تحقیقات منابع آب ایران
Iran-Water Resources Research

سال پانزدهم، شماره ۳، پاییز ۱۳۹۸
Volume 15, No.3, Fall 2019 (IR-WRR)
۶۲-۷۷

پایش داده‌های جریان در شبکه‌های توزیع آب با استفاده از روش‌های خوشه‌بندی مبتنی بر چگالی

ایمان مصلحی[۱]، محمدرضا جلیلی قاضی‌زاده[۲*] و احسان یوسفی خوش‌قلب[۳]

چکیده

تشخیص داده‌های نویز (پرت یا غیرعادی) از داده‌های جریان در شبکه‌های توزیع آب در مرحله آماده‌سازی و پیش‌پردازش داده‌ها برای دستیابی به داده‌های تاریخی قابل اعتماد انجام می‌گیرد که در بهبود روش‌های ارزیابی و مدیریت نشت و بهره‌برداری مؤثر از شبکه، مهم و ضروری است. هدف از ارائه این مقاله توسعه یک متدولوژی جدید بر مبنای روش‌های یادگیری بدون نظارت، جهت شناسایی داده‌های نویز در یک مجموعه داده‌های جریان در شبکه‌های توزیع آب می‌باشد. متدولوژی توسعه داده شده شامل مراحل: ۱- جمع‌آوری داده‌های مورد نیاز، ۲- صحت‌سنجی و نرمال‌سازی داده‌ها و ۳- شناسایی و کشف داده‌های پرت یا غیرعادی با استفاده از الگوریتم خوشه‌بندی مکانی مبتنی بر چگالیِ مقاوم در مقابلِ نویز (DBSCAN) می‌باشد. متدولوژی پیشنهادی برای داده‌های جریان ورودی به یک منطقه در شبکه توزیع آب شهری تهران با تواتر زمانی برداشت داده ۱۵ دقیقه برای سال ۱۳۹۴ استفاده شد. نتایج نشان داد که متدولوژی توسعه‌یافته قابلیت شناسایی داده‌های پرت ناشی از انواع شکستگی‌ها و مصارف مجاز غیرمعمول نظیر مصارف ناشی از تغییر در الگوی مصرفی جمعیت یا مصارف مجاز غیرعادی را دارد. بنابراین این متدولوژی را می‌توان به عنوان ابزاری کاربردی و انعطاف‌پذیر برای پایش داده‌های جریان، شناسایی و حذف انواع داده‌های پرت از آن‌ها مورد استفاده قرار داد.

کلمات کلیدی: داده پرت، خوشه‌بندی مبتنی بر چگالی، الگوریتم DBSCAN، داده‌های جریان، شبکه توزیع آب.

تاریخ دریافت مقاله: ۹۷/۱۱/۵
تاریخ پذیرش مقاله: ۹۸/۳/۳

۱- دانشجوی دکتری مهندسی عمران- آب، دانشکده مهندسی عمران، آب و محیط زیست، دانشگاه شهید بهشتی تهران.
۲- استادیار دانشکده مهندسی عمران، آب و محیط زیست، دانشگاه شهید بهشتی تهران.
۳- دانشجوی کارشناسی ارشد مهندسی عمران- آب و سازه‌های هیدرولیکی، دانشکده مهندسی عمران، آب و محیط زیست، دانشگاه شهید بهشتی تهران.
*- نویسنده مسئول
بحث و مناظره (Discussion) در مورد این مقاله تا پایان زمستان ۱۳۹۸ امکانپذیر است.

Monitoring of Flow Data in Water Distribution Networks Using Density-Based Clustering Methods

I. Moslehi[1], M.R. Jalili Ghazizadeh[2], and E. Yousefi Khoshqalb[3]*

Abstract

Detection of noise (anomaly or outlier) from flow data in water distribution networks (WDNs) is implemented in data preparation and prepossessing to achieve reliable historical data. This improves the assessment and management of leakage and helps the efficient operation of the network. The main objective of this paper is to develop a new methodology based on unsupervised learning methods for noise detection in a flow data set in water distribution networks. The developed methodology includes three steps 1- required data acquisition, 2- data validation and normalization, and 3- anomaly or outlier detection using the density-based spatial clustering of application with noise (DBSCAN) algorithm. The proposed methodology is applied for inflow data into a zone in Tehran's urban water distribution network which has 15-min reading intervals for the year 1394 (April 2015-March 2016)). The results showed that the developed methodology is capable in detection of anomalies caused by different types of pipe breaks and unusual legitimate consumption such as water usage due to changes in water consumption pattern or unauthorized consumption. Therefore, this methodology can be used as an applied and flexible tool for flow data monitoring and detecting and eliminating different types of outliers.

Keywords: Outlier, Density-Based Clustering, DBSCAN Algorithm, Flow Data, Water Distribution Network.

Received: February 4, 2019
Accepted: May 24, 2019

1- Ph.D. Student in Civil Engineering, Faculty of Civil, Water and Environmental Engineering, Shahid Beheshti University, Tehran, Iran.
2- Assistant Professor, Department of Water and Wastewater, Faculty of Civil, Water and Environmental Engineering, Shahid Beheshti University, Tehran, Iran. Email: m_jalili@sbu.ac.ir
3- M.Sc. Student in Civil Engineering, Faculty of Civil, Water and Environmental Engineering, Shahid Beheshti University, Tehran, Iran.
*- Corresponding Author

۱- مقدمه

یکی از معضلات اصلی شبکه‌های توزیع آب، سطح بالای نشت از این تأسیسات است. همچنین نشت از شبکه علاوه بر هدررفت حجم قابل توجهی از آب تصفیه شده، می‌تواند منجر به هدررفت انرژی، کاهش درآمد، ورود آلودگی و قطع جریان آب نیز شود. از این رو یکی از وظایف اصلی شرکت‌های آب و فاضلاب در حوزه بهره‌برداری، مدیریت نشت در شبکه‌های توزیع آب می‌باشد. منظور از مدیریت نشت، یک فرآیند پیوسته و مداوم به منظور کنترل و کاهش تلفات آب از طریق اجرای یک سری اقدامات فنی است. روش‌های مربوط به مدیریت نشت را می‌توان به سه دسته کلی شامل ۱) روش‌های ارزیابی و پایش نشت، ۲) روش‌های تشخیص و کشف نشت و ۳) روش‌ها و مدل‌های کنترل نشت طبقه‌بندی کرد (Puust et al., 2010).

روش‌های اصلی ارزیابی و پایش مقدار نشت شامل: ۱- روش بالا به پایین (جدول بالانسینگ)، ۲- روش آنالیز مؤلفه و ۳- روش پایین به بالا یا روش آنالیز جریان شبانه می‌باشند (AL-Washali et al., 2019). از بین روش‌های ذکر شده، روش پایین به بالا تنها روشی است که از طریق اندازه‌گیری‌های میدانی (اندازه‌گیری جریان و فشار) در نواحی ایزوله شده (DMA) انجام می‌گیرد (Mutikanga et al., 2013). هدف از پایش نشت در این روش، اندازه‌گیری پیوسته و منظم جریان ورودی به یک زون یا ناحیه ایزوله به منظور تحلیل جریان شبانه و تعیین میزان نشت شبکه در ساعات حداقل جریان شبانه است که معمولاً بین ساعات ۲ تا ۴ صبح اتفاق می‌افتد (Farley and Trow, 2005). به علت عدم قطعیت‌های موجود در روش تحلیل جریان شبانه به خصوص در تخمین میزان مصرف شبانه مشترکین و نیز عدم امکان کاربرد این روش برای نواحی بزرگ یا کل شبکه، روش‌هایی دیگری نیز بر مبنای تحلیل آماری داده‌های جریان برای ارزیابی میزان نشت توسعه یافته‌اند. در این روش‌ها از داده‌های تاریخی جریان که اغلب توسط سامانه‌های اسکادا ثبت و جمع‌آوری شده، استفاده می‌شود و از طریق روش‌های آماری نظیر روش ادغام داده، روش‌های رگرسیونی، آنالیز آماری ترتیبی و غیره میزان نشت از شبکه تخمین زده می‌شود (Alkasseh et al., 2013; Buchberger and Nadimpalli, 2004; Mazzolani et al., 2017).

در روش تحلیل جریان شبانه با کسر مصارف مجاز شبانه از جریان حداقل شبانه، تخمینی از میزان نشتِ شبکه در مدت وقوع جریان شبانه بدست می‌آید؛ حال اگر طی این مدت در شبکه شکستگی و یا مصارف غیرعادی مثل مصارف مجاز غیرمعمول (حجم بالای آبیاری فضای سبز، تغییر ناگهانی در الگوی مصرف و غیره) و یا مصارف غیرمجاز وجود داشته باشد، دقت و قابلیت اطمینان نتایج این روش کاهش یافته و مقدار نشت شبانه با اختلاف از مقادیر واقعی تخمین زده می‌شود (Thornton et al., 2008). روش‌های آماری موجود نیز اغلب بر مبنای داده‌های جریان با توزیع آماری از قبل تعیین‌شده (برای مثال توزیع نرمال، لاگ‌نرمال) یا داده‌های جریان با الگوی مشابه توسعه داده شده‌اند و نیاز است که در آن‌ها داده‌های غیرعادی حذف گردند. با این وجود یکی از محدودیت‌های اصلی در استفاده از روش‌های آماری، شناسایی و حذف داده‌های جریان با الگوی غیرعادی و نامتعارف (داده پرت) است؛ نظیر روزهایی که شکستگی در آن اتفاق افتاده است. بنابراین توسعه روش‌هایی به منظور شناسایی و حذف داده‌های غیرعادی یا پرت از سری زمانی داده‌های جریان در مرحله پیش‌پردازش و آماده‌سازی داده‌ها به منظور به‌کارگیری در روش‌های ارزیابی و مدیریت نشت ضروری است.

در یک تعریف کلی، یک داده پرت یا غیرعادی، مشاهده‌ای است که انحراف زیادی از سایـــر مشاهدات داشته باشد؛ بطوریکه به احتمال زیاد این داده به وسیله یک مکانیزم متفاوت ایجاد شده است (Hawkins, 1980). به عبارت دیگر داده پرت به داده‌ای می‌گویند که با رفتارهای مورد انتظار از داده‌ها یا الگوی داده‌ها عادی تطابق ندارد. در شبکه‌های توزیع آب، داده‌های پرت موجود در داده‌های جریان می‌تواند به علت: ۱- تغییرات ناگهانی در الگوی مصرف نظیر مصارف خانگی و غیرخانگی غیرعادی، ۲- تغییرات در سیستم بهره‌برداری از شبکه، ۳- خطاهای ناگهانی در شبکه نظیر شکستگی در خطوط اصلی و انشعابات و ۴- خرابی کنتورها یا مشکلات مربوط به تله‌متری و سیستم اسکادا باشد (Loureiro et al., 2016).

داده‌های پرت ایجاد شده در نتیجه خرابی کنتورها، تله‌متری یا سیستم‌های اسکادا معمولاً به صورت یک تغییر کنترل‌نشده نظیر داده‌های ثبت‌شده با مقادیر بسیار بالا یا بسیار کم، مقادیر صفر، مقادیر منفی و یا یک دوره زمانی طولانی با مقادیر داده ثابت خود را نشان داده و بطور قابل توجهی از بقیه داده‌ها تفاوت دارند. داده‌های پرت ناشی از مصارف غیرعادی خانگی و غیرخانگی، تغییرات قابل‌توجه در رفتار یا الگوی مصرف‌کنندگان را به علت جابه‌جایی جمعیت (مثلاً در روزهای تعطیل رسمی) یا تغییر در مصرف آب (مثلاً به علت آبیاری فضای سبز یا پرکردن استخرها در فصول گرم سال) و یا مصارف غیرمجاز شبانه را نشان می‌دهد (Mazzolani et al., 2017). این داده‌ها همچنین می‌توانند در اثر تغییرات در بهره‌برداری از شبکه، نظیر مانور شیرآلات یا پمپ‌ها ایجاد شوند. در نهایت داده‌های پرت می‌توانند در اثر نشت‌ها و شکستگی‌ها به وجود آیند که منجر به افزایش ناگهانی در داده‌های جریان پایش‌شده می‌شوند (Li et al., 2014). علاوه بر

اهمیت کشف سریع نشت‌ها و شکستگی‌ها در کمترین زمان، تشخیص روند داده‌های پرت در سری زمانی داده‌های جریان می‌تواند در کشف و جلوگیری از وقوع شکستگی‌های جدید، کشف مصارف غیرعادی و در ارزیابی روش تحلیل جریان شبانه و روش‌های آماری با قابلیت اعتماد بیشتر و دقت بالاتر مفید واقع شود (Loureiro et al., 2016).

