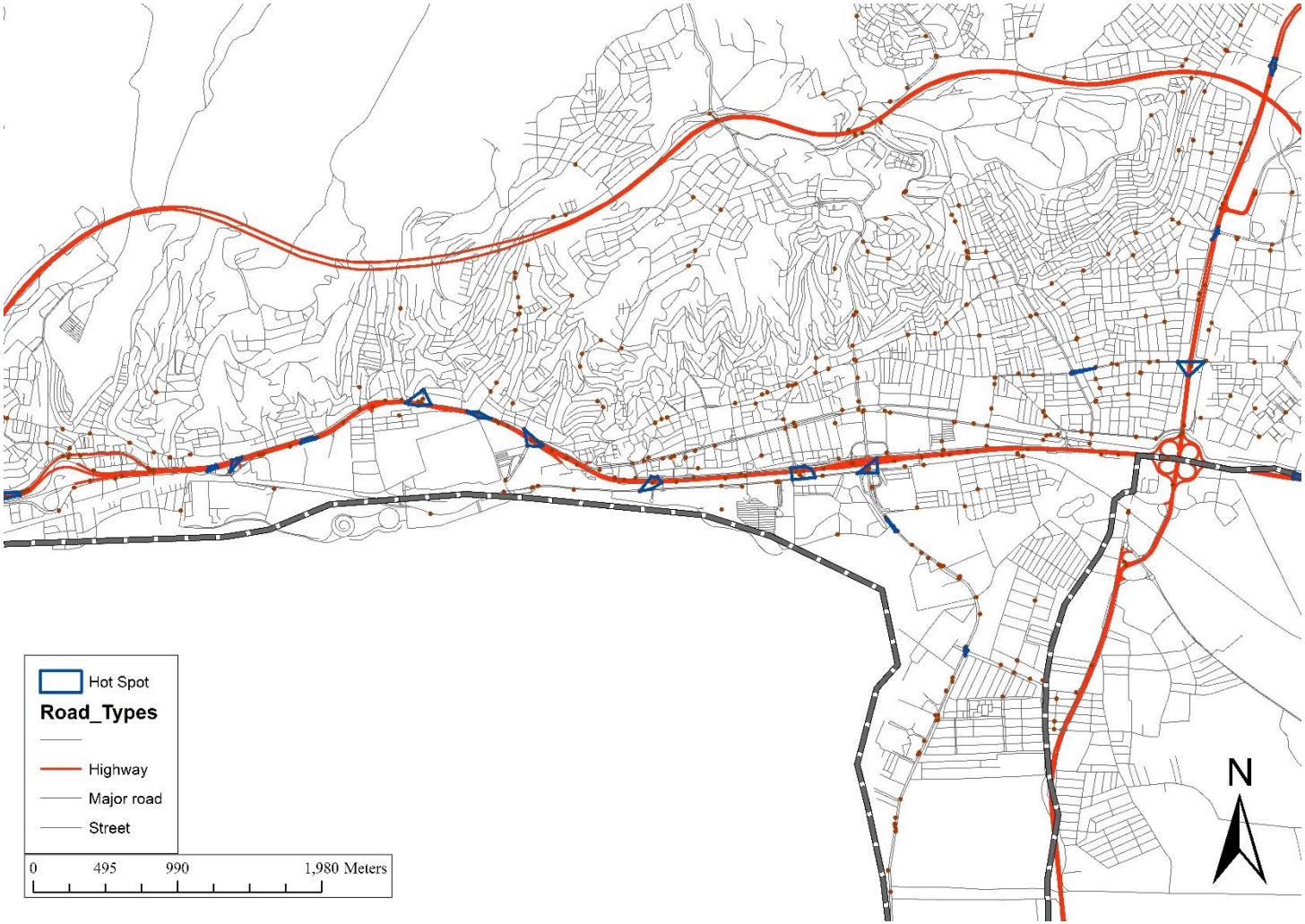


شکل۴- ۶:یکی از تقاطعات شهر ایزمیت که با معیار $d=100m$ و $n_{min} = 10$ به عنوان یک از خوشه های حادثه‌خیز بدست آمد.

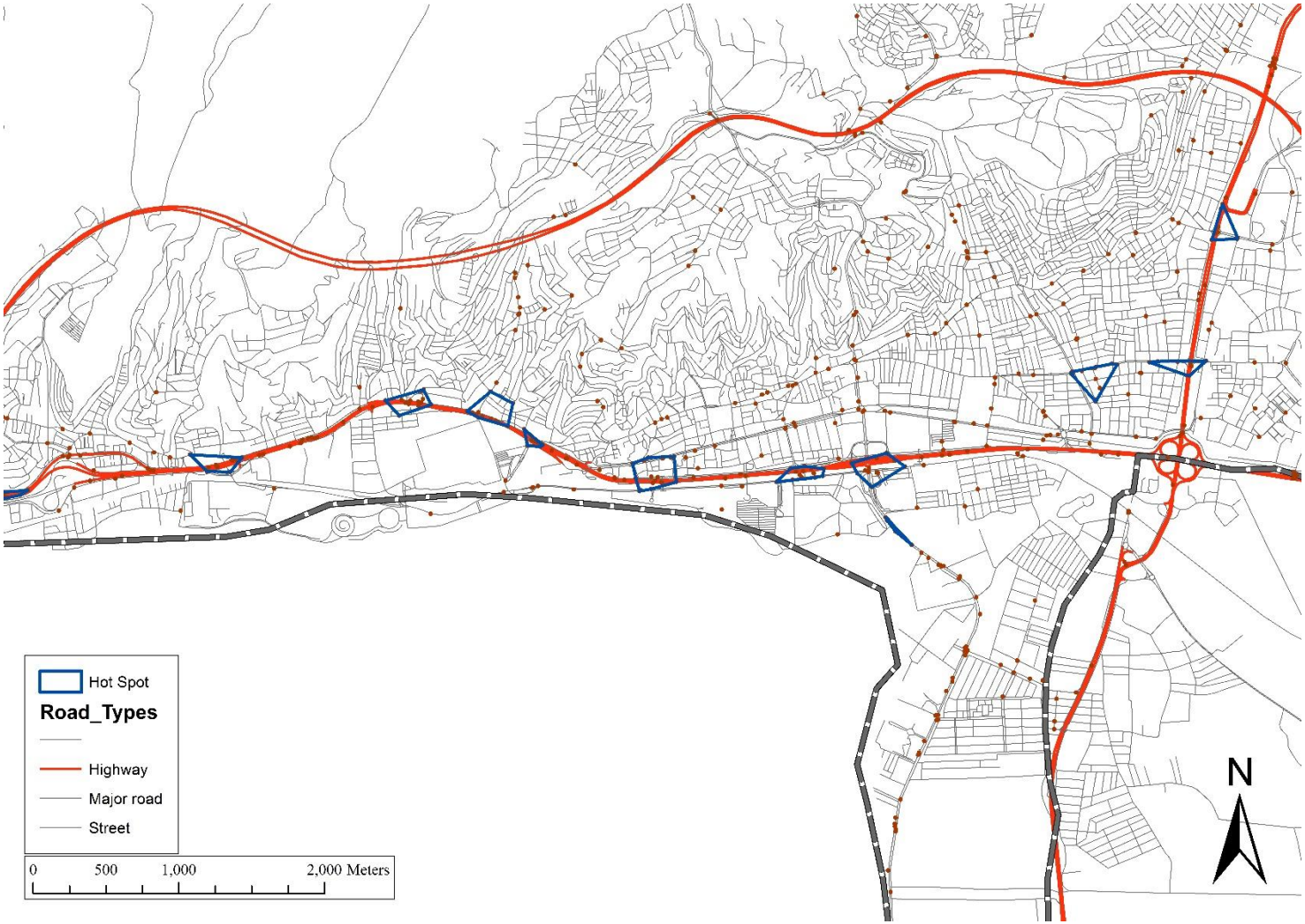
همانطور که در شکل ۴-۶ ملاحظه می‌کنید مقدار انتخابی d و n_{min} به خوبی نقاط تصادف در یک تقاطع را در برگرفته است. در ادامه برای مقایسه ی مقادیر مختلف n_{min} و d فاصله آستانه ی جستجو را برای درک تفاوت این دو پارامتر در خوشه بندی Nnh بررسی می‌کنیم.



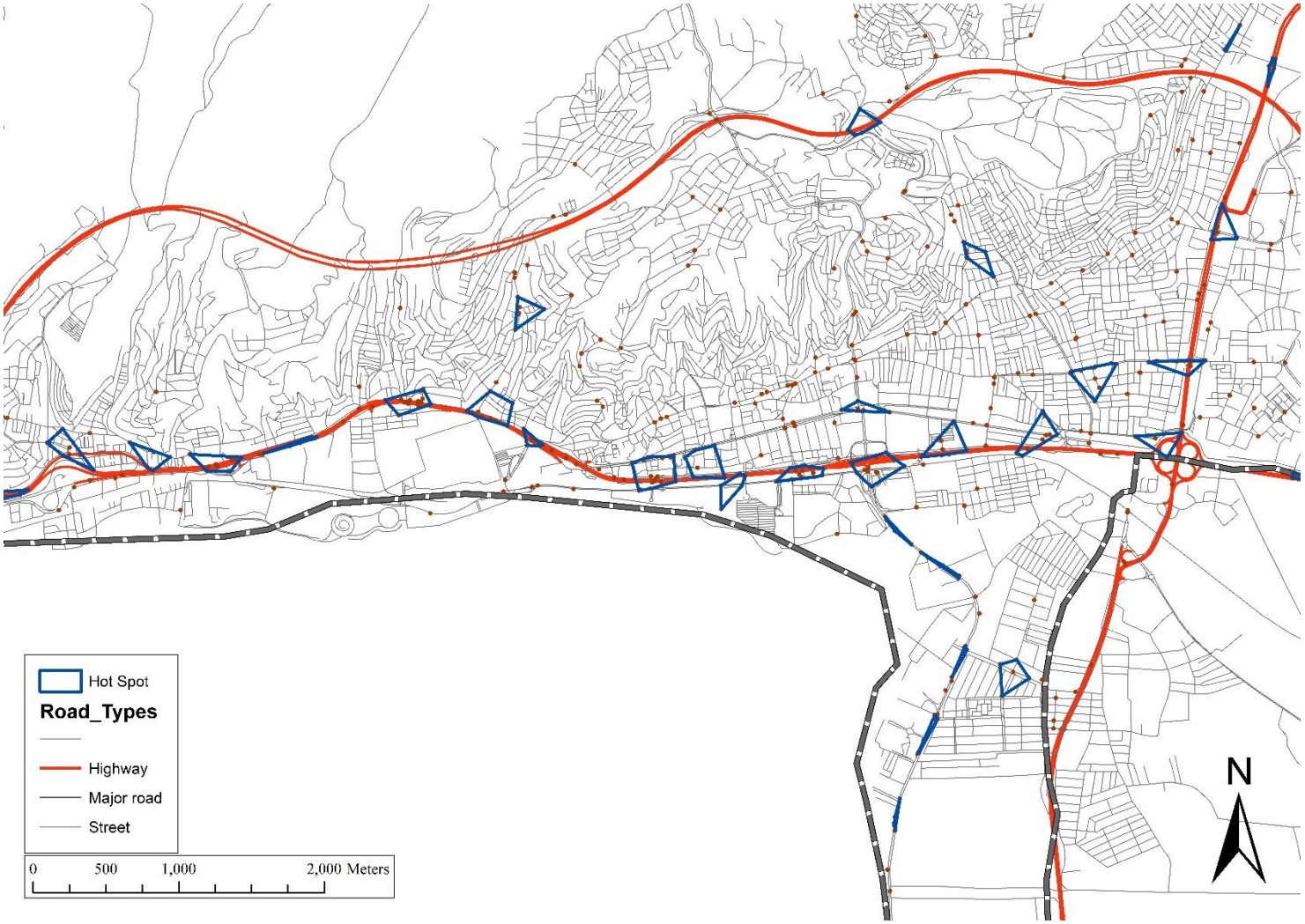
شکل۴- ۷: مقدار $n_{min} = 10$ و $d = 100$ برای خوشه بندی Nnh شهر ایزمیت



