نشریه علمی پژوهشی مهندسی آبیاری و آب ایران
سال دهم. شماره سی و هشتم، زمستان ۱۳۹۸


# شناسایی داده‌های غیرنرمال آزمایشات الگوی جریان در قوس با استفاده از روش‌های آماری

محمد واقفی[۱*]، کیومرث محمودی[۲] و مریم اکبری[۳]

تاریخ ارسال:۱۳۹۶/۰۸/۰۸
تاریخ پذیرش: ۱۳۹۷/۱۰/۲۶

مقاله پژوهشی

## چکیده

عوامل مختلفی مانند خطاهای انسانی و دستگاهی، شرایط اندازه‌گیری و طبیعت جریان در شرایط یکتا ممکن است سبب بروز داده‌هایی شود که با الگوی نرمال جامعه آماری در تناقض باشند؛ به گونه‌ای که این گمان به وجود آید که با یک روند متفاوت تولید شده‌اند. در یک تعریف کلی به این نوع از داده‌ها، داده‌های غیرنرمال (پرت یا خارج از محدوده) گفته می‌شود. شناسایی داده‌های پرت از جنبه‌های مختلف دارای اهمیت بوده و منجر به شناخت هرچه بهتر و دقیق‌تر الگوی جریان می‌شود. هدف اصلی از انجام این تحقیق، بررسی و شناسایی داده‌های پرت موجود در آزمایشات الگوی جریان در یک کانال قوسی با زاویه مرکزی ۱۸۰ درجه و عرض ۱ متر با و بدون وجود آبشکن در قوس با استفاده از روش‌های آماری است. کانال مورد نظر در آزمایشگاه هیدرولیک دانشگاه خلیج فارس قرار داشته و برای برداشت سرعت‌های سه بعدی جریان از سرعت‌سنج وکترینو استفاده شده است. به منظور شناسایی داده‌های پرت در این تحقیق از روش‌های انحراف مطلق میانگین، خوشه‌بندی K-Means، ضریب چگالی محلی و روش رای‌گیری استفاده شده است. نتایج حاصل از اجرای این روش‌ها بر داده‌های آزمایشگاهی برداشت شده نشان داد که کارایی بیشتر روش‌ها مناسب است. در این مقاله در نهایت برای حصول بهترین نتیجه، از روش رای‌گیری استفاده شده است. در این روش، داده‌هایی به عنوان کاندیدای داده پرت نهایی در نظر گرفته می‌شوند که توسط بیشتر روش‌ها به عنوان داده پرت شناسایی شده باشند.

**واژه‌های کلیدی**: **روش‌های آماری، داده‌های پرت، الگوی جریان، قوس ۱۸۰ درجه تند، سرعت سنج Vectrino.**

[۱] * دانشیار سازه های هیدرولیکی، گروه مهندسی عمران، دانشگاه خلیج فارس، بوشهر، ایران (نویسنده مسئول)، ۰۷۷-۳۱۲۲۲۴۰۱، vaghefi@pgu.ac.ir
[۲] دانشجوی دکتری، دانشکده مهندسی دریا، دانشگاه امیرکبیر، تهران، ایران،kumarsmahmoodi@aut.ac.ir
[۳] کارشناس ارشد سازه های هیدرولیکی، گروه مهندسی عمران، دانشگاه خلیج فارس، بوشهر، ایران، ۰۷۷-۳۳۵۵۶۴۷۸، m.akbari@pgu.ac.ir

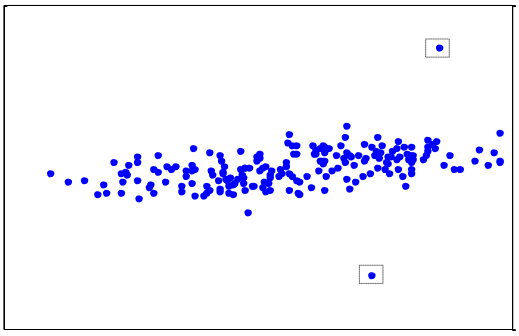
## مقدمه

در بسیاری از مسائل حاکم بر جریان‌های غیریکنواخت و ناپایدار، حرکت رسوب و مواردی که توام با استقرار سازه‌های هیدرولیکی است، مناسب‌ترین راه‌حل و در مواردی تنها راه حل ممکن، استفاده از مدل‌های فیزیکی و آزمایشگاهی است. در مدل‌های آزمایشگاهی آنچه که اهمیت زیادی دارد، اندازه‌گیری دقیق متغیرهای لازم است. به همین دلیل برای اندازه‌گیری دقیق پارامترهای جریان از وسایل الکترونیکی یا لیزری استفاده می‌شود.