در طول سال‌های اخیر با توسعه و افزایش دسترسی به تجهیزات اندازه‌گیری، سیستم‌های تله‌متری و اسکادا با قیمت و کیفیت مناسب، پایش داده‌های جریان در زون‌ها و نواحی ایزوله به یک عملیات معمول در سطح شرکت‌های آب و فاضلاب تبدیل شده است. معمولاً داده‌های جریان به صورت پیوسته در نقاط ورودی زون‌های شبکه با فواصل زمانی کمتر از یک ساعت (عموماً هر ۱۵ دقیقه) برداشت می‌شود که می‌تواند حجم بزرگی از داده‌های تاریخی را تولید کند. داده‌های تولید شده عموماً برای بهره‌برداری و کنترل آنلاین شبکه به کار می‌روند و اغلب بعد از مدتی حذف شده یا به اطلاعات با حجم کمتر تبدیل می‌شوند. از طرفی الگوی داده‌های جریان در بین نواحی مختلف یک شبکه، در طول فصول و هفته‌ها و روزها به علت تعداد و ناهمگنی مصرف‌کنندگان، شرایط و روش‌های بهره‌برداری از لوله‌های شبکه می‌تواند به مقدار قابل توجهی تغییر کند. وقوع داده‌های پرت نیز پیچیدگی داده‌ها را افزایش می‌دهد. از این رو تشخیص داده‌های غیرعادی و پرت در سری زمانی داده‌های جریان، به منظور جمع‌آوری داده‌های تاریخی قابل اعتماد جهت به‌کارگیری در روش‌های ارزیابی نشت، بهره‌برداری و برنامه‌ریزی پایدار شبکه‌های توزیع آب ضروری خواهد بود.

در یک دسته‌بندی کلی روش‌های تشخیص داده‌های پرت یا غیرعادی را می‌توان به روش‌های مبتنی بر توزیع یا روش‌های آماری، روش‌های مبتنی بر مجاورت و روش‌های مبتنی بر خوشه‌بندی طبقه‌بندی نمود (Chandola et al., 2009). از میان روش‌های ذکر شده، روش‌های مبتنی بر خوشه‌بندی که در دسته تکنیک‌های یادگیری غیرنظارت شده قرار دارند، ابزاری قدرتمند برای شناسایی داده‌های پرت می‌باشند. هدف از خوشه‌بندی، تقسیم‌کردن داده‌ها به صورتی است که در یک خوشه حداکثر مشابهت میان نقاط داده و بین داده‌ها در خوشه‌های گوناگون، کمترین مشابهت وجود داشته باشد. به عبارت دیگر، هدف از خوشه‌بندی قرار دادن نقاط داده در خوشه‌هایی با حداکثر مشابهت درون خوشه‌ای و حداقل مشابهت میان خوشه‌ای است (Cassisi et al., 2013). انواع مختلفی از الگوریتم‌های خوشه‌بندی وجود دارد که از بین آن‌ها، روش‌های خوشه‌بندی مبتنی بر چگالی به علت منطق قوی و سهولت در استفاده کاربرد بیشتری دارند (Lv et al., 2016). اولین الگوریتم بر اساس روش‌های مبتنی بر چگالی، با عنوان خوشه‌بندی مکانی مبتنی بر چگالی با کاربرد برای داده‌های نویز (DBSCAN) در سال ۱۹۹۶ ارائه شد (Ester et al., 1996). این روش به دلیل توانایی در تشخیص خوشه‌ها با شکل‌های مختلف و پیچیده، نبود اطلاعات اولیه در مورد مجموعه داده‌ها و نیز تشخیص خودکار داده‌های نویز مورد توجه زیادی قرار گرفت. منطق این الگوریتم شناسایی نواحی با چگالی بالا می‌باشد که داده‌های عادی یا نرمال به آن تعلق دارند. در حالیکه داده‌های پرت یا غیرنرمال به خوشه‌های کوچک و کم تراکم تعلق داشته یا به هیچ خوشه‌ای متعلق نیستند (Lv et al., 2016). در شبکه‌های توزیع آب از این الگوریتم بیشتر برای خوشه‌بندی مکانی شکستگی‌های شبکه به منظور تحلیل علل شکستگی لوله‌ها و شناسایی مناطق با خطر بالا به منظور اولویت‌بندی اقدامات بازسازی و نوسازی استفاده شده است (De Oliveira et al., 2011; Oliveira et al., 2009; Sun et al., 2014).

این مقاله یک متدولوژی جدید برای پایش داده‌های جریان و شناسایی داده‌های پرت یا غیرعادی بر مبنای الگوریتم خوشه‌بندی مبتنی بر چگالی ارائه می‌دهد. متدولوژی ارائه شده از سه بخش اصلی تشکیل شده است. در بخش اول داده‌های جریان و حوادث از منابع داده‌ای موجود در شرکت آب و فاضلاب جمع‌آوری می‌گردند؛ در بخش دوم این داده‌ها صحت‌سنجی و نرمال‌سازی گردیده و در گام سوم الگوریتم خوشه‌بندی DBSCAN به منظور شناسایی داده‌های پرت و غیرعادی به کار برده می‌شود. کارایی متدولوژی پیشنهادی نیز توسط کاربرد آن برای یک زون در شبکه توزیع آب تهران نشان داده شد. نتایج نشان‌داد که متدولوژی پیشنهادی، ابزاری قدرتمند در شناسایی و حذف داده‌های پرت یا غیرعادی است. از این روش همچنین می‌توان به عنوان پیش‌نیازی برای روش‌های آماری ارزیابی و پایش نشت نیز استفاده نمود؛ چرا که در اکثر این روش‌ها به داده‌های جریان که از آن داده‌های غیرعادی یا پرت شامل شکستگی‌ها و مصارف غیرمجاز حذف شده است، نیاز دارند.

۲- روش‌شناسی

در سال‌های اخیر با استقرار و توسعه سیستم‌های اسکادا و تجهیزات پیشرفته اندازه‌گیری، حجم انبوهی از داده‌های تاریخی شامل داده‌های جریان و فشار در شبکه‌های توزیع آب ثبت و جمع‌آوری شده‌اند که از طریق آن می‌توان اطلاعات مهمی از جمله نشت شبکه، وضعیت بهره‌برداری و کنترل شبکه را استخراج کرد. از این رو فرآیند پایش داده‌های جریان جهت جمع‌آوری یک مجموعه داده تاریخی قابل اعتماد و معتبر به منظور مدیریت نشت کارآمد و مؤثر، برنامه‌ریزی

پایدار، بهره‌برداری و کنترل بهینه شبکه ضروری و اجتناب‌ناپذیر است. در این مقاله یک متدولوژی سه مرحله‌ای به منظور پایش داده‌های جریان جهت شناسائی داده‌های غیرعادی یا پرت و حذف آنها ارائه شده که در شکل ۱ خلاصه شده است. متدولوژی ارائه شده شامل مراحل: ۱- جمع‌آوری داده، ۲- صحت‌سنجی و نرمال‌سازی داده‌ها و ۳- شناسائی و کشف ناهنجاری‌ها یا داده‌های پرت می‌باشد.

در مرحله اول، داده‌های لحظه‌ای جریان با تواتر زمانی مشخص (عموماً با گام‌های زمانی مساوی ۱۵ دقیقه) از سیستم‌های اسکادا یا تله‌متری موجود جمع‌آوری می‌شوند. علاوه بر این، به منظور صحت‌سنجی داده‌های جریان و توصیف داده‌های پرت، داده‌های مربوط به حوادث و تعمیرات شبکه (نظیر تاریخ وقوع حادثه، تاریخ تعمیر و نوع حادثه)، اطلاعات مربوط به مشخصات کنتورها (نظیر قطر، حداقل و حداکثر جریان عبوری از کنتور) و اطلاعات سیستم ورود و نقل و انتقال داده‌ها (نظیر تواتر زمانی قرائت داده و دقت آن) نیز جمع‌آوری می‌گردند.

در مرحله دوم، داده‌های جریان برای یک گام زمانی معین اعتبارسنجی، پاک‌سازی و نرمال‌سازی خواهند شد. منظور از اعتبارسنجی داده‌ها در این مرحله، شناسایی و تصحیح داده‌های غیرعادی یا پرت به علت مشکلات مربوط به سیستم تله‌متری یا کنتورها می‌باشد. این مشکلات می‌تواند مربوط به انتقال ناقصِ داده‌ها (نظیر خرابی باتری دیتالاگرها)، ورود داده‌های نامناسب (نظیر جریان بالاتر یا کمتر از دامنه اندازه‌گیری کنتورها) یا محدودیت‌های مربوط به ذخیره داده باشد. در این مرحله داده‌های بدون مقدار، مقدار منفی، مقدار صفر و داده‌های با مقادیر کمتر یا بیشتر از آستانه کمینه و بیشنه کنتورها حذف می‌شوند. نرمال‌سازی داده‌ها در ایجاد یک سری داده با یک گام زمانی منظم کمک خواهد کرد. در این مرحله یک گام زمانی مناسب (۱۵ دقیقه‌ای) و یک ماکزیمم فاصله داده‌ای مجاز (۶۰ دقیقه) تعریف می‌شود که بیشتر از این فاصله زمانی درون‌یابی انجام نخواهد گرفت. در مورد داده‌های مفقود در فاصله زمانی کمتر از یک ساعت، مقادیر داده جریان شروع و پایان در هر گام زمانی بوسیله درون‌یابی بدست آمده و مقدار جریان متوسط از روی آن محاسبه شده و قرار داده خواهد شد. در مورد داده‌های مفقود با گام زمانی بیش از یک ساعت، یک رکورد خالی در سری زمانی داده‌ها اضافه می‌شود. اگر چندین کنتور ورودی برای یک ناحیه وجود داشته باشد، داده‌های نرمال شده کنتورها با یکدیگر جمع می‌شوند. اگر در ناحیه، مصرف کنندگان بزرگ مثل کارخانجات وجود داشته باشد که جریان ورودی به آنها به طور پیوسته ثبت می‌شود، مصارف نرمال‌سازی شده آن‌ها از مقادیر جریان ورودی به ناحیه کم می‌شود؛ چرا که رفتار این مصرف‌کنندگان با دیگر مصرف‌کنندگان ناحیه کاملاً فرق می‌کند. پس از نرمال‌سازی داده‌ها، میانگین روزانه (یعنی طی مدت یک شبانه‌روز) و میانگین شبانه (در ساعات وقوع حداقل جریان شبانه) داده‌های جریان به منظور پایش و تشخیص داده‌های غیرعادی و پرت برای مدت یک سال بدست می‌آیند.