شکل۴- ۸: مقدار$n_{min} = 5$ و $d = 100$ برای خوشه بندی Nnh شهر ایزمیت



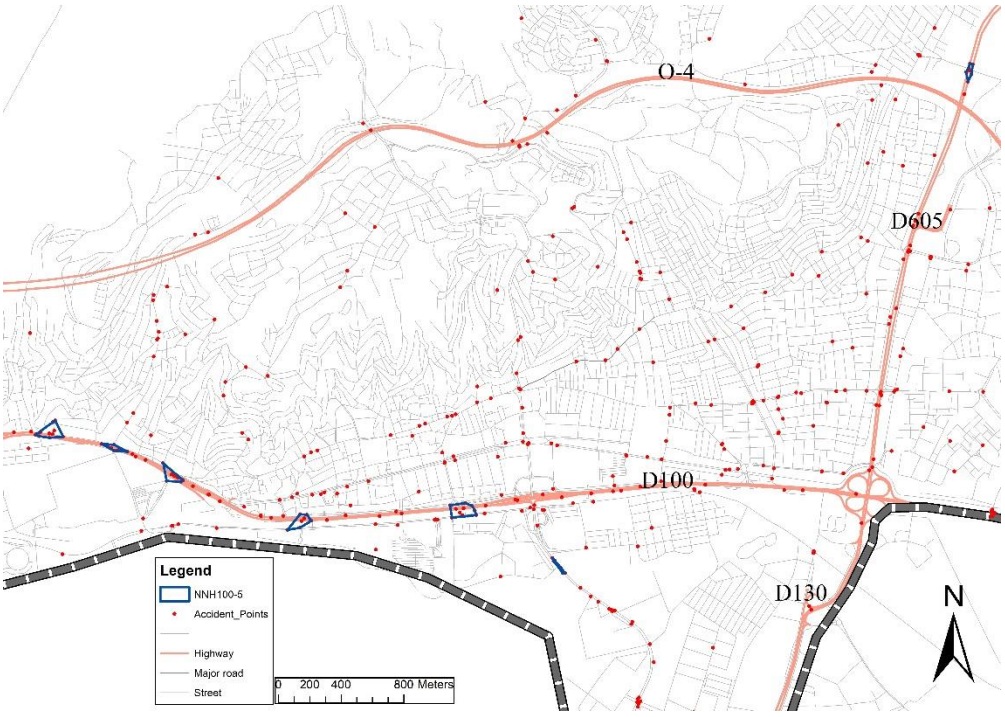
شکل۴- ۹: مقدار $n_{min} = 10$ و $d = 200$ برای خوشه بندی Nnh شهر ایزمیت



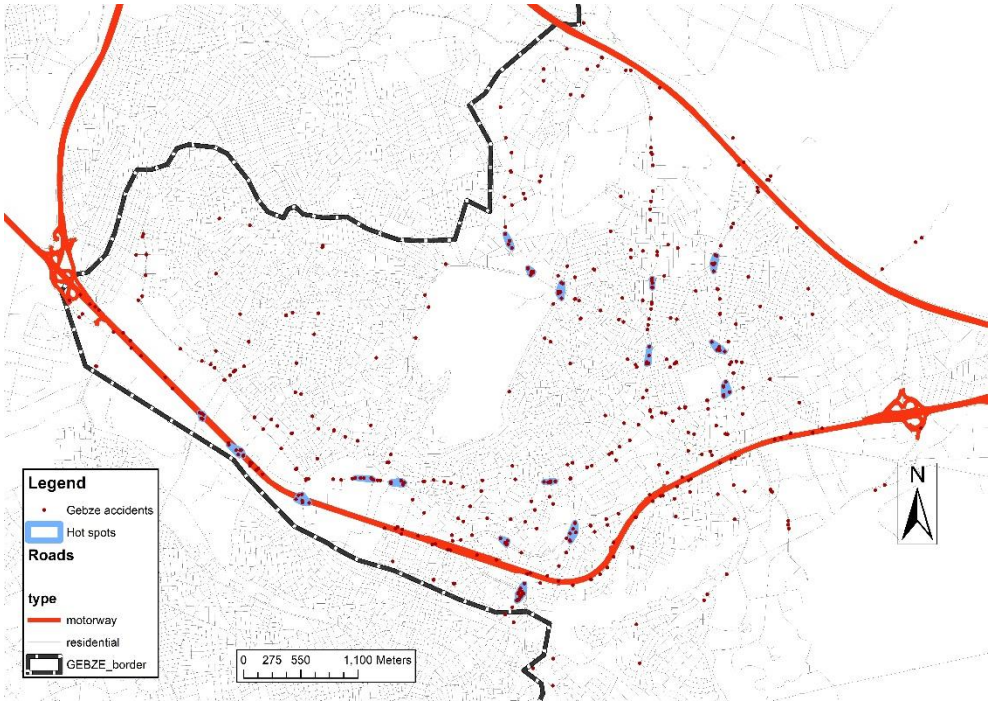
شکل۴- ۱۰: مقدار $n_{min} = 5$ و $d = 200$ برای خوشه بندی Nnh شهر ایزمیت

همانطور که در شکل‌های ۴-۷ تا ۴-۱۰ مشاهده می‌شود، با در نظر گرفتن فاصله آستانه ۱۰۰ متر و n_{min} های ۵ و۱۰برای مقایسه مقادیر مختلف n_{min} در هر دو شهر تعداد مختلفی از خوشه‌های حادثه خیز بدست آمد. همانطور که در شکل ۴-۷ و ۴-۸ قابل ملاحظه است به ازای کمترشدن مقدار n_{min}تعداد خوشه ها بیشتر می‌شود. بار دیگر مقدار فاصله آستانه را ۲۰۰ متر در نظر گرفتیم و مقادیر n_{min} های ۵ و۱۰قرار دادیم که در شکل۴-۹ و ۴-۱۰ قابل ملاحظه است. این بار نیز با کمتر شدن n_{min} تعداد خوشه ها بیشتر شد و بدین ترتیب هر چه مقدار n_{min} بیشتر باشد خوشه دارای تعداد تصادفات بیشتری خواهد بود و برای بررسی عوامل تصادف از لحاظ آماری قابل قبول‌تر است اما تعداد خوشه ها کمتر می‌شود و داده‌های ناخوشه زیاد می‌شود.

پس از مقایسه مقادیر متفاوت d و n_{min} بهترین مقدار برای اینکه بتوان با دیگر روش‌های خوشه‌بندی مقایسه نمود فاصله ۱۰۰ متر و تعداد نقاط ۵ در نظر گرفته شد. چرا که در این مقدار خوشه‌ها شامل نقاط حادثه‌خیز با تراکم بیشتر هستند.



شکل۴- ۱۱: مقدار $n_{min} = 5$ و $d = 100$ برای خوشه بندی Nnh شهر ایزمیت

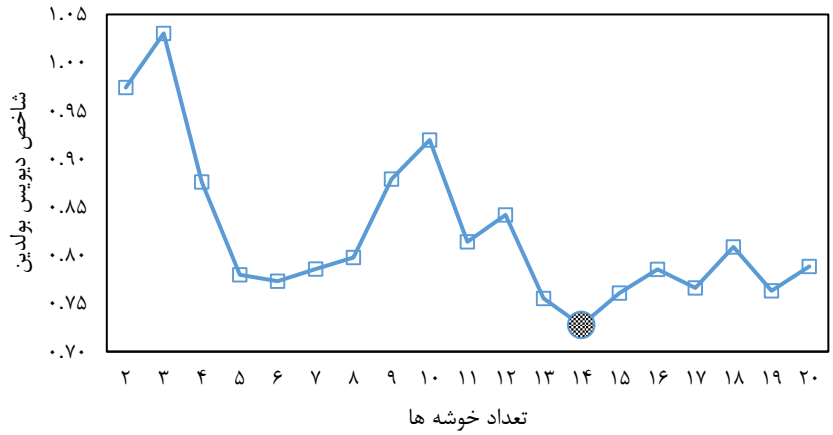


شکل۴- ۱۲:مقدار $n_{min} = 5$ و $d = 100$ برای خوشه بندی Nnh شهر ایزمیت

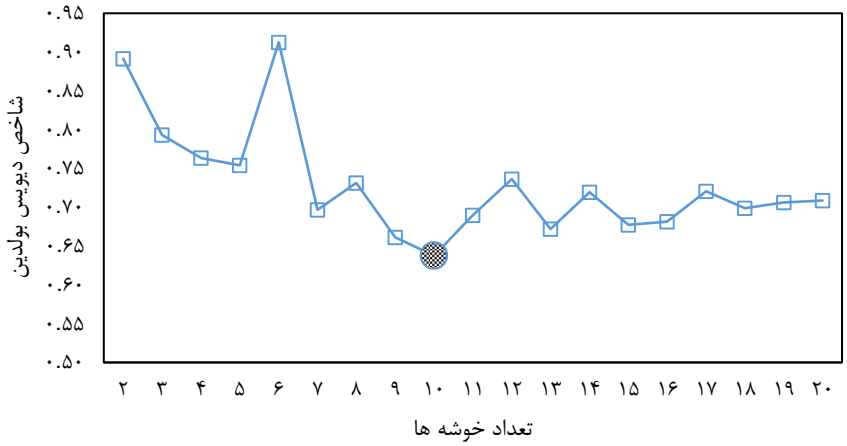
پس از بررسی خوشه‌ها مشاهده می‌شود که اکثر نواحی حادثه‌خیز در نزدیکی یا خود تقاطع‌ها، ورودی و خروجی های بزرگراه می‌باشند.تعداد کمی از این نواحی حادثه‌خیز در راه‌های فرعی و دور از مرکز شهر قرار دارند. در شهرهای با عرض معابر کم و توپوگرافی دارای اختلاف ارتفاع مانند گبزه نیز نقاط حادثه خیز در تقاطعات درون معابر دسترسی و راه‌های فرعی رخ می‌دهند (شکل ۴-۱۲). چرا که توپوگرافی خاص و بافت شهری گبزه مطمئنا رفتار ترافیکی متفاوتی نسبت به شهر ایزمیت که دارای خیابان‌های عریض است، خواهد داشت. در قسمت پایانی این فصل به تشریح خوشه‌ها مفصلا خواهیم پرداخت.

۴-۵-روش خوشه بندی K-Means

شاخص اعتبارسنجی دیویس-بولدین[1] برای بدست آوردن K بهینه در شکل ۴-۱۳ و ۴-۱۴ برای شهر گبزه و ایزمیت نمایش داده شده است. تـــعداد خوشه مناسب برای گبزه K=۱۴ و برای ایزمیت K=۱۰ به دست آمد.