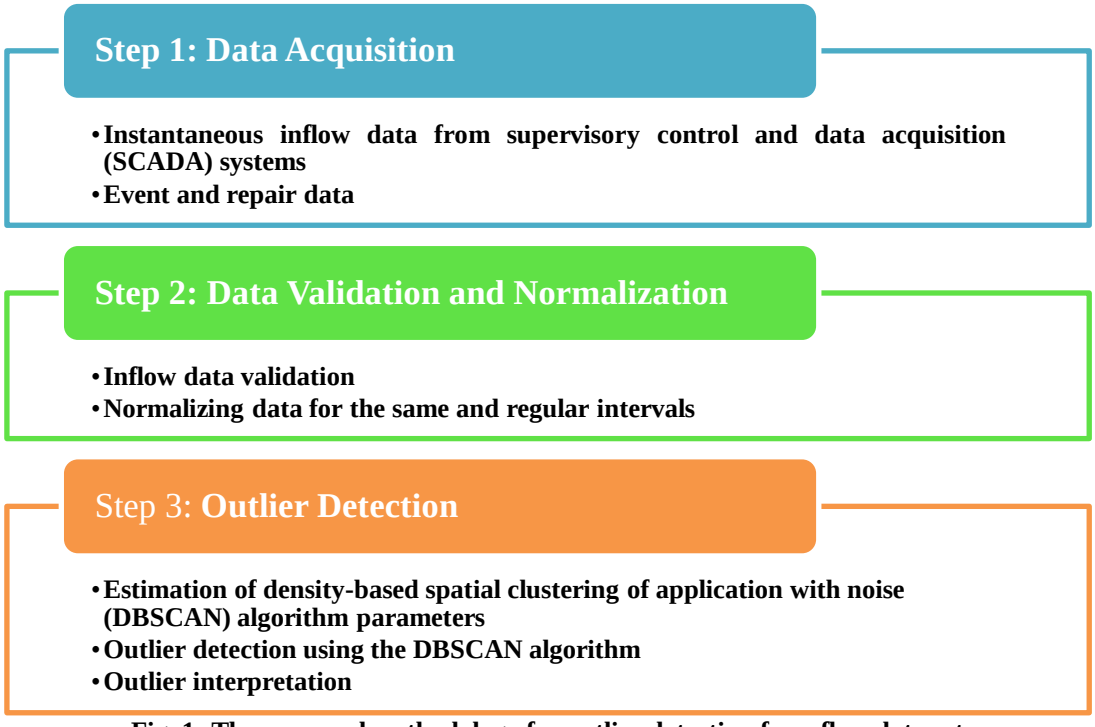


Fig. 1- The proposed methodology for outlier detection for a flow data set

شکل ۱- متدولوژی پیشنهادی جهت شناسایی داده‌های غیرعادی یا پرت برای یک مجموعه داده جریان

بدین منظور با استفاده از سری زمانی داده‌های جریان بدست آمده، ابتدا داده‌های جریان مربوط به جریان حداقل شبانه (بین ساعت ۲ تا ۴ بامداد) استخراج شده و متوسط جریان حداقل شبانه برای هر روز محاسبه می‌شود؛ سپس متوسط جریان روزانه با میانگین‌گیری از سری داده‌های جریان بدست آمده برای هر روز محاسبه می‌شود. در متدولوژی پیشنهادی، از سری زمانی جریان روزانه و جریان حداقل شبانه به صورت هم‌زمان استفاده می‌گردد؛ چرا که عموماً شکستگی‌ها در زمان جریان حداقل شبانه به علت افزایش فشار به وجود آمده رخ می‌دهند و مصارف مجاز غیرعادی نیز اغلب در این ساعات اتفاق می‌افتند. علاوه بر این تغییرات ناگهانی در الگوی مصرف جمعیت را در این حالت بهتر می‌توان شناسایی نمود. بنابراین برای هر روز یک داده دو بعدی شامل متوسط جریان ورودی روزانه و متوسط جریان حداقل شبانه بدست می‌آید که یک سری زمانی داده را برای یکسال تشکیل می‌دهند.

انواع الگوریتم‌های خوشه‌بندی شامل: ۱- الگوریتم‌های سلسله‌مراتبی (نظیر DIANA, AGNES)، ۲- الگوریتم‌های ناحیه‌ای (نظیر k-medoids, k-means)، ۳- الگوریتم‌های بر پایه شبکه‌بندی (CLIQUE, STING)، ۴- الگوریتم‌های مبتنی بر چگالی (DBSCAN, OPTICS)، ۵- الگوریتم‌های گراف‌پایه، ۶- الگوریتم‌های مبتنی بر مدل و ۷- الگوریتم‌های ترکیبی می‌باشند. از بین الگوریتم‌های خوشه‌بندی، الگوریم‌های مبتنی بر چگالی بر اساس چگالی محلی نقاط و اتصال‌ها، خوشه‌بندی را انجام می‌دهند. هدف از الگوریتم‌های خوشه‌بندی مبتنی بر چگالی، تعیین خوشه‌ها با چگالی‌های مختلف و جداسازی آن‌ها از سایر خوشه‌ها بر اساس این چگالی‌ها می‌باشد. در این الگوریتم یک خوشه به وسیله تعدادی از نقاط تعریف می‌شود که نسبت به نقاط مرکزی در یک حداقل فاصله معین قرار دارند. نقاطی که به هیچ خوشه‌ای متعلق نباشند به عنوان نقاط نویز شناخته می‌شوند. روش خوشه‌بندی مبتنی بر چگالی، ویژگی‌های منحصر به فردی به این نوع الگوریتم‌ها می‌دهد که تفاوت آن با دیگر روش‌های خوشه‌بندی را رقم می‌زند. این ویژگی‌ها شامل موارد زیر است، ۱- این الگوریتم خود را به شکل خوشه‌ها محدود نمی‌کنند و در نتیجه توانایی تشخیص خوشه‌ها با شکل و اندازه دلخواه را دارند، ۲- نیازی نیست از قبل تعداد خوشه‌ها را مشخص نمود، ۳- فهم و درک این الگوریتم‌ها و پیاده‌سازی آن‌ها ساده است، ۴- در حضور نویز عملکرد بسیار خوبی دارند و به سادگی نویز را تشخیص می‌دهند (تشخیص و حذف اتوماتیک نویز) و ۵- در یک مرحله انجام می‌گیرند (Jain, 2010; Jain et al., 1999; Lv et al., 2016; Soni and Ganatra, 2012a, 2012b). در مرحله سوم، از الگوریتم خوشه‌بندی مکانی مبتنی بر چگالی با قابلیتِ تشخیص نویز (DBSCAN) برای کشف داده‌های غیرنرمال یا پرت از سری زمانی داده جریان استفاده خواهد شد. الگوریتم DBSCAN، الگوریتم پایهٔ روش‌های خوشه‌بندی مبتنی بر چگالی است که با استفاده از دو پارامتر اصلی تعریف می‌شود. اولین پارامتر، ماکزیمم شعاع همسایگی از یک نقطه مشاهده‌ای (ε) می‌باشد و پارامتر دیگر تعداد حداقل نقاط همسایگی (MinPts) است که شامل کمترین تعداد نقاط داده موجود در این همسایگی می‌باشد. در حقیقت با استفاده از این دو پارامتر حداقل چگالی یک خوشه تعیین می‌شود. با فرض اینکه، یک مجموعه داده با n نقطه یعنی $D=\{x_1,x_2,\ldots,x_n\}$ وجود داشته باشد، هر x_i تعداد d بعد خواهد داشت $x_i=\{x_{i1},x_{i2},\ldots,x_{id}\}$. در DBSCAN سه رابطه متفاوت بین هر دو نقطه (= داده) مختلف می‌تواند وجود داشته باشد که بر اساس این روابط، خوشه‌ها و داده‌های پرت تشخیص داده می‌شوند. این روابط به صورت زیر تعریف می‌شود.

تعریف ۱- دسترس‌پذیری چگالی مستقیم: نقطه p از نقطه q دسترس‌پذیری چگالی مستقیم است، اگر اولاً p جزء همسایه‌های q باشد و ثانیاً q یک نقطه مرکزی باشد (شکل ۲-الف). یعنی:

$$p \in N_\varepsilon(q),\ N_\varepsilon(q)=\{p \mid distance(q,p)\le\varepsilon\} \quad (۱)$$

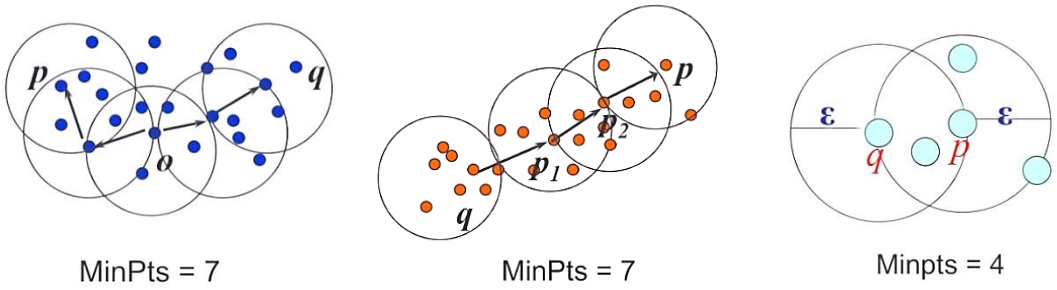
$$|N_\varepsilon(q)| \ge MinPts \quad (۲)$$

در رابطه ۱ فاصله نقطه p از q یا $distance(q,p)$ با توجه به توابع مختلفی که برای محاسبه فاصله موجود است (مانند تابع فاصله منهتن یا فاصله اقلیدسی) می‌تواند مقادیر متفاوتی داشته باشد.

تعریف ۲- دسترس‌پذیری چگالی: نقطه p از نقطه q نسبت به ε و MinPts دسترس‌پذیر چگالی است، اگر یک زنجیره از نقاط (داده) $p_1,p_2,\ldots,p_n$ وجود داشته باشد که در آن $p_1=q$ و $p_n=p$ باشد؛ بطوریکه p_{i+1} دسترس‌پذیری چگالی مستقیم از p_i نسبت به ε و MinPts برای $1\le i\le n$ و $p_i\in D$ باشد (شکل ۲- ب).

تعریف ۳- متصل چگالی: نقطه p متصل چگالی از نقطه q نسبت به ε و MinPts است، اگر یک نقطه مثل o وجود داشته باشد به گونه‌ای که هر دو نقطه p و q دسترس‌پذیری چگالی از نقطه o به نسبت ε و MinPts باشد (شکل ۲-ج).

با توجه به سه رابطه فوق، تمامی نقاط در الگوریتم DBSCAN را می‌توان در سه دسته اصلی شامل نقطه مرکزی، نقطه مرزی و نقطه نویز یا نقطه غیرعادی طبقه‌بندی نمود که به صورت زیر تعریف شده و در شکل ۳ نیز نشان داده شده است.



(c- ج) (b- ب) (a- الف)

Fig. 2- Graphical presentation of the relationship between two points in DBSCAN algorithm, a) Directly density reachable for MinPts= 4, b) density reachable for MinPts= 7 and c) density connected for MinPts= 7

شکل ۲- نمایش گرفیکی روابط دو داده در الگوریتم DBSCAN، الف) دسترس‌پذیری چگالی مستقیم برای MinPts= 4، ب) دسترس‌پذیری چگالی برای MinPts= 7 و ج) متصل چگالی برای MinPts= 7

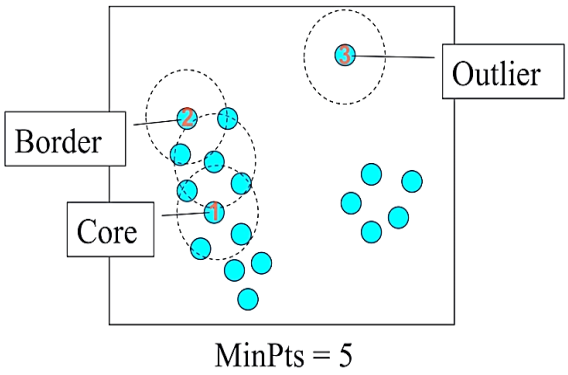


Fig. 3- Illustration of the points in DBSCAN algorithm 1- Core point, 2- Border point and 3- Noise or outlier for MinPts= 5

شکل ۳- تعریف نقاط در الگوریتم DBSCAN، ۱- نقطه مرکزی، ۲- نقطه مرزی و ۳- نویز یا داده پرت برای MinPts= 5

تعریف ۴- نقطه مرکزی: اگر تعداد نقاطی که از نقطه p دسترس‌پذیری چگالی مستقیم دارند، از کمترین تعداد نقاط (MinPts) در همسایگی شعاع ε نقطه p (یعنی $N_\varepsilon(p)$) بیشتر باشند، آنگاه نقطه p یک نقطه مرکزی است.

تعریف ۵- نقطه مرزی: اگر تعداد نقاط در همسایگی شعاع ε یک نقطه p (یعنی $N_\varepsilon(p)$) بیشتر از (MinPts) نباشد و نقطه P دسترس‌پذیری چگالی مستقیم از یک نقطه مرکزی داشته باشد، آنگاه نقطه p یک نقطه مرزی خواهد بود.