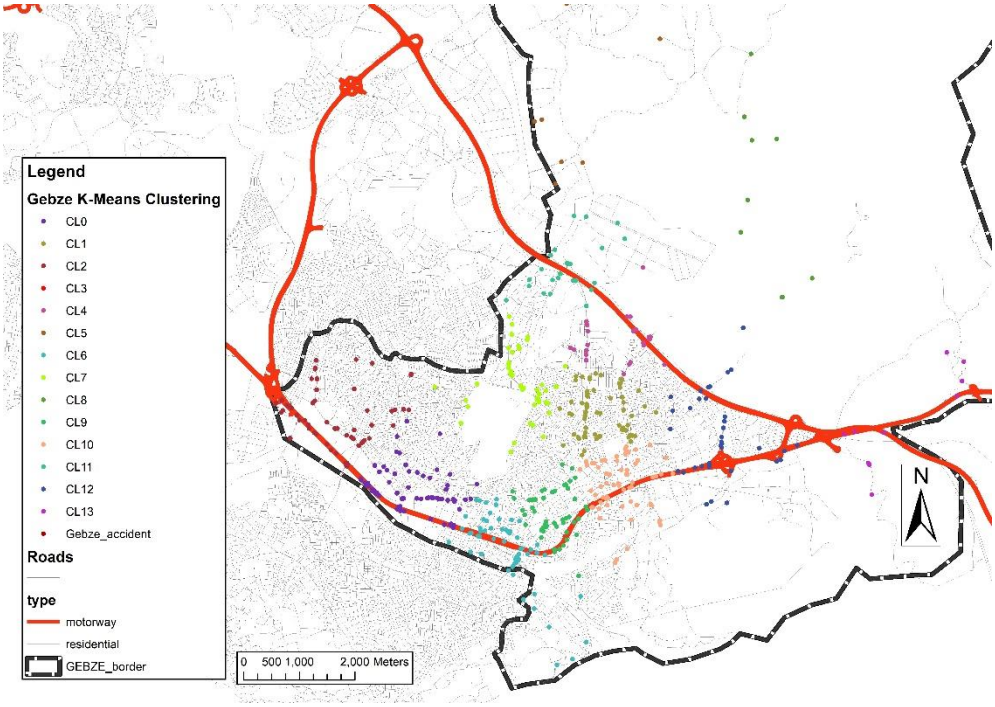
شکل۴- ۱۳:نمودار شاخص اعتبارسنجی دیویس-بولدین برای یافتن تعداد بهینه خوشه‌ها در روشK-Means برای شهر گبزه



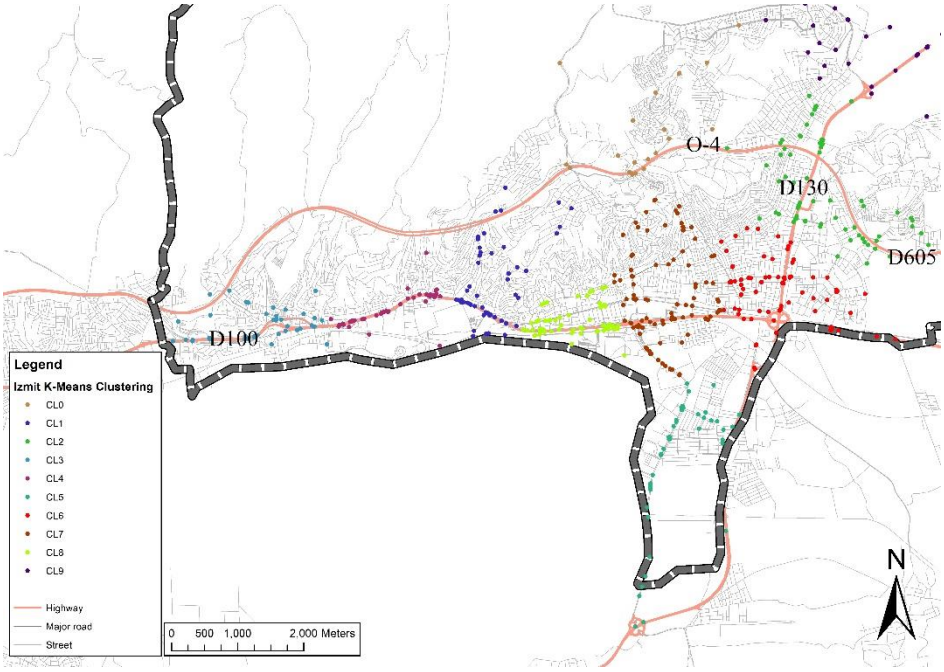
شکل۴- ۱۴: نمودار شاخص اعتبارسنجی دیویس-بولدین برای یافتن تعداد بهینه خوشه‌ها در روشK-Means برای شهر ایزمیت

[1] Davis-Bouldin Index

خروجی خو شه‌بندی به‌د ست‌آمده از خو شه‌بندی K-Means به ترتیب برای شهر گبزه و ایزمیت در شکل‌های ۴-۱۵ و ۴- ۱۶ قابل مشاهده است.