تعریف ۶- نویز: اگر نقطه P نه یک نقطه مرکزی و نه یک نقطه مرزی باشد، آنگاه نقطه p یک نقطه نویز یا یک داده غیرعادی (پرت) می‌باشد.

در الگوریتم DBSCAN یک خوشه (C) با شعاع همسایگی (ε) و حداقل تعداد نقاط همسایگی (MinPts) یک زیر مجموعه غیرتهی از D می‌باشد که دو شرط زیر را ارضا کند:

شرط حداکثر بودن: به ازای هر جفت نقطه p و q، اگر $p \in C$، یعنی p یکی از اعضای خوشه C باشد و q دسترس‌پذیر چگالی از p نسبت به شعاع همسایگی (ε) و کمترین تعداد نقاط همسایگی (MinPts) باشد، آنگاه q نیز متعلق به خوشه C است.

شرط اتصال: به ازای هر جفت نقطه p و q عضو خوشه C، نقطه p متصل چگالی از نقطه q نسبت به شعاع همسایگی ε و کمترین تعداد نقاط همسایگی MinPts خواهد بود.

الگوریتم DBSCAN از یک نقطه اختیاری مانند p از مجموع داده شروع می‌کند تا تعداد نقاط در شعاع همسایگی ε از نقطه p (دسترسی‌پذیری چگالی نقطه p) را محاسبه کند. اگر نقطه p یک نقطه مرکزی باشد، الگوریتم DBSCAN این نقطه را به عنوان یک خوشه جدید علامت‌گذاری می‌کند، سپس تمامی نقاط دسترس‌پذیر چگالی از نقطه p را بازیابی می‌کند و آنها را با همان برچسب خوشه p علامت‌گذاری می‌کند. در غیراین صورت، نقطه p به عنوان یک نقطه نویز برچسب‌گذاری می‌شود. در مرحله بعد، الگوریتم DBSCAN به صورت تکراری نقاطی که دسترس‌پذیر چگالی از نقاط مرکزی هستند را جمع‌آوری می‌کند. این فرآیند هنگامی به پایان می‌رسد که نقطه جدید دیگری را نتوان به هیچ خوشه‌ای اضافه کرد. در ضمن، اگر یک نقطه در یک همسایگی ε مشخص شده از هیچکدام از خوشه‌ها نباشد،

آن نقطه به عنوان یک نقطه نویز در نظر گرفته می‌شود که یک داده پرت محتمل یا امکان‌پذیر است. الگوریتم DBSCAN فقط یک مرتبه نقاط موجود در مجموعه داده D را اسکن می‌کند و به محاسبه فاصله هر جفت نقطه در مجموعه داده D نیاز است. بنابراین پیچیدگی محاسباتی کل الگوریتم از مرتبه $O(n^2)$ خواهد بود که در آن، n تعداد نقاط در مجموعه داده می‌باشد. اگر از ساختارهای شاخصی مؤثر و کارآمد استفاده شود و همچنین بُعد نقاط پایین باشد ($d \leq 5$)، پیچیدگی محاسباتی الگوریتم DBSCAN می‌تواند به مرتبه $O(n \log n)$ کاهش یابد.

همانگونه که بیان شد، به منظور اجرای الگوریتم DBSCAN به دو پارامتر ε و MinPts نیاز است. در این الگوریتم تخمین پارامتر حداقل تعداد نقاط همسایگی (MinPts) آسان‌تر است؛ که هدف آن هموار کردن تخمین چگالی هر یک از خوشه‌ها می‌باشد. Sander et al. (1998) پیشنهاد دادند که مقدار این پارامتر، دو برابر تعداد بُعد مجموعه داده یعنی MinPts=2×dim قرار داده شود. نتایج مطالعات قبلی نشان‌داده که برای مجموعه داده‌های با ابعاد بالا، با داده‌های تکراری و یا با تعداد داده‌های نویز زیاد، نتایج بدست آمده از الگوریتم با افزایش تعداد حداقل نقاط همسایگی بهبود می‌یابد (Daszykowski et al., 2001; De Oliveira et al., 2011; Schubert et al., 2017). تنظیم و تخمین پارامتر ε به مراتب مشکل‌تر بوده و باید تا حد امکان کوچک انتخاب شود. مقدار ε به تابع فاصله نیز بستگی دارد (به عنوان مثال فاصله اقلیدسی یا فاصله منهتن). در حالت ایده‌ال، باید دانش و آگاهی کافی در مورد نوع داده و دامنه تغییرات آن برای انتخاب این پارامتر بر اساس محیط کاربرد آن وجود داشته باشد (Han et al., 2001).

روش دیگری نیز برای انتخاب پارامتر ε بر مبنای ترسیم منحنی K–امین نزدیکترین همسایگی ارائه شده است (Ester et al., 1996; Sander et al., 1998). در این روش ابتدا مقدار K با توجه به رابطه $MinPts = k + 1$ بین حداقل تعداد نقاط همسایگی در نظر گرفته شده و K–امین نزدیکترین همسایگی بدست می‌آید. علت وجود این رابطه آن است که در K–امین نزدیکترین همسایگی برای نقطه p، این نقطه در نظر گرفته نمی‌شود. اما نقطه p برای ایجاد خوشه یا تخمین چگالی به حساب آورده می‌شود (Schubert et al., 2017). سپس تابع فاصله K–ام محاسبه گردیده که فاصله K–امین نزدیکترین همسایگی برای هر نقطه در مجموعه داده می‌باشد. در نهایت نقاط مجموعه داده با توجه به مقادیر فاصله K–امین نزدیکترین همسایگی هر نقطه به صورت نزولی مرتب شده و روی یک نمودار ترسیم می‌گردند تا منحنی فاصله K–ام مرتب شده به دست آید. این منحنی می‌تواند پیشنهاداتی در مورد توزیع چگالی در مجموعه داده و نحوه انتخاب پارامتر ε را ارائه دهد. هدف از ترسیم این منحنی تشخیص یک نقطه آستانه با بیشترین مقدار فاصله K–ام در کم‌جمعیت‌ترین یا کم‌تراکم‌ترین خوشه در مجموعه داده می‌باشد، که مقدار مطلوب پارامتر ε را بدست می‌دهد (Sander et al., 1998; Schubert et al., 2017). در حالت کلی، می‌توان مقدار صحیح پارامتر ε را حول نقطه انحناء یا خمیدگی منحنی در محل تغییر چشمگیر و شدید این نقطه روی گراف بدست آورد (Sun et al., 2014). شکل ۴، گراف فاصله سوم مرتب شده را برای یک مجموعه داده نشان می‌دهد. نقطه آستانه برای تشخیص نقاط نویز از نقاط عادی نیز در این شکل با توجه به مفاهیم گفته شده، نشان داده شده است. پس از تعیین پارامترهای الگوریتم DBSCAN و اجرای الگوریتم، دو شاخص جهت سنجش کیفیت نتایج بدست آمده ارائه شده که شامل شاخص درصد داده‌های نویز و اندازه بزرگترین خوشه می‌باشد. نتایج مطالعات قبلی پیشنهاد داده‌اند که درصد داده‌های نویز بین ۳۰–۱ درصد داده‌ها باشد. اندازه بزرگترین خوشه به نوع داده و بُعد آن بستگی داشته و عموماً بین ۸۰–۲۰ درصد داده‌ها گزارش شده است (Ester et al., 1996; Han et al., 2001; Liu et al., 2007; Schubert et al., 2017).

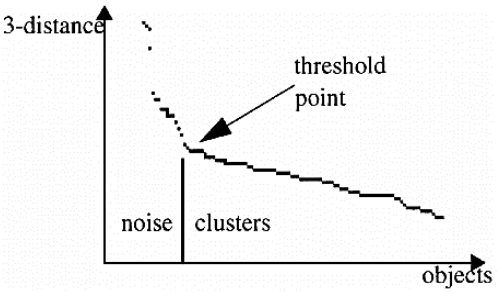


Fig. 4- The sorted K-st curve for k=3
شکل ۴- منحنی K–امین فاصله مرتب شده برای K=3

۳- مطالعه موردی

متدولوژی پیشنهادی در این پژوهش برای یک ناحیه (که در اینجا ناحیه الف نامیده می‌شود) در شبکهٔ توزیع آب شهر تهران به کار برده شد. ناحیه مذکور به صورت ثقلی تغذیه گردیده؛ طول کل خطوط اصلی حدود ۹۶ کیلومتر و تعداد انشعابات ناحیه ۱۸۵۶۰ انشعاب بود. علاوه بر این اندازه‌گیری جریان در نقطه ورودی به ناحیه انجام گرفت. مرحله اول و دوم از متدولوژی پیشنهادی مربوط به جمع‌آوری، پیش‌پردازش و آماده‌سازی داده‌های جریان، مربوط می‌شود. در مرحله اول داده‌های خام جریان ورودی به زون الف با تواتر زمانی ۱۵ دقیقه برای سال ۱۳۹۴ از سیستم اسکادا اخذ گردید. علاوه بر این، داده‌های مربوط به

حوادث لوله‌های اصلی شبکه توزیع به منظور صحت‌سنجی متدلوژی پیشنهادی جمع‌آوری و مرتب‌سازی گردید. در مرحله دوم داده‌های جمع‌آوری شده اعتبارسنجی، پاکسازی و نرمال‌سازی شدند. بدین منظور ابتدا داده‌های با مقادیر منفی، صفر و بدون مقدار به منظور اعتبارسنجی داده‌ها، شناسایی و حذف گردید. سپس مقادیر میانگین برای داده‌های بدون‌مقدار یا مقدار حذف‌شده در بازه زمانی کمتر از یک ساعت محاسبه و مقدار آن قرار داده شد. در ادامه سری زمانی داده‌های جریان حداقل شبانه از سری زمانی داده‌های جریان بدست آمده استخراج گردیده و متوسط جریان حداقل شبانه برای هر روز محاسبه گردید؛ متوسط جریان روزانه نیز با استفاده از سری زمانی داده‌های جریان بدست آمده محاسبه شد. در نهایت یک سـری زمانی دوبعدی شامل متوسط جریان حداقل شبانه و متوسط جریان روزانه برای یک‌سال بدست آمد. شکل ۵، نحوه پراکندگی داده‌ها را برای سال ۱۳۹۴ نشان می‌دهد. شکل ۶ نیز، سری زمانی داده‌های متوسط جریان روزانه و متوسط جریان حداقل شبانه را برای سال ۱۳۹۴ نشان می‌دهند.

۴-تعیین پارامترهای الگوریتم DBSCAN

نتایج حاصل از الگوریتم DBDCAN با تنظیم دو پارامتر ε و MinPts کنترل می‌گردند. از این رو تعیین صحیح این دو پارامتر تأثیر بسزایی در دقت و قابلیت اطمینان نتایج بدست آمده از این الگوریتم خواهد گذاشت.