شکل۴- ۱۵: خوشه‌بندی K-Means شهر گبزه

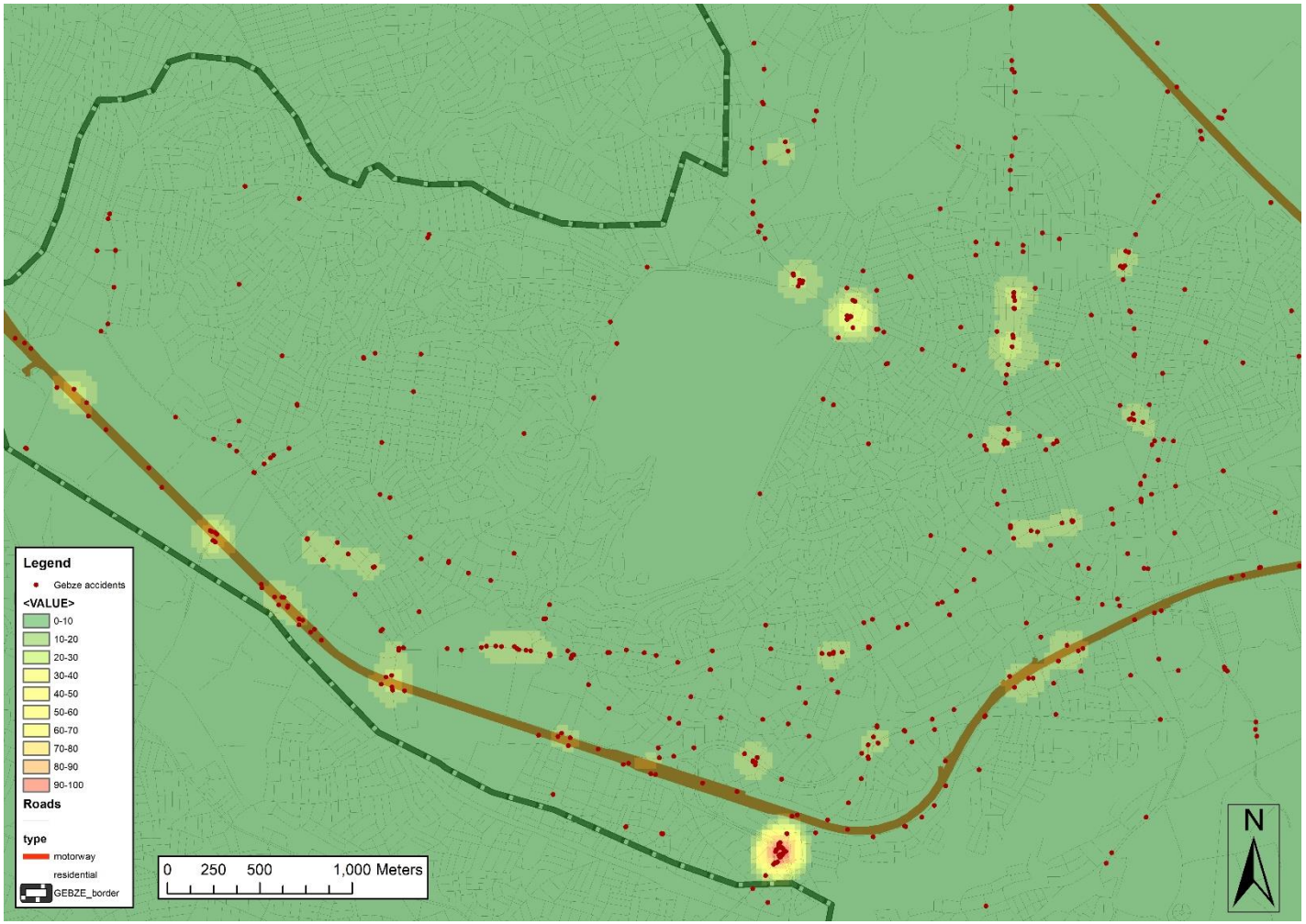


شکل۴- ۱۶:خوشه‌بندی K-Means شهر ایزمیت

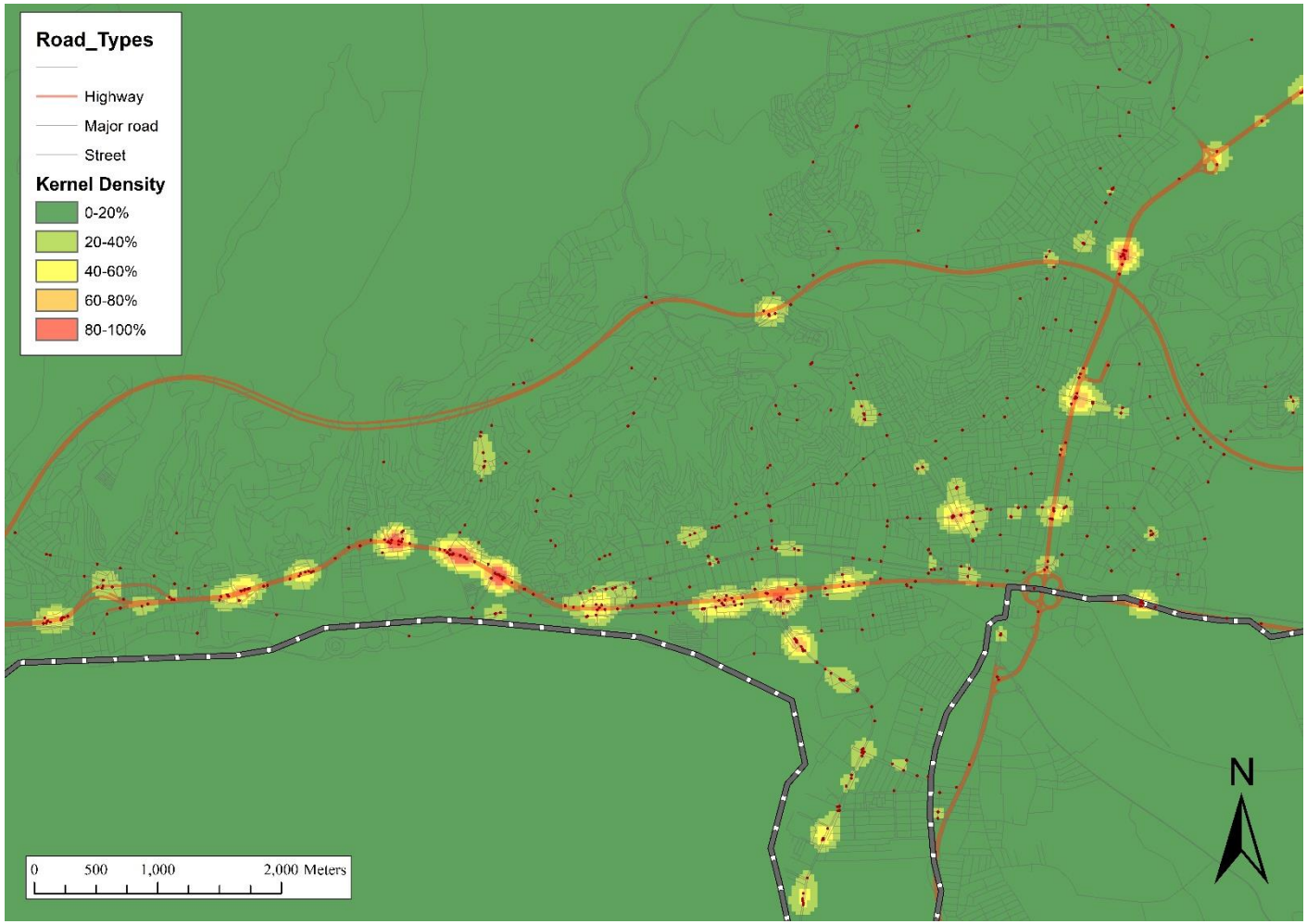
۴-۶-روش تخمین چگالی کرنل

همانطور که در روش تحقیق ذکر شده است، اندازه هر سلول و اندازه پهنای باند (شعاع جستجو) برای محاسبه نقشه KDE از محاسبه نقاط حادثه‌خیز شناسایی شده با روش خوشه بندی Nnh حاصل می شود.بدین ترتیب، کل مساحت convex hulها و حداکثر فاصله بین نقاط در نقاط حادثه‌خیز در نظر گرفته می شود. برای این مطالعه، شعاع جستجو نیز با توجه به نتایج خوشه بندی Nnh انتخاب شده است. برای انجام این کار، حداکثر فاصله تصادفات موجود در خوشه اندازه گیری می شود. حداکثر فاصله بین تصادفات ۲۰۰ متر است. به همین دلیل، برای هر نوع KDE، ۲۰۰ متر به عنوان شعاع جستجو استفاده می شود.

با توجه به این نقشه، مناطق قرمز رنگ که به مقدار تراکم نسبی ٪۸۰-۱۰۰٪ می‌باشد به عنوان نقاط حادثه‌خیز شناخته می شوند.در شهر اکثر تصادفات در محور D100 بزرگراه ایزمیت به استانبول رخ داده است. این تصادفات هم در خط اصلی بزرگراه و هم در رمپ‌ها مشاهده شده است.



شکل۴- ۱۷: نقشه تخمین ترام کرنل با شعاع جستجوی ۲۰۰ متر برای شهر گبزه



شکل۴- ۱۸: نقشه تخمین ترام کرنل با شعاع جستجوی ۲۰۰ متر برای شهر ایزمیت

تعداد نقاط قرار گرفته در مناطق حادثه خیز را محاسبه و مقدار نرخ ضربه و دقت پیش‌بینی را برای روش کرنل نیز محاسبه می‌کنیم. از سوی دیگر، در برخی از شریان های کوچک، مناطق متراکم دیده می شود که به دلیل ویژگی های کاربری زمین مانند مسکونی، تجاری و غیره در آن فعالیت های عابر پیاده ومناطق مسکونی قابل مشاهده است برای مثال یکی از نقاط حادثه‌خیز در منطقه مسکونی در شکل ۴-۱۹ قابل مشاهده است.

شکل۴- ۱۹: تقاطعی در منطقه مسکونی که در همه‌ی روش‌های خوشه‌بندی به عنوان نقطه حادثه‌خیز شناخته شده است.

۴-۷- نتایج روش‌های خوشه‌بندی و بررسی عوامل موثر بر تصادفات در نقاط حادثه‌خیز

با ادغام انواع خوشه بندی در محیط GIS نقاط حادثه‌خیز تصادفات را به راحتی می‌توان تشخیص داد. سپس برای درک رابطه بین پارامترهای مؤثر بر تصادفات و تعداد مجروحین در خوشه ها در روش

خوشه‌بندی GriDBSCAN، با استفاده از آزمون ANOVA در نرم‌افزار SPSS هر خوشه را مورد بررسی قرار دادیم و در هر خوشه تعدادی از پارامترها به عنوان پارامتر معنی‌دار شناخته شد. در این آزمون تعداد افراد جراحت دیده به عنوان داده وابسته و بقیه پارامترهای ذکر شده در فصل سوم به عنوان داده مستقل در نظر گرفته شد.

۴-۷-۱- توصیف خوشه‌های حادثه‌خیز شهر گبزه

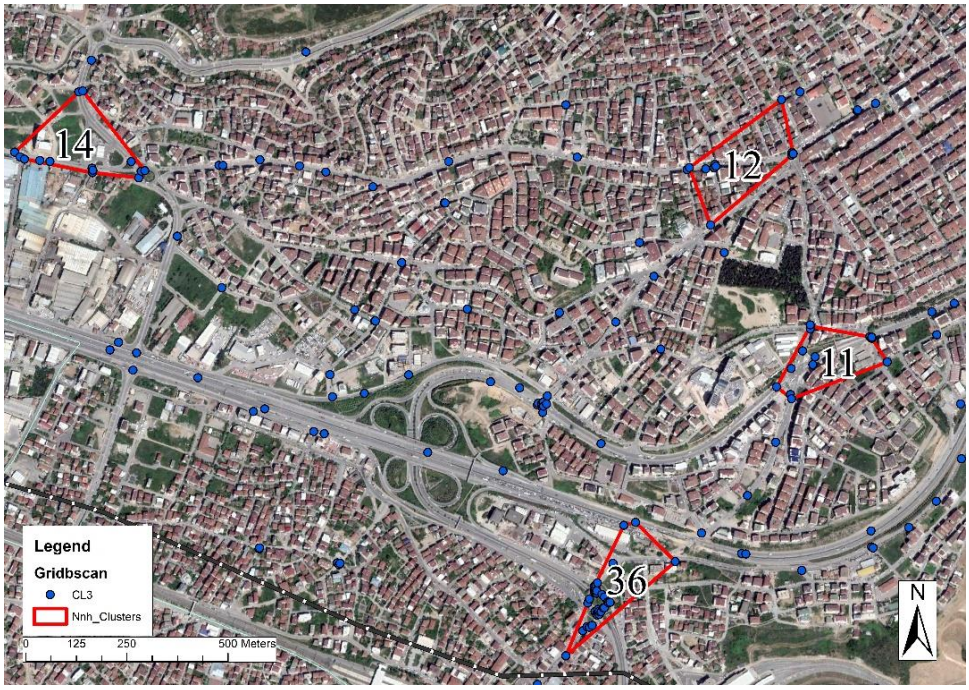
در این قسمت به بررسی و تشریح خوشه‌های حادثه‌خیز شهر گبزه می‌پردازیم و با ادغام خوشه‌ها ویژگی‌های مناطق را می‌کنیم. ابتدا پارامترهای معنادار در خوشه بندی GriDBSCAN را از طریق آزمون ANOVA دریافت کرده و پس از ادغام خوشه بندی Nnh و KDE پارامترها را مجددا بررسی می‌نماییم. همچنین با در اختیار بودن ابزار دیدخیابان در نرم‌فزار گوگل‌ارث به بررسی میدانی شرایط تقاطع می‌پردازیم.

خوشه CL3

این خوشه با بیشترین تعداد تصادف(۱۷۵) در جنوب غربی شهر قرار دارد. با تـــوجه به جدول ۴-۱ دو پارامتر، نوع تقاطع که تصادف در آن محل اتفاق افتاده و وجود یا عدم وجود گاردریل به عنوان داده تأثیر گذار بر تعداد افراد مجروح بدست آمده است. با توجه به به شکل ۴-۲۰ نقاط آبی تمام نقاط داخل خوشه CL3 می‌باشد و خطوط قرمز خوشه بندی Nnh و شماره‌ها تعداد تصادفات در داخل آن را نشان می‌دهد. با توجه به اینکه نوع تقاطع رابطه معناداری با تعداد مجروحین دارد و موقعیت تصادفات در نقشه‌ی ماهواره ای نشان می‌دهد که بیشترین تصادفات در تقاطعات رخ داده است با بررسی مجدد تصادفات موجود در خوشه‌های Nnh چند نقطه به عنوان پرحادثه ترین نقاط در داخل این خوشه ها بدست آمد.

جدول۴- ۱:نتایج آزمون ANOVA برای خوشه بندی GridBSACN خوشه‌های شهر گبزه

نام خوشه	نقاط درون خوشه	پارامترهای معنی دار با تعداد مجروحین	مجموع مربعات	درجه آزادی	مجذور میانگین	F	Sig.	اتا	مجذور اتا
CL0	۱۶	نوع تصادف	۲/۲۲۳	۴	۰/۵۵۶	۳/۵۶۶	۰/۰۴۲	۰/۷۵۱	۰/۵۶۵
		روز تصادف (روزهای هفته)	۳/۱۸۸	۵	۰/۶۳۸	۸/۵۰۰	۰/۰۰۲	۰/۹۰۰	۰/۸۱۰
CL1	۱۰	-	-	-	-	-	-	-	-
CL2	۳۶	وجود یا عدم وجود شانه راه	۷/۲۵۹	۱	۷/۲۵۹	۴/۳۸۴	۰/۰۴۴	۰/۳۳۸	۰/۱۱۴
		طبقه راه(خیابان یا بزرگراه شهری)	۸/۶۱۷	۱	۸/۶۱۷	۵/۳۳۳	۰/۰۲۷	۰/۳۶۸	۰/۱۳۶
		نوع راه (یکطرفه، دوطرفه جدا شده و جدانشده)	۱۰/۸۸۹	۲	۵/۴۴۴	۳/۴۱۱	۰/۰۴۵	۰/۴۱۴	۰/۱۷۱
CL3	۱۷۵	نوع تقاطع	۹/۰۲۸	۴	۲/۲۵۷	۲/۷۷۹	۰/۰۲۹	۰/۲۴۸	۰/۰۶۱
		وجود یا عدم وجود گاردریل	۷/۵۷۸	۱	۷/۵۷۸	۴/۰۴۵	۰/۰۴۷	۰/۱۸۰	۰/۰۳۲
CL4	۳۱	-	-	-	-	-	-		
CL5	۱۲۳	نوع تصادف	۲۵/۵۶	۵	۵/۱۱۲	۲/۸۶۵	۰/۰۱۸	۰/۳۳۰	۰/۱۰۹



شکل۴- ۲۰: ترکیب خوشه CL3 از روش GriDBSCAN و خوشه‌های حاصل از Nnh

جدول۴- ۲: توزیع نقاط در تقاطعات مختلف در خوشه CL3 شهر گبزه

تقاطع	تعداد	واریانس
چهارراه	۴۲	۰/۲۰۹
مسیر مستقیم	۵۰	۰/۶۸۹
میدان	۳۸	۲/۱۲۹
تقاطع T شکل	۳۸	۰/۴۱۳
تقاطع Y شکل	۷	۰/۲۸۶
مجموع	۱۷۵	۰/۸۴۶

همانطور که در جدول۴-۲ مشاهده می‌شود تعداد تصادفات در تقاطعات قابل توجه می‌باشد. با توجه به تصاویر دریافتی از طریق نرم‌افزار گوگل‌ارث[1] می‌توان به صحت تجمع نقاط در این تقاطعات پی

[1] Google Earth

برد. همچنین با ابزار دید خیابان[1] در گوگل ارث می‌توان از نزدیک این مکان ها را مشاهده نمود. تعدادی از این تقاطعات را در تصاویر زیر می‌توان مشاهده کرد.

شکل۴- ۲۱: تصاویر هوایی و تصاویر دید خیابان در نرم‌افزار گوگل ارث

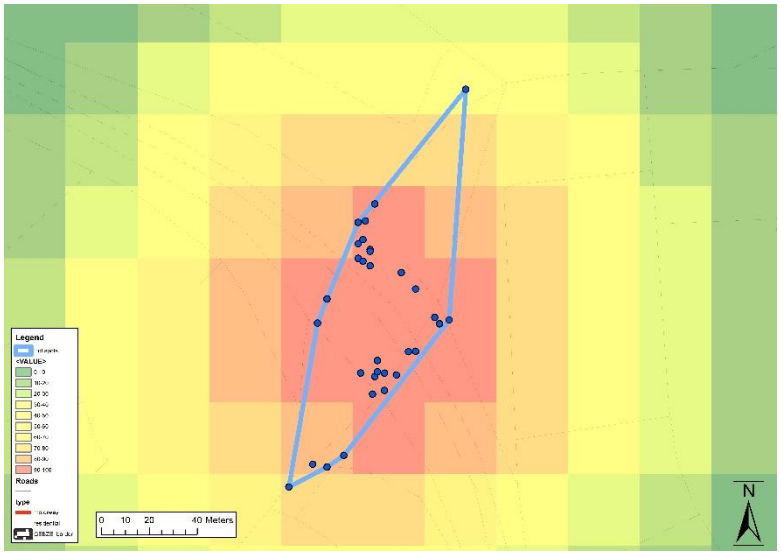
پارامتر گاردریل نیز به تبع به علت وقوع تصادفات در نقاط تقاطع شهری عدم وجود گاردریل به عنوان داده معنی‌دار به عنوان نتیجه ای بدیهی محسوب می‌شود.

با ادغام خوشه‌بندی‌های مختلف در GIS که در شکل ۴- ۲۲ قابل مشاهده است می‌توان نقاط حادثه‌خیز را دقیق‌تر مورد بررسی قرار داد. پس از بررسی مجدد داده‌های درون خوشه‌بندی حاصل از Nnh برای مثال برای میدانی که به عنوان نقطه حادثه‌خیز شناخته شده است پارامتر علائم چراغ‌دار

[1] Street view

راهنمایی به عنوان داده معنی دار بدست آمد . با توجه به شکل تقاطع و به علت تداخل حرکت راننده‌ها در این میدان تصادفات زیادی دراین منطقه اتفاق افتاده است. این مکان به عنوان یکی از پرحادثه‌ترین نقاط شهر گبزه می‌باشد.

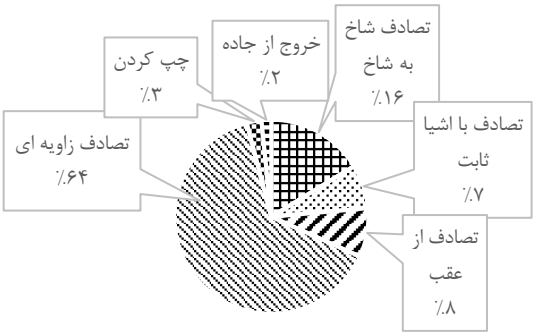
شکل۴- ۲۲: میدان پرتصادف در شهر گبزه



شکل۴- ۲۳: ادغام خوشه بندی Nnh، KDE و GriDBSCAN

خوشه CL5

با ۱۲۵ تصادف دارای بیشترین تصادف بعد از خوشه CL3 می‌باشد. در این خوشه پارامتر نوع تصادف با تعداد مجروحین رابطه معناداری دارد. همانطور که در شکل۴-۲۴ قابل مشاهده است در این خوشه تصادفات بیشتر از نوع تصادف زاویه‌ای یا از بغل می‌باشد.



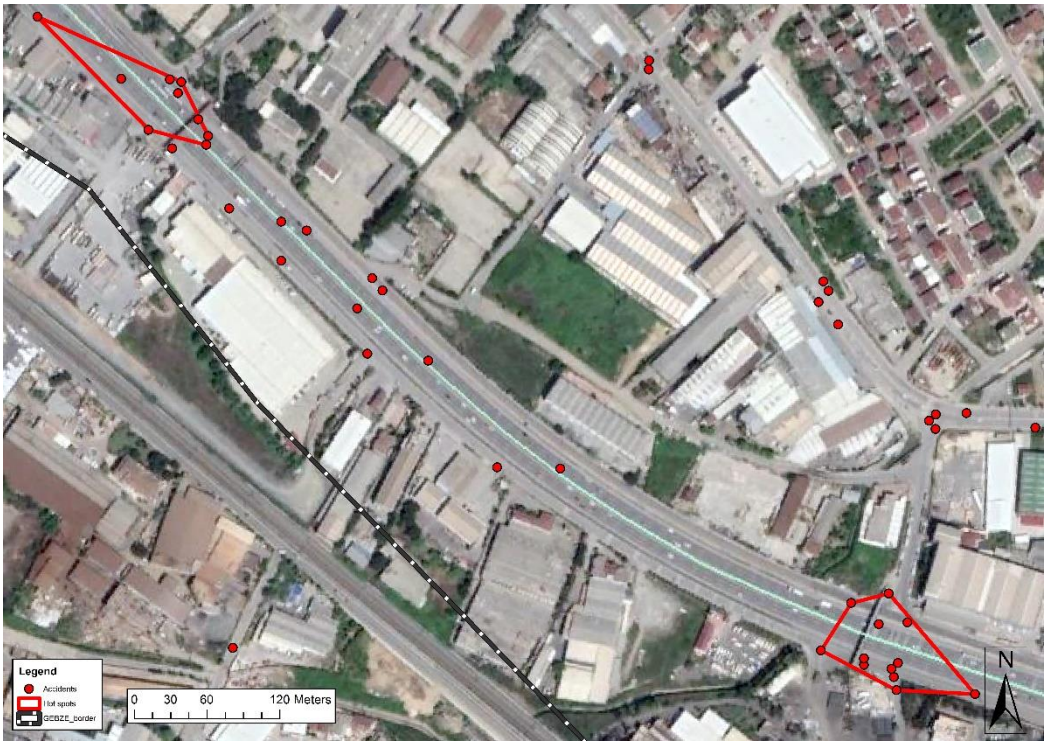
شکل۴- ۲۴: توزیع تصادفات در خوشه CL5 در روش GriDBSCAN

پس از بررسی مجدد این خوشه و ادغام دیگر روش‌های خوشه‌بندی پارامتر زمانی شب و روز به عنوان داده معنی دار با مقدار معنی‌داری ۰/۰۱۶با آلفا ۰/۰۵ بدست آمد. با بررسی تصاویر دید خیابان می‌توان ویژگی‌های محیطی این مناطق را مشاهده کرد. این منطقه از مناطق پرتراکم ساختمانی گبزه است و عرض خیابان ۹ متر و مسیر دو طرفه بدون میانه راه است، و دارای پارک حاشیه‌ای است.

خوشه CL2

تصادفات در این خوشه در محدوه‌ی بزرگراه ایزمیت-استانبول (D100) قرار دارد. برای مثال برای یکی از تقاطعات در این خوشه با توجه به خوشه بندی Nnh که درشکل ۴-۲۴ قابل مشاهده است با بررسی ویژگی‌های تصادف در این منطقه از لحاظ مکانی پارامتر نوع راه و طبقه راه به عنوان داده معناداری به‌دست آمد و باتوجه به اینکه موقعیت تصادفات این خوشه در محدوده بزرگراه می‌باشد معناداری این پارامترها بدیهی است. اما از لحاظ زمانی بیشتر تصادفات در این منطقه در روز اول کاری

یعنی روز دوشنبه اتفاق افتاده است و این موضوع می‌تواند دلیل مهم‌تری برای حادثه‌خیز بودن این محدوده باشد.



شکل۴- ۲۵: تصادفات خوشه CL2 و خوشه‌های روش Nnh در شهر گبزه

۴-۷-۲- توصیف خوشه‌های حادثه‌خیز شهر ایزمیت

در این بخش به شرح خوشه‌های شهر ایزیمت می‌پردازیم که این شهر نیز به دو بخش شهری و محور بزرگراه استانبول-ایزمیت تقسیم می‌شود. همانطور که در شکل ۴-۲۶ قابل مشاهده است تصادفات بیشتر در قسمت بزرگراه D100 ایزمیت به استانبول قرار دارند و تعداد دیگری از تصادفات نیز در تقاطعات و مناطق پرتراکم مسکونی قرار دارند.

شکل۴- ۲۶: تصویر ماهواره‌ای محور D۱۰۰ عبوری از شهر ایزمیت و توزیع نقاط تصادف

با توجه به جدول ۴-۳ خوشه‌هایی که دارای بیشترین تعداد تصادفات هستند را مورد بررسی قرار می‌دهیم.

جدول۴- ۳:نتایج آزمون Anova برای خوشه‌های شهر ایزمیت

نام خوشه	نقاط درون خوشه	پارامترهای معنی دار با تعداد مجروحین	مجموع مربعات	درجه آزادی	مجذور میانگین	F	Sig.	اتا	مجذور اتا
CL0	۱۰	-	-	-	-	-	-	-	-
CL1	۴۸	-	-	-	-	-	-	-	-
CL2	۱۰	نوع راه (یکطرفه، دوطرفه جدا شده و جدانشده)	۰/۷۱۱	۱	۰/۷۱۱	۶/۴۰۰	۰/۰۳۵	۰/۶۶۷	۰/۴۴۴
		طبقه راه(خیابان یا بزرگراه شهری)	۰/۷۱۱	۱	۰/۷۱۱	۶/۴۰۰	۰/۰۳۵	۰/۶۶۷	۰/۴۴۴
		مسیر مستقیم یا قوس	۰/۷۱۱	۱	۰/۷۱۱	۶/۴۰۰	۰/۰۳۵	۰/۶۶۷	۰/۴۴۴
CL3	۱۰	-	-	-	-	-	-	-	-
CL4	۲۹	-	-	-	-	-	-	-	-
CL5	۲۳	خط کشی‌های راه	۱۱/۴۵۸	۱	۱۱/۴۵۸	۱۶/۲۶۳	۰/۰۰۱	۰/۶۵۲	۰/۴۲۵
CL6	۱۰	شرایط نور مسیر	۰/۴	۱	۰/۴	۶/۴۰۰	۰/۰۳۵	۰/۶۶۷	۰/۴۴۴
		مسیر مستقیم یا قوس	۰/۴	۱	۰/۴	۶/۴۰۰	۰/۰۳۵	۰/۶۶۷	۰/۴۴۴
CL7	۱۶	وجود یا عدم وجود شانه راه	۶/۳۴۹	۱	۶/۳۴۹	۵/۶۸	۰/۰۳۲	۰/۵۳۷	۰/۲۸۹
CL8	۲۴	-	-	-	-	-	-	-	-
CL9	۱۳	تعداد خودرو درگیر تصادف	۱۰/۶۷۴	۲	۵/۳۳۷	۱۳/۰۳۲	۰/۰۰۲	۰/۸۵۰	۰/۷۲۳
CL10	۱۱	مسیر صاف یا شیبدار	۲/۶۲۷	۱	۲/۶۲۷	۱۱/۲۶۰	۰/۰۰۸	۰/۷۴۵	۰/۵۵۶
CL11	۵۴	تعداد خودرو درگیر تصادف	۱۵/۸۶۲	۳	۵/۲۸۷	۴/۰۷۴	۰/۰۱۲	۰/۳۳	۰/۱۰۹
CL12	۱۴	طبقه راه(خیابان یا بزرگراه شهری)	۳/۸۵۷	۱	۳/۸۵۷	۱۰/۲۸۶	۰/۰۰۸	۰/۶۷۹	۰/۴۶۲
		وجود یا عدم وجود گاردریل	۳/۸۵۷	۱	۳/۸۵۷	۱۰/۲۸۶	۰/۰۰۸	۰/۶۷۹	۰/۴۶۲
		شرایط نور مسیر	۳/۸۵۷	۱	۳/۸۵۷	۱۰/۲۸۶	۰/۰۰۸	۰/۶۷۹	۰/۴۶۲
CL13	۱۰	طبقه راه(خیابان یا بزرگراه شهری)	۸/۸۱۷	۱	۸/۸۱۷	۱۱/۵۹۵	۰/۰۰۹	۰/۷۶۹	۰/۵۹۲
		وجود یا عدم وجود شانه راه	۹/۳۴۴	۱	۹/۳۴۴	۱۳/۴۵۶	۰/۰۰۶	۰/۷۹۲	۰/۶۲۷