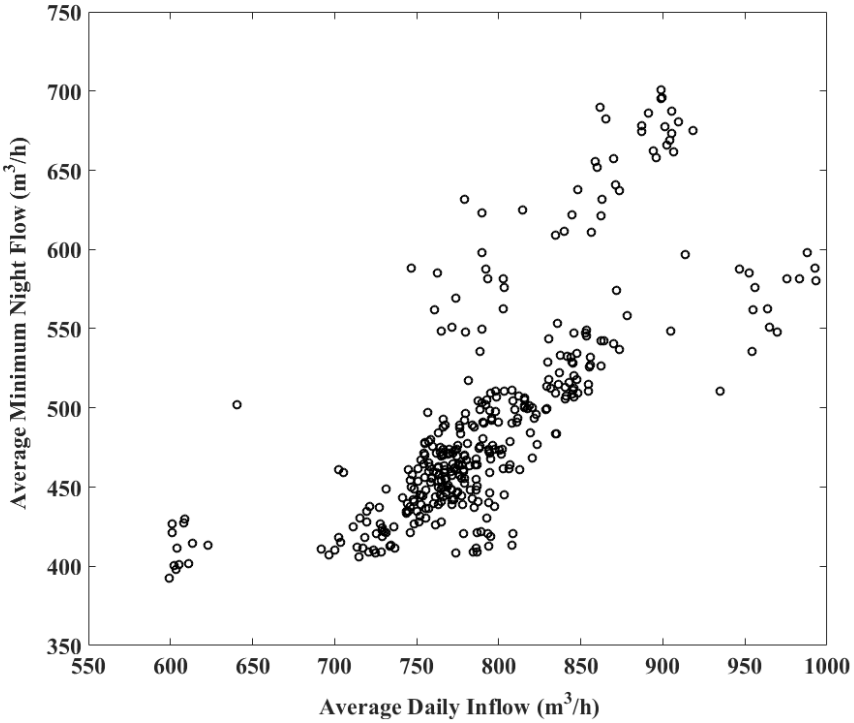


Fig. 5- Two-dimensional flow data (average daily vs. average nightly flow) in the study area for the year 1394

شکل ۵- متوسط روزانه و شبانه داده‌های جریان ورودی به ناحیه مورد مطالعه در سال ۱۳۹۴

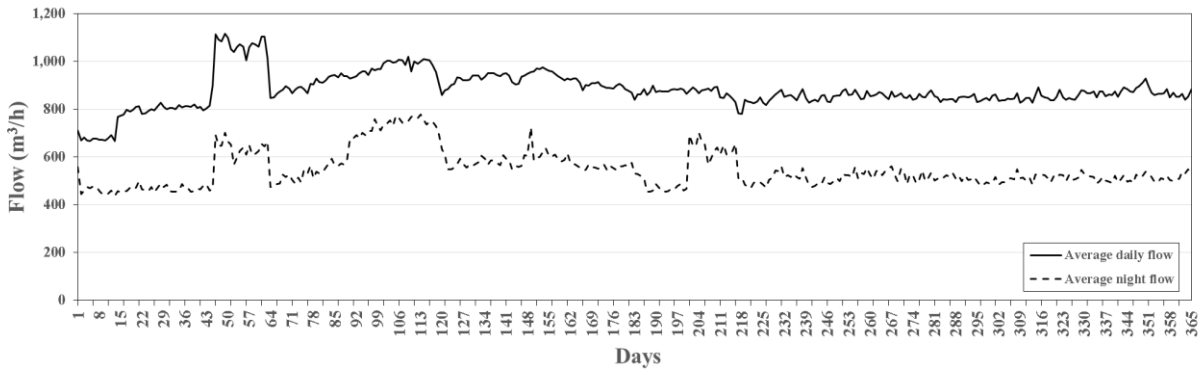


Fig. 6- Average daily and nightly flow data time series in the study area for the year 1394

شکل ۶- داده‌های متوسط دبی روزانه و شبانه سال ۱۳۹۴ در منطقه مورد مطالعه

در اولین گام برای تعیین تعداد حداقل نقاط همسایگی و با توجه به بُعد داده‌های جریان در این مطالعه، شامل متوسط جریان حداقل شبانه و متوسط جریان روزانه برای یک‌سال، dim=2 در نظر گرفته شد؛ تعداد حداقل نقاط همسایگی، دو برابر بعد داده‌ها یعنی $MinPts = 4$ قرار داده شد. علاوه بر این تعداد نقاط همسایگی ۶ و ۱۰ که به ترتیب سه و پنج برابر بُعد داده‌های جریان بودند، به منظور تعیین تعداد حداقل نقاط همسایگی مناسب، در این مطالعه نیز در نظر گرفته شد. سپس با استفاده از حداقل نقاط همسایگی در نظر گرفته شده، گراف K–امین نزدیکترین همسایه (K=MinPts-1) به منظور تعیین دامنه شعاع همسایگی بهینه ترسیم گردید. شکل ۷، منحنی K–امین نزدیکترین همسایگی را برای مقادیر $k = 3,5,9$ نشان می‌دهد. با استفاده از هر یک از منحنی‌های K–امین نزدیکترین همسایگی می‌توان یک دامنه برای شعاع همسایگی با توجه به محدوده انحناء یا خمیدگی منحنی‌ها بدست آورد که شعاع همسایگی بهینه در آن دامنه قرار دارد. همانگونه که در شکل ۷ مشخص است با افزایش تعداد نقاط حداقل همسایگی از ۴ تا ۱۰، محدوده انحناء یا خمیدگی نیز در منحنی‌های K–امین نزدیکترین همسایگی افزایش می‌یابد. چرا که با افزایش مقادیر K (K=MinPts-1)، مقادیر فواصل K–امین نزدیکترین همسایگی برای هر یک از داده‌های جریان نیز افزایش خواهد یافت. با توجه به شکل ۷، دامنه مقادیر فاصله K–امین نزدیکترین همسایگی در محدوده انحناء یا خمیدگی منحنی‌ها به صورت تقریبی و محافظه‌کارانه می‌تواند بین مقادیر ۵ تا ۲۵ تغییر نماید. از این رو مقدار پارامتر شعاع همسایگی بهینه (ε) برای هر یک از نقاط همسایگی انتخاب شده به احتمال زیاد در این دامنه قرار دارد. علاوه بر این در شکل ۷ می‌توان مشاهده نمود که برای حدود ۷۰ درصد از داده‌ها در سه منحنی رسم شده در شکل ۷، روند تغییرات شیب آنها مشابه و نزدیک به هم بوده و پس از آن انحراف منحنی‌ها از یکدیگر زیاد می‌شود. این روند تغییرات نشان‌دهنده آن است که مقادیر بخش بزرگی از داده‌ها در محدوده نزدیکی از یکدیگر قرار داشته و در آن محدوده تراکم بالایی دارند. از این رو با احتمال زیادی می‌توان انتظار داشت که بزرگترین خوشه درصد زیادی از داده‌ها را به خود اختصاص می‌دهد.

با انتخاب یک نقطه اختیاری روی هر یک از منحنی‌های K–امین نزدیکترین همسایگی مرتب شده در شکل ۷ (برای مثال 3-Dist Curve) و با قرار دادن مقدار شعاع همسایگی (ε) برابر با فاصله سومین نزدیکترین همسایگی آن نقطه، تمامی نقاط با فاصله برابر یا کوچکتر از مقدار فاصله سومین نزدیکترین همسایگی نقطه انتخاب شده، یک نقطه مرکزی بوده و در یکی از خوشه‌ها قرار می‌گیرند.

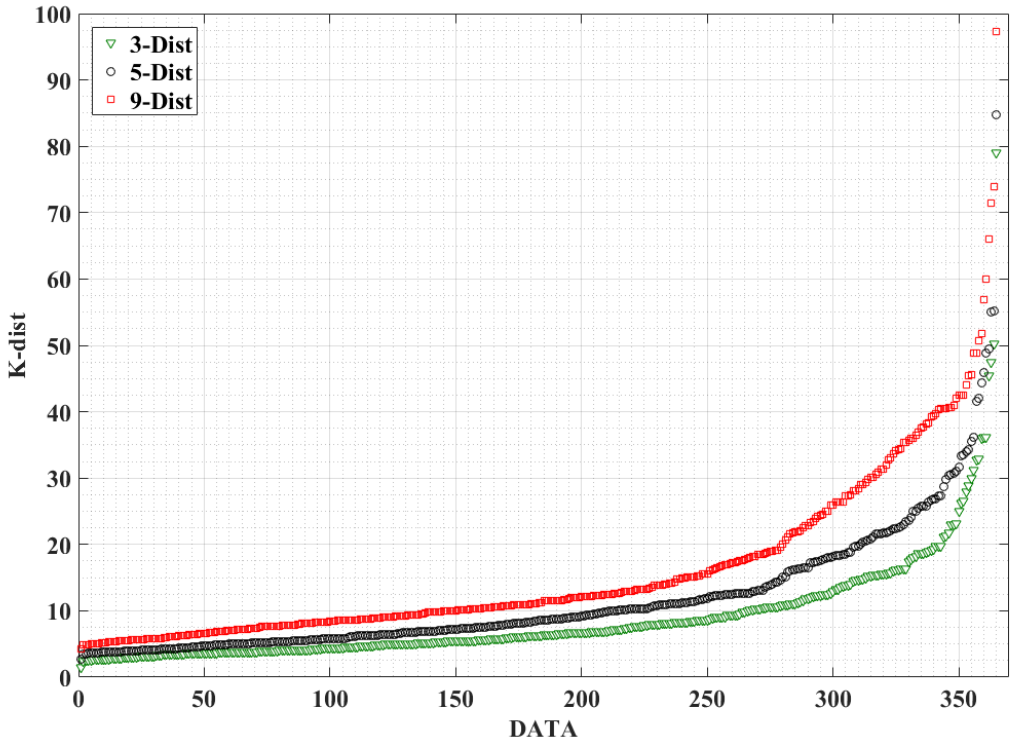


Fig. 7- The sorted K nearest neighbor curve for the inflow data set in the study area (K=3, 5, 9)

شکل ۷- منحنی K- امین نزدیکترین همسایگی مرتب شده برای مجموعه داده جریان ناحیه مورد مطالعه (K=3, 5, 9)

از این رو باید نقطه‌ای با بیشترین مقدار فاصله سومین همسایگی در کم تراکم‌ترین خوشه (یعنی خوشه‌ای با تعداد نقاط برابر ۴) را روی منحنی سومین نزدیک‌ترین همسایگی مرتب شده به منظور تعیین مقدار شعاع همسایگی بهینه پیدا نمود. همانگونه که قبلاً بیان گردید، این فاصله بهترین مقدار را برای شعاع همسایگی بهینه ارائه می‌دهد که اغلب حول نقطه انحناء یا خمیدگی منحنی یاد شده بدست می‌آید. تا قبل از رسیدن به این فاصله، با افزایش شعاع همسایگی رشد تعداد نقاط مرکزی سریع بوده که منجر به آن می‌شود که تعداد نقاط بیشتری در خوشه‌ها قرار گرفته (افزایش تعداد نقاط خوشه‌بندی) و اندازه بزرگترین خوشه رشد کند؛ در عین حال ممکن است منجر به آن بشود که چندین خوشه به یک خوشه پیوسته و اغلب تعداد خوشه‌ها کاهش یابند (Schubert et al., 2017). پس از شعاع همسایگی بهینه و در یک دامنه از شعاع‌های همسایگی بزرگتر، اغلب اندازه بزرگترین خوشه و تعداد خوشه‌ها ثابت مانده یا تغییرات اندکی دارند. چرا که نیازی نیست تا کل نقاط موجود در کم‌تراکم‌ترین خوشه، نقطه مرکزی باشند. علاوه بر این اگر این نقاط فقط متصل چگالی باشند نیز می‌توانند به کم‌تراکم‌ترین خوشه متعلق باشند. بنابراین یک دامنه برای شعاع‌های همسایگی بهینه می‌توان بدست آورد (Sander et al., 1998; Sun et al., 2014). در مرحله بعد، الگوریتم DBSCAN، با توجه به دامنه بدست آمده برای شعاع همسایگی و تعداد حداقل نقاط همسایگی انتخاب شده برای تعیین پارامتر شعاع همسایگی بهینه اجرا گردید. شکل ۸، نمودار درصد داده‌های خوشه‌بندی شده، داده‌های نویز و داده‌های موجود در بزرگترین خوشه را برای تعداد حداقل نقاط همسایگی ۴ و دامنه تعیین‌شده برای شعاع همسایگی را نشان می‌دهد. همانگونه که در شکل ۸ می‌توان مشاهده کرد، اندازه بزرگترین خوشه از شعاع همسایگی ۵ تا ۱۲ رشد سریعی داشته، بطوریکه اندازه آن از ۲۵ به حدود ۷۵ درصد از داده‌ها می‌رسد. از شعاع همسایگی ۱۹–۱۲ تقریباً درصد داده‌های بزرگترین خوشه تقریباً ثابت مانده و تغییرات اندکی دارد (بین ۷۷–۷۵ درصد از داده‌ها). بنابراین دامنه‌ای از شعاع همسایگی بهینه بدست می‌آید که در آن اندازه بزرگترین خوشه تقریباً ثابت است. علاوه بر این شاخص درصد داده‌های غیر عادی یا نویز، برای این دامنه در محدوده مجاز قرار داشته و بین ۱۵–۵ درصد تغییر می‌کند. مطالعات قبلی نشان داده‌اند که مقادیر کوچکتر برای شعاع همسایگی نتایج بهتری ارائه می‌دهند (Schubert et al., 2017). از این رو در این پژوهش برای تعداد حداقل نقاط همسایگی $MinPts = 4$، مقدار شعاع همسایگی بهینه $\varepsilon = 12$ پیشنهاد می‌گردد.