خوشه CL11

با بیشترین تعداد تصادف (54) در جنوب شهر و محور D100 ایزمیت به استانبول قرار دارد و در اطراف یکی از تقاطعات غیر همسطح اصلی در ورودی شهر می‌باشد(شکل۴-۲). در این خوشه تعداد خودروهای درگیر تصادف به عنوان پارامتر مؤثر بر تعداد مجروحین مشخص شده است. همانطور که در جدول زیر قابل مشاهد است تصادفات دو خودرویی آمار قابل توجهی را نشان می‌دهد.

جدول۴- ۴: آمار توصیفی تعداد خودروهای درگیر تصادف در خوشه CL11

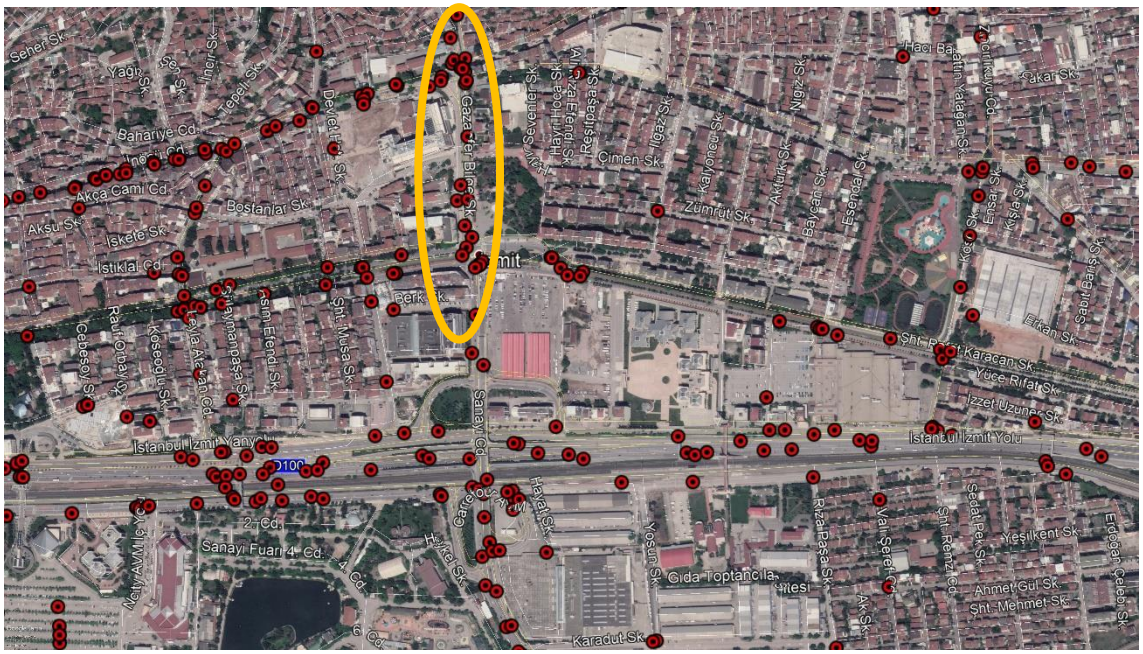
تعداد خودرو درگیر تصادف	N	واریانس
۱	۱۲	۱
۲	۳۲	۱/۱۱۲
۳	۷	۰/۹۰۵
۴	۳	۷
مجموع	۵۴	۱/۵۲۴

یکی از نقاط شناسایی شده درون این خوشه در خیابان تورکچاکال در مقابل آتش‌نشانی ایزمیت می‌باشد. در این نقطه بیشتر تصادفات بین دو خودرو بوده است.

شکل۴- 27: نقطه حادثه خیز در نزدیکی آتش‌نشانی ایزمیت

خوشه CL6

این خوشه در یکی محدوده پرترافیک شهری ایزمیت در خیابان غضنفربیلگه قرار دارد. این خیابان در امتداد خیابان سنایی که از خیابان‌های پرتردد و مسیر کناردریا مرمره می‌باشدکه به شهر باشیسکله منتهی می‌شود.



شکل۴-۲۸: محدوده خوشه CL6 شهر ایزمیت

با توجه به جدول ۳-۴ در این خوشه داده های معنی‌دار بدست آمده، یعنی قرارگیری در مسیر مستقیم یا قوس و شرایط نور مسیر می‌باشد. این پارامترها را نمی‌توان به راحتی به عنوان عوامل موثر بر تصادفات در این خوشه به حساب آورد چرا که باید به صورت میدانی نیز شرایط نور مسیر را در این مکان بررسی کرد. اما با توجه به داده‌های موجود تصادفات در این منطقه بیشتر در تایم شب اتفاق افتاده است. در این خوشه تقاطعی وجود دارد که پنج خیابان را به هم متصل می‌کند(شکل ۴-۲۹).

شکل۴- ۲۹: تقاطع غضنفربیلگه شهر ایزمیت خوشه CL6

خوشه CL2

این خوشه در محور اصلی استانبول-ایزمیت در قسمت غربی شهر قرار دارد. در این خوشه پارامتر مسیر قوس به عنوان داده موثر بر تصادفات شناخته شده است. با توجه به طرح هندسی راه این خوشه در مسیر قوس خطرناکی قرار گرفته است که آمار تصادفات در این ناحیه قابل توجه است(شکل۴-۳۰).

شکل۴- ۳۰: خوشه CL2 شهر ایزمیت

۴-۸-مقایسه روش‌های خوشه‌بندی و شاخص‌های ارزیابی خوشه‌ها

همانطور که در فصل دو اشاره گردید برای بررسی کیفیت خوشه بندی باید تعدادی از معیارهای درونی و بیرونی خوشه‌بندی را مورد بررسی قرار داد. در جدول ۴-۵ کیفیت خوشه‌بندی GriDBSCAN که در نرم افزار Elki انجام شد، قابل مشاهده است.

جدول۴- ۵: جدول شاخص‌های خوشه‌بندی GriDBSCAN

		گبزه	ایزمیت
معیارهای شمارش زوجی	جاکارد	۰/۱۱۵	۰/۲۷۹
	اندازه F	۰/۲۰۶	۰/۴۳۷
	دقت	۱/۰۰۰	۱/۰۰۰
	حساسیت	۰/۱۱۵	۰/۲۷۹
	رند	۰/۱۱۵	۰/۲۷۹
	رند اصلاحی	۰/۰۰۰	۰/۰۰۰
	فولکز-مالوز	۰/۳۳۹	۰/۵۲۹
معیارهای مبتنی بر آنتروپی	NMI	۰/۲۴۰	۰/۳۴۰
	مجذور NMI	۰/۴۹۰	۰/۵۸۰
معیار مبتنی بر BCubed	اندازه F	۰/۴۲۸	۰/۴۳۷
	حساسیت	۱/۰۰۰	۱/۰۰۰
	دقت	۰/۲۷۲	۰/۲۷۹
معیار مبتنی بر تطابق مجموعه داده	اندازه F	۰/۵۶۷	۰/۶۷۰
	شاخص خلوص	۰/۳۹۶	۰/۵۰۳
	شاخص خلوص معکوس	۱/۰۰۰	۱/۰۰۰
معیار های مربوط به فاصله خوشه‌ها	اندازه F	۰/۵۶۴	۰/۶۶۲
	دقت	۰/۳۹۴	۰/۵۰۲
	حساسیت	۰/۹۸۹	۰/۹۷۴
	شاخص جینی	±۰/۳۶۴۱= میانگین	±۰/۳۶۰۵= میانگین

همانطور که در جدول ۴-۶ مشاهده می‌شود روش خوشه‌بندی GriDBSCAN در مدت زمان کمتری نسبت به الگوریتم معمولی DBSCAN اجرا می‌شود و برای داده‌هایی همچون داده‌های تصادفات که دارای حجم زیادی هستند، مدت زمان اجرای الگوریتم را به طور قابل توجهی کاهش می‌دهد. تفاوت عمده خوشه‌بندی DBSCAN معمولی با GriDBSCAN در اجرای سریعتر الگوریتم است.

جدول۴- ۶: مقادیر مورد استفاده ه برای الگوریتم GriDBSCAN

پارامترهای الگوریتم GriDBSCAN	Gebze	Izmit
پارامتر Eps(m)	۳۵۰	۲۵۰
minPts	۱۵	۱۰
d شبکه(m)	۷۰۰	۵۰۰
تعداد خوشه‌ها	۶	۱۴
GriDBSCAN Runtime (msec)	۱۰	۱۲
DBSCAN Runtime (msec)	۲۰	۱۹

در جدول۴-۶ برای هر شهر مقادیر مورد استفاده برای الگورریتم و نتایج به دست آمده از الگوریتم(تعداد خوشه ها، تعداد داده های خوشه‌بندی شده و ...)نشان داده‌شده است. پس از بررسی تمام نتایج مقدار نرخ ضربه و شاخص دقت پیش‌بینی را برای مشخص کردن بهترین روش خوشه‌بندی استفاده شده در تحقیق حاضر در جدول ۴-۷ آمده است.