شکل ۹، تعداد خوشه‌های تولید شده در الگوریتم DBSCAN را بر حسب شعاع همسایگی برای تعداد حداقل نقاط همسایگی $MinPts = 4$ نشان می‌دهد. همانگونه که در شکل مشخص است، برای شعاع همسایگی ۸ تا ۱۲ با افزایش شعاع همسایگی، تعداد خوشه‌ها از ۱۳ به ۵ خوشه کاهش یافته که دلیل آن می‌تواند رشد سریع بزرگترین خوشه به علت تراکم بالای درصد بزرگی از داده‌ها در این محدوده از شعاع همسایگی باشد.

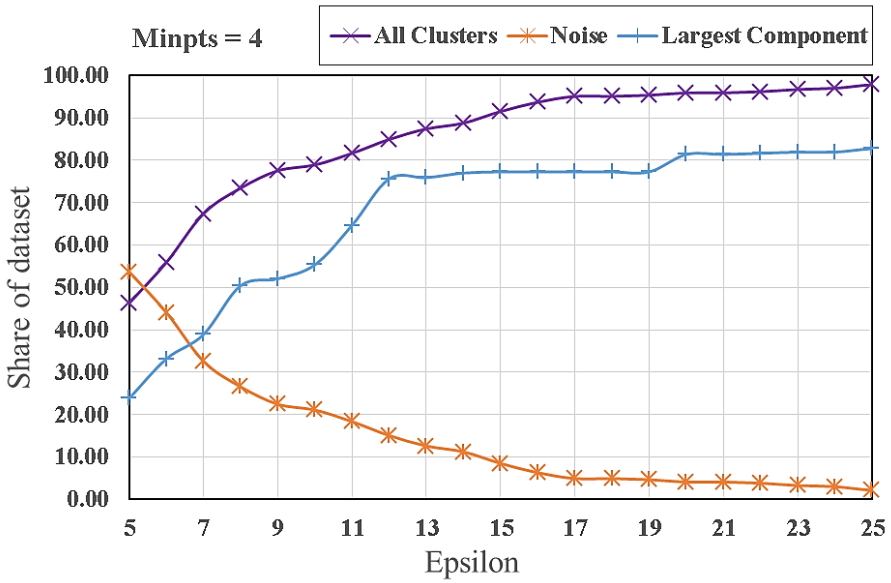


Fig. 8- The percent of clustered data, the size of largest cluster and the percent of outliers by DBSCAN for MinPts= 4

شکل ۸- درصد داده‌های خوشه‌بندی شده، اندازه بزرگترین خوشه و درصد داده‌های نویز و توسط الگوریتم DBSCAN برای MinPts= 4

اما در دامنه شعاع همسایگی بهینه (یعنی از ۱۹–۱۲) تعداد خوشه‌ها روند افزایشی داشته و از ۵ به ۸ خوشه رسیده است، در صورتیکه اغلب در دامنه شعاع همسایگی بهینه تعداد خوشه‌ها ثابت یا تغییرات اندکی دارد؛ این حالت به علت تشکیل خوشه‌های زائد به وجود آمده است. افزایش تعداد خوشه‌ها با افزایش شعاع همسایگی در این دامنه می‌تواند به علت تعداد حداقل نقاط همسایگی پایین (MinPts=4) در تشکیل یک خوشه و نزدیکی و تراکم داده‌ها به یکدیگر باشد که منجر به عدم تشخیص صحیح خوشه‌ها توسط الگوریتم می‌شود. علاوه بر این تعداد داده‌های نویز بالای مجموعه داده (۱۵–۵ درصد) نیز تأثیر گذار است؛ بطوریکه افزایش شعاع همسایگی در دامنه ذکر شده احتمالاً منجر به تولید خوشه‌های جدیدی شده است. یکی از روش‌ها برای بهبود نتایج در چنین حالتی افزایش تعداد حداقل نقاط همسایگی است. بنابراین در این مطالعه، الگوریتم DBSCAN برای شعاع همسایگی با تعداد حداقل نقاط همسایگی MinPts=6,10 که به ترتیب ۳ و ۵ برابر بعد داده‌ها می‌باشد نیز اجرا گردید.

شکل ۱۰، درصد داده‌های بزرگترین خوشه، داده‌های نویز و کل داده‌های خوشه‌بندی شده را برای تعداد حداقل همسایگی MinPts=6,10 نشان می‌دهد. همانگونه که در شکل ۱۰–الف مشخص است، برای تعداد حداقل نقاط همسایگی MinPts=6، در دامنه شعاع همسایگی ۲۵–۱۳ اندازه بزرگترین خوشه تقریباً ثابت می‌ماند و ۷۸–۷۶ درصد از داده‌ها را به خود اختصاص می‌دهد. در حالیکه این دامنه برای شعاع همسایگی MinPts=10 در شکل ۱۰–ب بین ۲۵–۱۶ تغییر نموده که همان درصد از داده‌ها را شامل می‌شود. بنابراین شعاع همسایگی بهینه می‌تواند در دامنه‌های بدست آمده قرار گیرد. از طرفی با توجه به شکل ۱۰ می‌توان مشاهده نمود که دامنه شعاع همسایگی بهینه با افزایش تعداد حداقل نقاط همسایگی از ۶ به ۱۰ کاهش می‌یابد. این روند به دلیل کاهش نقاط مرکزی و افزایش تعداد نقاط نویز ناشی از افزایش حداقل نقاط همسایگی رخ می‌دهد. در دامنه‌های بدست آمده تغییرات درصد داده‌های غیرعادی برای MinPts=6 بین (۲۰–۴ درصد) و برای MinPts=10 بین (۲۴–۱۴ درصد) می‌باشد که برای هر دو در محدوده مجاز برای درصد داده‌های نویز قرار دارد. فاکتور محدود کننده دوم در دامنه شعاع همسایگی بهینه، تعداد خوشه‌های تشکیل شده توسط الگوریتم است. شکل ۱۱، تعداد کل خوشه‌های تولید شده را برای تعداد حداقل نقاط همسایگی MinPts=6,10 و دامنه شعاع همسایگی بهینه بدست آمده را ارائه می‌دهد. همانگونه که در شکل ۱۱–الف مشخص است، برای تعداد حداقل نقاط همسایگی MinPts=6 و دامنه شعاع همسایگی ۱۷–۱۵ تعداد خوشه‌ها ثابت مانده است. بنابراین با در نظر گرفتن اندازه بزرگترین خوشه و تعداد خوشه‌های تولید شده، دامنه شعاع همسایگی بهینه برای MinPts=6 بین ۱۷–۱۴ بدست می‌آید. همچنین با توجه به شکل ۱۱–ب و در نظر گرفتن فاکتورهای ذکر شده دامنه شعاع همسایگی بهینه برای MinPts=10 بین ۲۵–۱۸ بدست می‌آید. همانگونه که قبلاً ذکر گردید با توجه به پیشنهادات مطالعات قبلی، مقدار شعاع همسایگی بهینه کمترین مقدار در دامنه شعاع همسایگی بهینه انتخاب می‌گردد. از این رو در این مطالعه برای MinPts=6,10 به ترتیب شعاع همبستگی $\varepsilon = 15,19$ پیشنهاد می‌شود. علاوه بر این به منظور کاهش حجم محاسبات بهتر است، تعداد حداقل نقاط همسایگی کمتر در نظر گرفته شود.

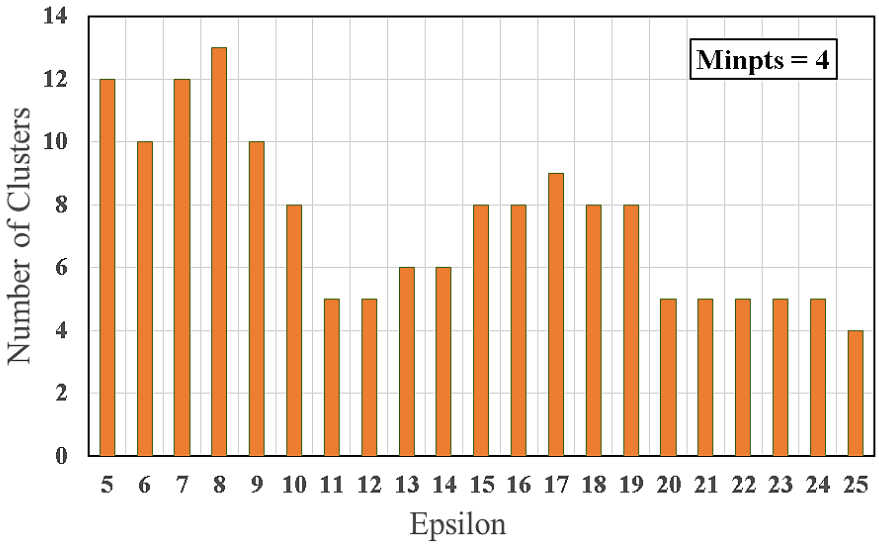


Fig. 9- The number of generated clusters by DBSCAN algorithm for MinPts= 4

شکل ۹- تعداد خوشه‌های تولید شده توسط الگوریتم DBSCAN برای MinPts= 4

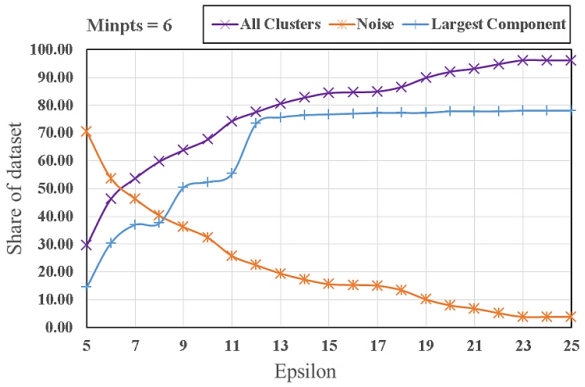


Fig a –شکل الف

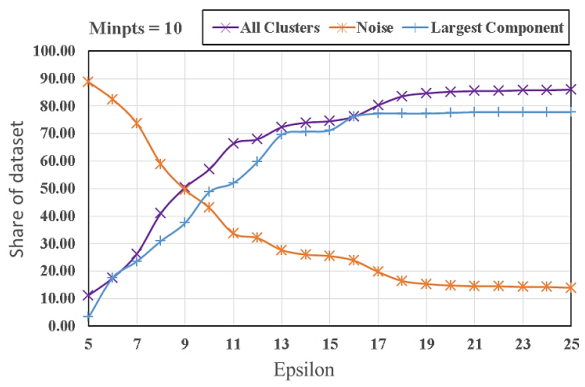


Fig b –شکل ب

Fig. 10- The percent of clustered data, the size of the largest cluster and the percent of outliers by DBSCAN for, a) MinPts= 6 and b) MinPts= 10

شکل ۱۰- درصد داده‌های خوشه‌بندی شده، اندازه بزرگترین خوشه و تعداد داده‌های نویز و توسط الگوریتم DBSCAN برای الف) MinPts= 6 و ب) MinPts= 10

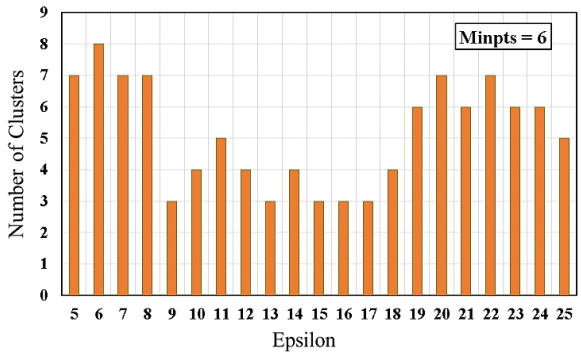


Fig a –شکل الف

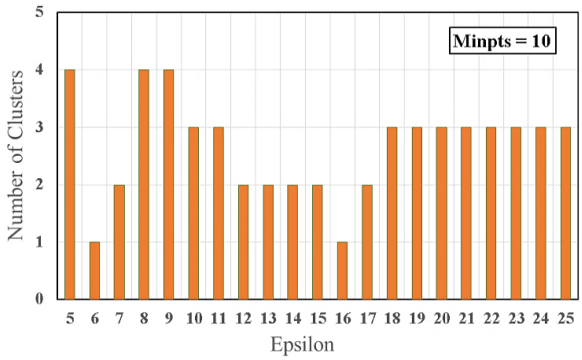


Fig b –شکل ب

Fig. 11- The number of generated clusters by DBSCAN algorithm for, a) MinPts= 6 and b) MinPts= 10

شکل ۱۱- تعداد خوشه‌های تولید شده توسط الگوریتم DBSCAN برای الف) MinPts= 6 و ب) MinPts= 10

شکل ۱۲ نمایشی گرافیکی از نتایج اجرای الگوریتم DBSCAN برای حداقل تعداد همسایگی MinPts=4 و شعاع همسایگی بهینه $\varepsilon = 12$ را ارائه می‌دهد. به ازای پارامترهای تعیین شده، تعداد خوشه‌های بدست آمده برابر ۵ و تعداد داده‌های نویز ۵۵ داده می‌باشد که حدود ۱۵ درصد از کل داده‌ها را تشکیل می‌دهد. علاوه بر این تعداد داده‌های بزرگترین خوشه، ۲۷۶ داده بوده که حدود ۷۶ درصد از کل داده‌ها را شامل می‌شود. علت ایجاد چنین خوشه‌ای، وجود الگوی مصرف تقریباً مشابه در اکثر روزهای سال است که سبب شده بیش از سه‌چهارم داده‌ها در بزرگترین خوشه تولید شده قرار گیرند. دلایل به وجود آمدن چنین الگویی در ناحیه مورد مطالعه عبارتند از: ۱- خصوصیات زون مورد مطالعه از نظر میزان تقاضای روزانه و دامنه تغییرات آن و جمعیت تحت پوشش. متوسط جریان ورودی به زون در سال ۱۳۹۴ حدود ۱۹۰۰۰ مترمکعب در روز بوده و ضریب تغییرات آن (انحراف معیار به میانگین) در کل سال حدود ۸ درصد می‌باشد. علاوه بر این، جمعیت تحت پوشش زون مورد مطالعه حدود هشتاد هزار نفر می‌باشد که این جمعیت در اغلب روزهای سال تقریباً ثابت بوده و تغییرات قابل توجهی ندارد. از این رو متوسط جریان روزانه نسبتاً بالایی با شدت تغییرات پایین وارد این ناحیه می‌شود، ۲- نوع مصارف غالب در ناحیه مورد مطالعه که عمدتاً از نوع مصارف خانگی و تجاری بوده و مصرف کنندگان بزرگ با الگوی مصرف خاص در آن وجود ندارد و ۳- رژیم بهره‌برداری از شبکه در طول سال نرمال و تقریباً ثابت بوده که به دلیل استفاده از سامانه‌های مدیریت فشار هوشمند و استقرار سامانه اسکادا حاصل شده است. شکل ۱۲ همچنین نشان می‌دهد که در این مجموعه داده، دو خوشه (شماره ۳ و ۵) با حداقل تعداد نقاط همسایگی در اطراف داده‌های نویز تشکیل شده که علت تشکیل آنها می‌تواند وجود تعداد بالای داده‌های نویز در این مجموعه داده و تعداد پایین حداقل نقاط

همسایگی برای تشکیل یک خوشه باشد. از این رو همانگونه که قبلاً بیان شد برای بهبود نتایج الگوریتم DBSCAN، این الگوریتم برای تعداد حداقل نقاط همسایگی ۶ و ۱۰ و شعاع همسایگی بهینه تعیین شده نیز مجدداً اجرا گردید که نمایش گرافیکی نتایج آن در شکل ۱۳ ارائه شده است. شکل ۱۳–الف و ب نشان می‌دهد که با افزایش تعداد حداقل نقاط همسایگی و برای شعاع همسایگی بهینه تعیین شده، تعداد خوشه‌های تولید شده برای هر دو حالت کاهش یافته و برابر ۳ خوشه به دست آمده است.

با دقت در خوشه‌های تولید شده در شکل ۱۳ مشخص گردید که خوشه شماره یک، دارای ۱۳ عضو می‌باشد. با استخراج روزهای مربوط به این خوشه از نتایج الگوریتم و انطباق آن با روزهای سال معلوم گردید که کل اعضای این خوشه مربوط به ۱۳ روز اول سال شمسی می‌باشد. علت تشکیل چنین خوشه‌ای می‌تواند تغییرات قابل توجه جمعیت مسکونی در تعطیلات نوروز در زون مورد مطالعه باشد که منجر به کاهش شدید جریان روزانه و شبانه ورودی به زون گردیده است. علاوه بر این با استخراج روزهای مربوط به خوشه شماره سه، مشخص گردید که اعضای این خوشه (۱۶ عضو) بین روزهای ۱۰۰ تا ۱۲۰ سال شمسی می‌باشند که مطابق با ماه مبارک رمضان هستند. تشکیل این خوشه را می‌توان با تغییر در رفتار مصرفی جمعیت و سرانه مصرفی مشترکین به خصوص در ساعات جریان حداقل شبانه (ساعات نزدیک به سحر) به دلیل عادات مذهبی مرتبط دانست. بنابراین می‌توان بیان داشت که تغییرات قابل توجه در رفتار مصرفی جمعیت به علت ماه مبارک رمضان و نیز تغییرپذیری شدید جمعیت در ایام تعطیلات نوروز منجر به تولید خوشه‌های اول و سوم شده است. خوشه شماره دو بزرگترین خوشه تولید شده است که تعداد اعضای آن برای حداقل نقاط همسایگی ۶ و ۱۰ به ترتیب ۲۸۰ و ۲۸۲ بدست آمده و حدود ۷۷ درصد از کل داده‌ها را شامل می‌شود.

شکل ۱۳ همچنین نشان می‌دهد که با افزایش تعداد حداقل نقاط همسایگی از ۶ به ۱۰ و انتخاب شعاع همسایگی بهینه برای هرکدام از آنها تغییر قابل‌توجهی در نتایج بدست آمده از الگوریتم DBSCAN حاصل نمی‌شود. از این رو برای کاهش حجم محاسبات پیشنهاد شده که تعداد حداقل نقاط همسایگی تا جایی که امکان دارد، کوچک انتخاب شود (Sander et al., 1998; Schubert et al., 2017). بنابراین برای اجرای الگوریتم DBSCAN پارامترهای حداقل نقاط همسایگی MinPts=6 و $\varepsilon = 15$ برای ناحیه مورد مطالعه پیشنهاد می‌گردند. بطورکلی می‌توان بیان داشت که انتخاب حداقل نقاط همسایگی به نوع داده، بُعد آن و درصد داده‌های نویز مجموعه داده بستگی دارد. با توجه به نتایج بدست آمده از این مطالعه برای داده‌های جریان، حداقل تعداد نقاط همسایگی بین ۲ تا ۵ برابر بعد داده‌های جریان پیشنهاد می‌شود.

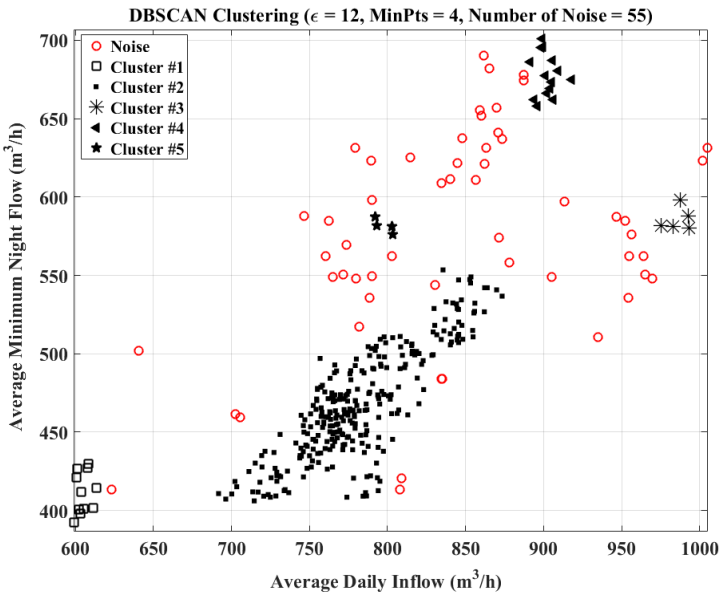


Fig. 12- The graphical presentation of the obtained results in accordance with the DBSCAN algorithm determined parameters for inflow data set in the study area MinPts=4 and ε=12

شکل ۱۲- نمایش گرافیکی نتایج بدست آمده با توجه به پارامترهای تعیین شده برای الگوریتم DBSCAN برای مجموعه داده‌های جریان ورودی به ناحیه مورد مطالعه (MinPts= 4 و ε=12)

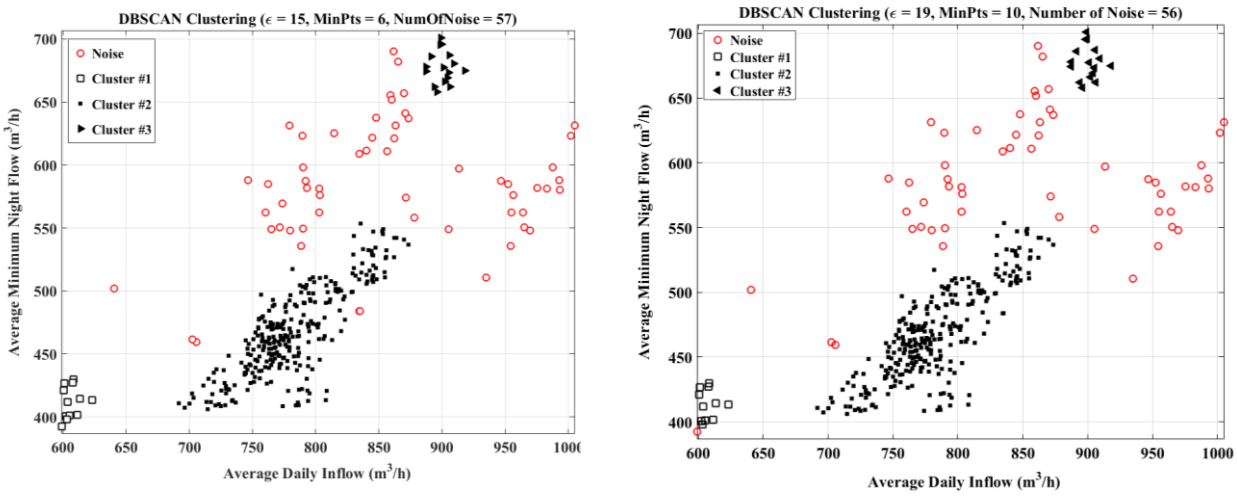


Fig a –شکل الف

Fig b –شکل ب

Fig. 13- The graphical presentation of the obtained results in accordance with the DBSCAN algorithm determined parameters for inflow data set oinf the study area a) MinPts= 6 and ε= 15 , b) MinPts= 10 and ε= 19

شکل ۱۳- نمایش گرافیکی نتایج بدست امده با توجه به پارامترهای تعیین شده برای الگوریتم DBSCAN برای مجموعه داده‌های جریان ورودی به ناحیه مورد مطالعه، الف) MinPts= 6 و ε= 15، ب) MinPts= 10 و ε= 19

علاوه بر این انتخاب شعاع همسایگی بهینه کاملاً وابسته به خصوصیات داده‌های جریان و تراکم آنها بوده و باید با استفاده از منحنی K–امین نزدیکترین همسایگی دامنه آن بدست آید. سپس مقدار بهینه با توجه به شاخص‌های اندازه بزرگترین خوشه، تعداد خوشه‌های تولید شده و تعداد داده‌های نویز کشف شده قابل تعیین است. بطور کلی با افزایش تعداد حداقل نقاط همسایگی به خصوص در مجموعه داده‌ها با تعداد داده‌های نویز بالا نتایج الگوریتم بهبود می‌یابد. اما ملاحظات محاسباتی را نیز باید مد نظر قرار داد؛ از طرفی شعاع همسایگی بهینه کوچکتر اغلب نتاج بهتری ارائه می‌دهد.