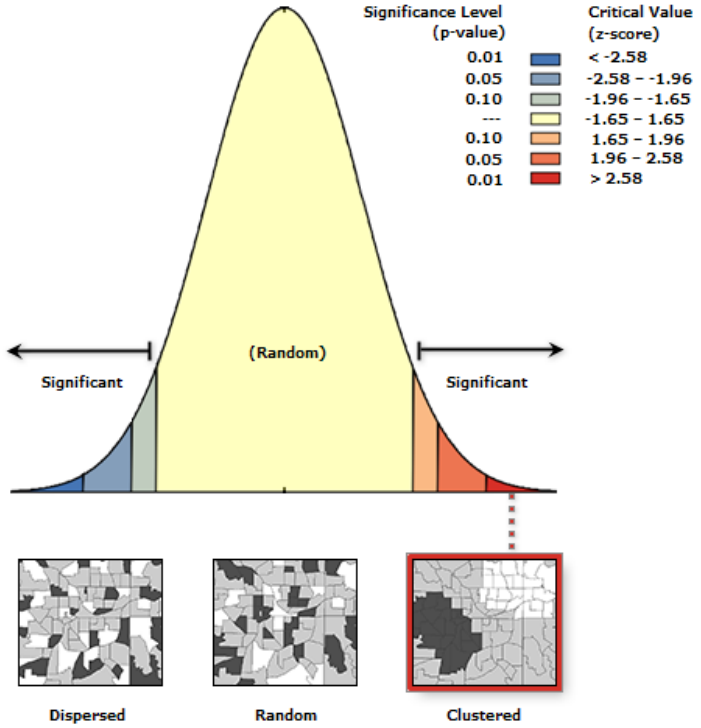
جدول۴- ۷: مقادیر نرخ ضربه و شاخص دقت پیش‌بینی روش‌های مختلف

	GriDBSCAN		Nnh		KDE	
نتایج حاصل از خوشه بندی	گبزه	ایزمیت	گبزه	ایزمیت	گبزه	ایزمیت
مساحت منطقه مورد مطالعه	۶۷۸۰۲۹۳۸	۳۱۴۲۵۲۲۰	۶۷۸۰۲۹۳۸	۳۱۴۲۵۲۲۰	۶۷۸۰۲۹۳۸	۳۱۴۲۵۲۲۰
مساحت خوشه‌بندی شده(متر مربع)	۶۴۶۵۸۸۳	۱۲۸۰۴۲۲	۲۹۱۸۲۸	۳۱۹۹۲۹	۲۵۶۱۵۴۶	۱۷۲۷۹۸۰
تعداد نقاط خوشه‌بندی شده	۳۹۱	۲۸۴	۲۵۴	۱۷۶	۲۸۰	۲۲۵
تعداد نقاط کل	۶۴۷	۵۷۲	۶۴۷	۵۷۲	۶۴۷	۵۷۲
نرخ ضربه	۶۰/۴۳	۴۹/۶۵	۳۹/۲۶	۳۰/۷۷	۴۳/۲۸	۳۰/۷۷
شاخص PAI	۶/۳۷۷	۱/۱۷۸	۹۱/۲۱	۳۰/۲۲	۱۱/۴۵	۲/۹۲

با توجه به نتایج جدول۴-۷ با توجه به مقادیر PAI مدل Nnh روش دقیق‌تری برای شناسایی نقاط حادثه‌خیز بدست آمده است. با استفاده از خوشه‌بندی روش GridBSCAN می‌توان مناطق متراکم تر را پیدا کرده و ویژگی‌های کلی منطقه‌ی تصادف را بدست‌آورد و سپس با روش Nnh برای شناسایی نقاط حادثه‌خیز اقدام نمود. الگوریتم GriDBSCAN برای داده‌های انبوه بسیار مناسب است و برای شناسایی مناطق تراکم داده‌ها کاربردی است و داده های پرت را از خوشه‌ها جدا می‌سازد.

۴-۹-نتایج حاصل از توابع موجود در GIS

در این قسمت نتایج توابع خودهمبستگی موران و گتیس‌ارد که برای تعیین الگوی پراکنش داده‌ها است آمده است. با توجه به توضیحات ذکر شده در فصل دو این تابع نشان می‌دهد آیا داده‌های مورد نظر دارای الگوی خوشه‌ای هستند یا نه. در شـــکل ۴-۳۱ مقادیر شـــاخص موران، امتیاز استانداردZ(Z-Score) وP-Value برای شهر ایزمیت قابل مشاهده است.

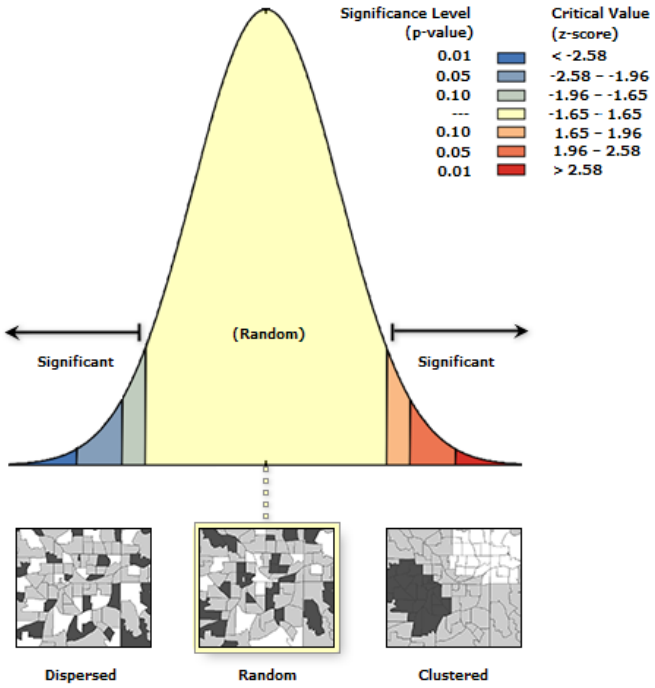


شکل۴- ۳۱: نتیجه تابع خود همبستگی موران برای شهر ایزمیت

جدول۴- ۸: نتایج تابع خودهمبستگی موران برای شهر ایزمیت

Izmit		**Gebze**	
Moran's Index:	۰/۰۵۷۱۱۷	**Moran's Index:**	-۰/۰۰۳۳۶۲
z-score:	۳/۱۳۳۸۴۰	**z-score:**	-۰/۱۲۵۵۸۸
p-value:	۰/۰۰۱۷۲۵	**p-value:**	۰/۹۰۰۰۵۸

با توجه به مقادیر به‌دست‌آمده P-Value که مقدار کمتری دارد و ولی شاخص موران نزدیک به ۱ و امتیاز استاندارد Z مثبت داده‌های مورداستفاده در شهر ایزمیت دارای الگوی خوشه‌ای هستند. تابع موران الگوی خوشه‌ای بودن داده‌های مورد نظر را تایید می‌کند. اما در در شهر گبزه که شاخص موران منفی بدست‌آمده و مقدارP-Value بالاتری دارد، دارای الگوی تصادفی هستند و خوشه‌بندی داده‌ها این شهر الگوی تصادفی دارد.



شکل۴- ۳۲: نتیجه تابع خود همبستگی موران برای شهر ایزمیت

فصل ۵

بحث و نتیجه‌گیری

۵-۱- مقدمه

تاکنون تحقیقات زیادی به‌منظور ایجاد ارتباط بین مبحث خوشه‌بندی و تحلیل تصادفات انجام‌گرفته است. همان‌طور که در پیشینه تحقیق بیان شد؛ محققین از روش‌های مختلف داده‌کاوی همچون طبقه‌بندی و خوشه‌بندی برای تحلیل تصادفات استفاده کرده‌اند. در بیشتر مطالعات تصادفات، تصادفات برون‌شهری مدنظر بوده و تصادفات درون‌شهری به علت ساختار پیچیده و به علت شبکه‌ای بودن راه در مناطق شهری کمتر موردمطالعه قرارگرفته است. اما به علت افزایش تصادفات درون‌شهری در کشورهای درحال‌توسعه تصمیم بر این شد تا نقاط حادثه‌خیز تصادفات را شناسایی کرده و عوامل مؤثر بر آن‌ها را موردبررسی قرار دهیم. تحلیل مکانی تصادفات و اهمیت سیستم اطلاعات جغرافیایی در تحلیل‌های مکانی سبب شد که در این تحقیق به تحلیل مکانی تصادفات با الگوریتم‌های خوشه‌بندی مرتبط با بحث‌های مکانی پرداخته شود.

۵-۲- خلاصه نتایج و جمع‌بندی

با توجه به نتایج به‌دست‌آمده از روش‌های خوشه‌بندی، روش‌های خوشه‌بندی غیر آماری GriDBSCAN، Nnh و روش خوشه‌بندی آماری برآورد تراکم کرنل انتخاب شدند. با توجه به اینکه روش‌های مبتنی بر تراکم برای تحلیل نقاط حادثه‌خیز روشی قابل‌اطمینان‌تر و مطلوب‌تر هستند و نتایج نسبتاً پایداری را به کاربر ارائه می‌دهند. داده‌های تصادفات با استفاده از این روش‌ها تحلیل شدند. درنهایت مناطق حادثه‌خیز شهری به‌خوبی شناسایی شدند و این مناطق شناسایی‌شده با توجه به‌واقع شدن در مناطق مسکونی یا اداری و یا تقاطعات شهری ازلحاظ انطباق با دنیای واقعی نیز در مکان‌های مستعد تصادف قرارگرفته‌اند.

باتوجه به مقدار شاخص دقت پیش‌بینی PAI برای روش‌های خوشه‌بندی انجام‌شده روش Nnh به عنوان بهترین روش برای شناسایی نقاط حادثه‌خیز بدست‌آمد. تفسیر خوشه‌ها برای هر کدام از شهرها

نشان داد که هر شهر ویژگی‌های مختص به خود را دارد و نقاط حادثه‌خیز نتیجه‌ی تجمیع عوامل مختلف ترافیکی، هندسی، و دیگر عوامل در یک منطقه است.

در تحقیق حاضر درشهر ایزمیت اغلب نقاط حادثه‌خیز تصادف در محور اصلی D100 ایزمیت-استانبول اتفاق افتاده است و بیشترین تصادفات در این محور در نزدیکی خروجی‌ها و ورودی‌ها بود و همینطور در محور وسط این بزرگراه در نزدیکی ایستگاه‌های اتوبوس و پل عابرپیاده تصادفات رخ‌داده است. در شبکه شهری ایزمیت نیز تصادفات بیشتر در نزدیکی تقاطعات بود. روش‌های مورد استفاده همگی توانستند به خوبی نقاط تمرکز تصادف را خوشه بندی کنند.

در شهر گبزه نیز با توجه به توپوگرافی متنوع و شیب‌دار و همچنین تراکم بالای کاربری‌های مختلف در عرض معابر کم تصادفات به صورت پراکنده در اکثر تقاطعات و خیابان‌های محلی و دسترسی اتفاق افتاده است .

با توجه به توزیع زمانی داده‌ها در هر دو شهر، اکثر تصادفات در زمان اوج ترافیک عصر و سپس در زمان اوج ترافیک صبح اتفاق افتاده است. سازمان‌های ترافیکی و ارگان‌های مرتبط، باید باتوجه به عوامل به‌دست آمده مؤثر بر تصادفات در این دو شهر در جهت رفع این عوامل تصمیم‌گیری کرده و اقدام نمایند.

در انتها با توجه به مقدار شاخص دقت پیش‌بینی بادقت ترین روش، روش Nnh بدست آمد. با توجه به عنوان تحقیق سعی بر آن بود تا روش‌های مختلف خوشه‌بندی تصادفات را مورد بررسی قرار دهیم و روشی کاربردی ارئه دهیم. در این تحقیق الگوریتمی جدید به عنوان GriDBSCAN مورد استفاده گردید. این روش نسبت به DBSCAN معمولی از لحاظ زمانی ۵۰٪ درصد سریعتر اجرا می‌شود.