شکل ۱۳ نشان می‌دهد که تعداد داده‌های نویز شناسایی شده برای حداقل نقاط همسایگی ۶ و ۱۰ بسیار نزدیک به هم و به ترتیب ۵۷ و ۵۶ نقطه بدست آمده است که حدود ۱۵ درصد از کل داده‌ها را شامل می‌شوند. در مرحله بعد و به منظور تفسیر داده‌های پرت شناسایی شده، این داده‌ها با داده‌های حوادث ثبت شده در سال ۱۳۹۴ مقایسه گردیدند. در سال ۱۳۹۴ در منطقه مورد مطالعه، تعداد ۲۷ شکستگی گزارش شده بود که با توجه به داده‌های اخذ شده از شرکت آب و فاضلاب تمامی آنها توسط متدولوژی پیشنهادی شناسایی گردید. از طرفی این متدولوژی داده‌های نویز (پرت یا غیرعادی) بیشتری را نسبت به شکستگی‌های گزارش شده شناسایی کرده که اختلاف به وجود آمده سه دلیل اصلی دارد؛ ۱- داده‌های مربوط به حوادث توسط شرکت آب و فاضلاب به صورت کامل ثبت و گزارش نشده‌اند؛ ۲- تعدادی از داده‌های نویز می‌تواند مربوط به شکستگی‌های گزارش نشده باشند که تا کشف آن به صورت تصادفی یا انجام عملیات نشت‌یابی قابل شناسایی نمی‌باشند. از این رو معمولا زمان بین وقوع نشت و زمان آگاهی فاصله زمانی زیادی وجود داشته باشد؛ ۳- تعدادی از داده‌های نویز کشف شده می‌تواند منشاء دیگری غیر از شکستگی‌ها داشته و به تغییرات قابل توجه در الگوی مصرفی جمعیت به علت اعیاد و عادات مذهبی مربوط می‌شود. به عنوان نمونه تعداد ۹ داده نویز کشف شده منطبق بر اعیاد رسمی کشور بود؛ ۴- تعدادی از داده‌های نویز می‌تواند به مصارف مجاز غیرمعمول یا استثنایی نظیر مصارف شرکت آب و فاضلاب در شستشوی مخازن یا مصارف غیرمجاز شبانه مرتبط گردند.

در این مطالعه، تأثیر داده‌های نویز در تخمین مقدار نشت شبکه با استفاده از روش تحلیل جریان حداقل شبانه مورد ارزیابی قرار گرفت. به این منظور نشت شبکه با استفاده از روش جریان حداقل شبانه، در حضور داده‌های نویز و پس از حذف آن‌ها تخمین زده شد و درصد خطای محاسبات بدست آمد. نتایج نشان داد که مقدار نشت از شبکه در حضور داده‌های نویز ۴۰۰/۱۱ مترمکعب بر ساعت و پس از حذف داده‌های نویز مقدار ۳۸۳/۴۳ مترمکعب بر ساعت می‌باشد. بنابراین حدود ۱۶/۶۸ مترمکعب بر ساعت یا ۰/۲۱ میلیون مترمکعب بر سال، در محاسبات نشت اختلاف به وجود می‌آید؛ که در این صورت حدود ۴/۳۵ درصد خطا در محاسبات مربوط به نشت برای ناحیه مورد مطالعه

از شبکه توزیع آب شهر تهران وارد می‌شود. ذکر این نکته ضروریست که درصد خطای محاسبه شده تنها برای یک ناحیه مطالعاتی از شبکه توزیع آب شهر تهران بدست آمده است. از این رو خطای تجمعی ناشی از عدم حذف داده‌های نویز در محاسبه نشت شبکه می‌تواند منجر به اختلاف قابل توجهی در تخمین نشت کل شبکه گردد.

۵- نتیجه‌گیری

یک مرحله مهم و ضروری در پایش داده‌های جریان، شناسایی و حذف داده‌های پرت یا غیرعادی، به منظور دستیابی به یک مجموعه داده تاریخی قابل اعتماد و معتبر می‌باشد که به مدیریت و بهره‌برداری بهینه و کارآمد شبکه‌های توزیع آب کمک می‌کند. در این مقاله یک متدولوژی جدید سه مرحله‌ای بر مبنای روش‌های یادگیری بدونِ نظارت با استفاده از الگوریتم خوشه‌بندی مبتنی بر چگالی مقاوم در مقابلِ نویز (DBSCAN) برای شناسایی و حذف داده‌ای پرت از یک مجموعه داده ارائه گردید. پارامتر اصلی الگوریتم DBSCAN یعنی پارامتر شعاع همسایگی با استفاده از مفهوم منحنی K–امین نزدیکترین همسایگی مرتب شده و شاخص‌های اندازه بزرگترین خوشه و درصد داده‌های نویز تعیین گردید. علاوه بر این پارامتر دیگر یعنی حداقل تعداد همسایگی که نشان دهنده کمترین تعداد نقاط برای تشکیل یک خوشه است، با توجه به خصوصیات و بُعد داده‌ها انتخاب گردید. در نهایت با استفاده از پارامترهای تعیین شده برای الگوریتم، تعداد خوشه‌ها و تعداد داده‌های نویز بدست آمد.

متدولوژی پیشنهادی در این مقاله برای یک ناحیه از شبکهٔ توزیع آب تهران به کار برده شد. با توجه به نتایج بدست آمده برای الگوریتم DBSCAN و در نظر گرفتن شاخص‌های کارآمدی آن حداقل نقاط همسایگی MinPts=6 و شعاع $\varepsilon = 15$ به عنوان پارامترهای بهینه الگوریتم DBSCAN برای داده‌های جریان ناحیه مورد مطالعه در سال ۱۳۹۴ بدست آمد. با اعمال این پارامترها و اجرای الگوریتم تعداد ۳ خوشه و ۵۷ داده غیرعادی یا پرت تولید گردید. دلیل اصلی ایجاد این تعداد خوشه و اندازه بزرگترین خوشه، نزدیکی و تراکم نقاط داده‌ای به علت وجود یک الگوی مصرف مشابه در اکثر روزهای سال در ناحیه مورد مطالعه می‌باشد که به خصوصیات جامعه، نوع مصرف غالب در این ناحیه و رژیم بهره‌برداری از شبکه بستگی دارد؛ که از بررسی داده‌ها نیز بدست آمد. شاخص‌های کارآمدی الگوریتم نشان داد که اندازه بزرگترین خوشه (۷۶ درصد از داده‌ها) و درصد داده‌های نویز (۱۵ درصد) در محدوده مجاز پیشنهادی قرار دارد. نتایج بدست آمده بیانگر آن بودند که متدولوژی پیشنهادی در این مطالعه قادر به کشف داده‌های پرت یا غیرعادی ناشی از تغییرات الگوی مصرفی جمعیت، شکستگی‌ها و مصارف مجاز غیرعادی یا مصارف غیرمجاز در سری داده‌های جریان می‌باشد. از این رو این متدولوژی می‌تواند در پایش داده‌های جریان و دستیابی به داده‌های جریان قابل اعتماد و با الگوی جریان مشابه مورد استفاده قرار گیرد، تا از داده‌های بدست آمده در روش‌های آماری ارزیابی نشت یا بهبود دقت روش تحلیل جریان حداقل شبانه استفاده شود.

۶- مراجع

AL-Washali T, Sharma S, AL-Nozaily F, Water MH- and 2019 U (2019) Modelling the leakage rate and reduction using minimum night flow analysis in an intermittent supply system. Water (MDPI) 11(1):1-15

Alkasseh JMA, Adlan MN, Abustan I, Aziz HA and Hanif ABM (2013) Applying minimum night flow to estimate water loss using statistical modeling: A case study in Kinta Valley, Malaysia. Journal of Water Resources Management 27(5):1439-1455

Buchberger SG and Nadimpalli G (2004) Leak estimation in water distribution systems by statistical analysis of flow readings. Journal of Water Resources Planning and Management 130(4):321-329

Cassisi C, Ferro A, Giugno R, Pigola G and Pulvirenti A (2013) Enhancing density-based clustering: Parameter reduction and outlier detection. Journal of Information Systems 38(3):317-330

Chandola V, Banerjee A and Kumar V (2009) Anomaly detection: A survey. Journal of ACM Computing Surveys 41(3):1-58

Daszykowski M, Walczak B, and Massart D (2001) Looking for natural patterns in data: Part 1. Density-based approach. Journal of Chemometrics and Intelligent Laboratory Systems 56(2):83-92

De Oliveira DP, Garrett JH and Soibelman L (2011) A density-based spatial clustering approach for defining local indicators of drinking water distribution pipe breakage. Journal of Advanced Engineering Informatics 25(2):380-389

Ester M, Hans-Peter K, Jorg S, and Xiaowei X (1996) Density-based clustering algorithms for discovering clusters. In: Proc. of The Second International Conference on Knowledge Discovery and Data Mining (KDD-96) 2:226-231

Farley M and Trow S (2005) Losses in water distribution networks: a practitioner's guide to assessment, monitoring and control. London: IWA Publishing, 296p

data. Journal of Water Resources Planning and Management 143(6):1-11

Mutikanga H, Sharma SK, and Vairavamoorthy K (2013) Methods and tools for managing losses in water distribution systems. Journal of Water Resources Planning and Management (April):166-174

Oliveira D, Garrett JH, and Soibelman L (2009) Spatial clustering analysis of water main break events. Journal of Computing in Civil Engineering 338-347

Puust R, Kapelan Z, Savic DA, and Koppel T (2010) A review of methods for leakage management in pipe networks. Urban Water Journal 7(1):25-45

Sander J, Ester M, Kriegel HP, and Xu X (1998) Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. Journal of Data Mining and Knowledge Discovery 2(2):169-194

Schubert E, Sander J, Ester M, Kriegel HP, and Xu X (2017) DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. Journal of ACM Transactions on Database Systems 42(3):1-21

Soni N and Ganatra A (2012a) Comparative study of several Clustering Algorithms. International Journal of Advanced Computer Research 2(4):1-37

Soni N and Ganatra A (2012b) Categorization of several Clustering algorithms from different perspective: A review. International Journal of Advanced Research in Computer Science and Software Engineering 2(8):63-68

Sun J, Wang R, Wang X, Yang H, and Ping J (2014) Spatial cluster analysis of bursting pipes in water supply networks. Journal of Procedia Engineering 70:1610-1618

Thornton J, Sturm R, and Kunkel G (2008) Water loss control. McGraw Hill Professional, 700p

Han J, Kamber M, and Tung AKHH (2001) Spatial clustering methods in data mining: A survey. in: Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, 1-29

Hawkins DM (1980) Identification of outliers. Chapman and Hall, 194p

Jain AK (2010) Data clustering: 50 years beyond K-means. Journal of Pattern recognition letters. Elsevier 31(8):651-666

Jain AK, Murty MN, and Flynn PJ (1999) Data clustering: a review. Journal of ACM computing surveys 31(3):264-323

Li R, Huang H, Xin K, and Tao T (2014) A review of methods for burst/leakage detection and location in water distribution systems. Journal of Water Science and Technology: Water Supply 15(3):429-441

Liu P, Zhou D, and Wu N (2007) VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise. In: Proc. of 2007 International Conference on Service Systems and Service Management. Chengdu, China: IEEE, 1-4

Loureiro D, Amado C, Martins A, Vitorino D, Mamade A, and Coelho ST (2016) Water distribution systems flow monitoring and anomalous event detection: A practical approach. Journal of Urban Water Journal 13(3):242-252

Lv Y, Ma T, Tang M, Cao J, Tian Y, Al-Dhelaan A and Al-Rodhaan M (2016) An efficient and scalable density-based clustering algorithm for datasets with complex structures. Journal of Neurocomputing 171:9-22

Mazzolani G, Berardi L, Laucelli D, Simone A, Martino R, and Giustolisi O (2017) Estimating leakages in water distribution networks based only on inlet flow