۵-۳- پیشنهادات کارهای آینده

- باتوجه به نبود داده‌هایی همچون حجم ترافیک، عرض دقیق معابر توصیه می‌شود تأثیر این دو عامل بر تصادفات را نیز در تحقیقات آتی در نظر گرفته‌شود.
- از الگوریتم GriDBSCAN برای خوشه‌بندی تصادفات برون‌شهری و برای آزادراه‌ها و بزرگراه نیز استفاده گردد تا مشخص گردد این الگوریتم درکدام مناطق کاربردی‌تر است.
- با توجه به زمان اجرای بالای این الگوریتم بهتر است برای داده‌های بزرگ[1] همانند تصادفات یا جرایم در مناطق درون‌شهری که دارای ویژگی‌های مکانی هستند نیز استفاده شود.
- وضعیت جوی دقیق نیز در ماه های مختلف سال و ساعات مختلف روز تغییر می کند؛ لذا بررسی نقش عامل فوق با روش های حاضر در این تحقیق در تصادفات قابل انکار نیست و باید مورد توجه قرار گیرند.

-ارائه راهکار جهت کاهش تعداد و خسارات تصادفات عابرین پیاده با توجه به اثرات عوامل موثر و انجام مطالعات قبل-بعد با همکاری نهاد های اجرایی جهت ارزیابی میزان کارآیی این راهکار ها.

[1] Big Data

مراجع:

احدی, م., زایرزاده, ع., ۱۳۸۴. پارامترهای مؤثر بر خستگی رانندگان و نقش آن در وقوع تصادفات. اولین کنفرانس بین‌المللی حوادث رانندگی و جاده‌ای.

جمال شهرابی, ۱۳۹۲. داده‌کاوی. جهاد دانشگاهی (دانشگاه صنعتی امیرکبیر).

سجادی, ح., ۱۳۸۸. تحلیل تصادفات جاده‌ای در سیستم اطلاعات جغرافیایی با تأکید بر خصوصیات جاده و محیط (مطالعه موردی: محور قزوین -رشت). وزارت علوم، تحقیقات و فناوری - دانشگاه شهید بهشتی - دانشکده علوم زمین.

عفتی, م., رجبی, م.ع., حکیم پور, ف., شعبانی, ش., ۱۳۹۳. تحلیل عوامل مکانی مؤثر بر تمرکز تصادفات در راه‌های برون‌شهری با استفاده از GIS و داده‌کاوی. علوم و فنون نقشه‌برداری سال چهارم ۲ , ۸۷–۱۰۲.

محمد سلمانی, مهدی رمضان زاده لسبویی, مسلم دریکوند, فرخ ثابتی, ۱۳۸۷. بررسی عوامل مؤثر بر تصادفات جاده‌ای و ارائه راه‌کارهایی برای کاهش آن ؛ موردمطالعه : منظومه روستایی جنوب خور و بیابانک. پژوهش‌های جغرافیای انسانی سال چهلم ۶۵ , ۸۷–۱۰۴.

معتمد, م., ۱۳۸۸. مدل‌سازی تصادفات ترافیکی جهت ارتقاء ایمنی با استفاده از روش‌های آماری پیشرفته. وزارت علوم، تحقیقات و فناوری - دانشگاه صنعتی خواجه‌نصیرالدین طوسی - دانشکده عمران.

Alotaibi, A.S., 2018. Density-based clustering for road accident data analysis. Int. J. Adv. Appl. Sci. 5 8 , 113–121.

Anderson, T.K., 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. Accid. Anal. Prev. 41 3 , 359–364.

Andersson, A.K., Chapman, L., 2011. The impact of climate change on winter road maintenance and traffic accidents in West Midlands, UK. Accid. Anal. Prev. 43 1 , 284–

289.

Association, W.R., 2016. Road Safety Manual: Recommendations from the World Road Association (PIARC).

Bradley, P.S., Mangasarian, O.L., Street, W.N., 1997. Clustering via concave minimization, in: Advances in Neural Information Processing Systems. pp. 368–374.

Chainey, S., Tompson, L., Uhlig, S., 2008. The utility of hotspot mapping for predicting spatial patterns of crime. Secur. J. 21 1–2 , 4–28.

Chang, L.-Y., Chen, W.-C., 2005. Data mining of tree-based models to analyze freeway accident frequency. J. Safety Res. 36 4 , 365–375.

Chen, S., Saeed, T.U., Alinizzi, M., Lavrenz, S., Labi, S., 2019. Safety sensitivity to roadway characteristics: a comparison across highway classes. Accid. Anal. Prev. 123, 39–50.

Chimba, D., Sando, T., 2009. The prediction of highway traffic accident injury severity with neuromorphic techniques. Adv. Transp. Stud. 2009 19 , 17–26.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38 3 , 434–444.

Dobra, A., 2002. Classification and regression tree construction. Retrieved Sept. 18, 2011.

Dunham, M.H., 2003. Data Mining: Introductory and Advanced Topics, Pearson Education India. Pearson Education India.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in: Kdd. pp. 226–231.

Fowlkes, E.B., Mallows, C.L., 1983. A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. 78 383 , 553–569.

Gaonkar, M.N., Sawant, K., 2013. AutoEpsDBSCAN: DBSCAN with Eps automatic for

large dataset. Int. J. Adv. Comput. Theory Eng. 2 2 , 11–16.

Geurts, K., Wets, G., Brijs, T., Vanhoof, K., 2003. Profiling of high-frequency accident locations by use of association rules. Transp. Res. Rec. J. Transp. Res. Board 1840 , 123–130.

Grabusts, P., Borisov, A., 2002. Using grid-clustering methods in data classification, in: Parallel Computing in Electrical Engineering, 2002. PARELEC'02. Proceedings. International Conference On. IEEE, pp. 425–426.

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

Ivan, J.N., Wang, C., Bernardo, N.R., 2000. Explaining two-lane highway crash rates using land use and hourly exposure. Accid. Anal. Prev. 32 6 , 787–795.

Jaccard, P., 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat 37, 547–579.

Kaufman, L., Rousseeuw, P.J., 1990. Partitioning around medoids (program pam). Find. groups data an Introd. to Clust. Anal. 68–125.

Kovács, F., Legány, C., Babos, A., 2005. Cluster validity measurement techniques, in: 6th International Symposium of Hungarian Researchers on Computational Intelligence. Citeseer.

Kumar, S., Toshniwal, D., 2016. Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). J. Big Data 3 1 , 1–11. doi:10.1186/s40537-016-0046-3

Kunt, M.M., Aghayan, I., Noii, N., 2011. Prediction for traffic accident severity: comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods. Transport 26 4 , 353–366.

Levine, N., 2010. CrimeStat III: A Spatial Statistics Program for the Analysis of Crime Incident Locations (version 3.3). Ned Levine & Associates: Houston, National Institute of TX.

Loh, W., 2008. Classification and regression tree methods. Wiley StatsRef Stat. Ref.

Online.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Oakland, CA, USA, pp. 281–297.

Mahran, S., Mahar, K., 2008. Using grid for accelerating density-based clustering, in: Computer and Information Technology, 2008. CIT 2008. 8th IEEE International Conference On. IEEE, pp. 35–40.

Moghaddam, F.R., Afandizadeh, S., Ziyadi, M., 2011. Prediction of accident severity using artificial neural networks. Int. J. Civ. Eng. 9 1 , 41.

Moreira, A., Santos, M.Y., Carneiro, S., 2005. Density-based clustering algorithms–DBSCAN and SNN. Univ. Minho-Portugal.

Oh, C., Oh, J.-S., Ritchie, S., Chang, M., 2001. Real-time estimation of freeway accident likelihood, in: 80th Annual Meeting of the Transportation Research Board, Washington, DC.

Oh, J.-S., Oh, C., Ritchie, S.G., Chang, M., 2005. Real-time estimation of accident likelihood for safety enhancement. J. Transp. Eng. 131 5 , 358–363.

Polat, K., Durduran, S.S., 2011. Subtractive clustering attribute weighting (SCAW) to discriminate the traffic accidents on Konya–Afyonkarahisar highway in Turkey with the help of GIS: A case study. Adv. Eng. Softw. 42 7 , 491–500.

Prabha, S.L., Shanavas, A.R.M., 2014. Educational data mining applications. Oper. Res. Appl. An Int. J. 1 1 , 1–6.

Prasannakumar, V., Vijith, H., Charutha, R., Geetha, N., 2011. Spatio-temporal clustering of road accidents: GIS based analysis and assessment. Procedia-Social Behav. Sci. 21, 317–325.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. 66 336 , 846–850.

Sabel, C.E., Kingham, S., Nicholson, A., Bartie, P., 2005. Road traffic accident

simulation modelling-a kernel estimation approach, in: The 17th Annual Colloquium of the Spatial Information Research Centre University of Otago, Dunedin, New Zealand. Citeseer, pp. 67–75.

Sagberg, F., 2018. Characteristics of fatal road crashes involving unlicensed drivers or riders: implications for countermeasures. Accid. Anal. Prev. 117, 270–275.

Sasaki, Y., 2007. The truth of the F-measure. Teach Tutor mater 1 5 , 1–5.

Shafabakhsh, G.A., Famili, A., Bahadori, M.S., 2017. GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran. J. traffic Transp. Eng. (English Ed. 4 3 , 290–299.

Tan, P.-N., Steinbach, M., Kumar, V., 2005. Introduction to data mining. 1st.

Tian, R., Yang, Z., Zhang, M., 2010. Method of road traffic accidents causes analysis based on data mining, in: Computational Intelligence and Software Engineering (CiSE), 2010 International Conference On. IEEE, pp. 1–4.

Tseng, W.-S., Nguyen, H., Liebowitz, J., Agresti, W., 2005. Distractions and motor vehicle accidents: Data mining application on fatality analysis reporting system (FARS) data files. Ind. Manag. Data Syst. 105 9 , 1188–1205.

Turkstat, T.S.I., 2018. Turkish Statistical Institute Road Traffic Accident Statistics 2014.

World-Health-Organization, 2015. Global status report on road safety 2015. World Health Organization.

Yi-ling, Z., Hongbo, X., Shuo, B.A.I., 2008. OPTICS-Plus for Text Clustering. J. Chinese Inf. Process. 22 1 , 51–60.

Abstrcat

Based on traffic accident statistics, it is clear that traffic accidents impose the harmful effects and unnecessary costs on society; thus, researchers try to analysis traffic accidents, identify accident locations, and determine significant parameters.Today, data mining techniques, specifically spatial clustering methodes, are used to analyze accident data and find their spatial patterns. In this research, a combined method called GriDBSCAN algorithm was used for clustering accident data. The DBSCAN algorithm applied for its spatial characteristics and the GRID algorithm used by taking data into a grid mesh increased the accuracy and execution time in big data such as accident data. Other clustering methods such as K-Means, Nnh, KDE and spatial autocorrelation were also used to be compareed with GriDBSCAN method. The results showed that Nnh algorithm was the most accurate method for spatial clustering of traffic accident points, and GriDBSCAN algorithm was also applicable to the separation of high density areas of accidents in high volume data. The clusters obtained from GriDBSCAN algorithm properly differentiated the accidents of different urban areas according to the density. By examining the parameters in these clusters, it is possible to reduce the number of accidents and understand the factors affecting them.

Keywords: Urban accidents, data mining algorithms, spatial clutering analysis, GriDBSCAN algorithm

Shahrood University of Technology

Faculty of Civil Engineering

M.Sc. Thesis in Transportation Engineering

Presenting a spatial clustering method for crash data

By: Amin Ganjali Khosrowshahi

Supervisor:

Dr. Iman Aghayan

Advisor:

Dr. Abdolahad Choupani

Jan 2019