رودخانه‌های قوس‌دار را می‌توان به عنوان یکی از مواردی دانست که جریان آب بسیار پیچیده در آن برقرار است. این پیچیدگی نه فقط به خاطر آشفتگی و طبیعت سه بعدی شدید آن بلکه به علت توپوگرافی و تغییرات عمق آن می‌باشد که در حالت عمومی علت آن، پروسه‌های فرسایش، انتقال رسوب و رسوب‌گذاری است. خطوط جریان در چنین میدانی نه تنها خطوط منحنی موازی هم نیستند بلکه این خطوط را می‌توان گفت که در هم تنیده‌اند. داشتن دانش هیدرودینامیکی از چنین جریان‌های انحنا دار، از لحاظ کاربردی بسیار مهم و ضروری می‌باشد که از آن جمله می‌توان به مسائل جلوگیری از پدیده رسوب‌گذاری، تعیین مسیر مناسب کشتی‌رانی، پایدارسازی دینامیکی توپوگرافی رودخانه‌ها، انتخاب محل مناسب آبگیر جانبی و مسئله پخش آلودگی اشاره کرد؛ در واقع می‌توان گفت که هدف اصلی از تعیین الگوی جریان در قوس رودخانه‌ها باید بررسی طرز کار جریان ثانویه و در نتیجه جریان حلزونی و یا مارپیچی باشد. چرا که وجود این جریان باعث متفاوت بودن الگوی جریان در خم با الگوی جریان در مسیرهای مستقیم می‌شود (واقفی و همکاران، ۱۳۸۷). به دلیل وجود جریان ثانویه در قوس رودخانه‌ها، مورفولوژی خم‌های آبرفتی دست‌خوش تغییرات زیادی خواهد شد؛ به گونه‌ای که باعث فرسایش زیاد در قوس خارجی شده و رسوبات شسته شده ناشی از آن را به صورت عرضی حرکت داده و در قوس داخلی ته‌نشین می‌نماید. علاوه بر آن، الگوی جریان حلزونی نیز آشفتگی شدید جریان را به همراه دارد. درحالی‌که یکی از اهداف مهم احداث آبشکن‌ها، حفاظت دیواره‌های خارجی قوس رودخانه و بهبود راستای دیواره و انحراف جریان از دیواره خارجی به میانه رودخانه است. پیچیدگی الگوی جریان در قوس همراه با پیچیدگی الگوی جریان حول آبشکن شرایط الگوی جریان حول آبشکن‌های موجود در قوس را بسیار پیچیده می‌کند (واقفی و همکاران، ۱۳۸۸). به همین دلیل در سال‌های اخیر، استفاده از دستگاه‌های سرعت‌سنج سه بعدی که داده‌های سرعت جریان را به صورت سری‌های زمانی نشان می‌دهند، برای تعیین الگوی جریان پیرامون سازه‌های مستقر در رودخانه‌ها بسیار مورد توجه قرار گرفته است. به عنوان نمونه می‌توان به مطالعات محققین زیر اشاره کرد: Giri et al. (2004) Fazli et al. (2008), Ghodsian and Vaghefi (2009), Duan et al. (2011), Vaghefi et al. (2012), Sukhodolov (2014), Keshavarzi et al. (2014), Tiwari and Sharma (2015), Vaghefi et al. (2017), Akbari and Vaghefi (2017).

در بسیاری از مجموعه داده‌های گردآوری شده از مطالعات آزمایشگاهی، شماری از داده‌ها به صورت متناقض و ناهمسان با الگوی نرمال کلی ایجاد می‌شوند. معمولاً این داده‌ها تحت شرایط خاصی به وجود می‌آیند و تعداد آن‌ها در مجموعه داده اندک است. در یک تعریف کلی به این نوع از داده‌ها، پرت[1] گفته می‌شود. داده پرت داده‌ای است که به طور قابل ملاحظه‌ای از سایر مشاهدات دور افتاده باشد، به گونه‌ای که این سوء ظن ایجاد شود که با یک روند متفاوت تولید شده است (Barnett and Lewis 1994). به عنوان مثال در شکل ۱، داده‌های مشخص شده با یک علامت مربع به دور آن‌ها نماینده داده‌های پرت است.

1. Outlier



**شکل (۱): یک مجموعه داده تصادفی با تعداد دو داده پرت (داده‌های مشخص شده با علامت مربع)**

در یک دسته‌بندی کلی می‌توان روش‌های شناسایی داده‌های پرت را در چهار دسته طبقه‌بندی کرد: ۱) روش‌های آماری؛ ۲) روش‌های مبتنی بر فاصله؛ ۳) روش‌های مستندی[1]و ۴) روش‌های مبتنی بر مدل.

در تکنیک‌های آماری داده‌ها معمولاً با استفاده از یک توزیع احتمالی مدل شده و بر حسب رابطه‌ای که با مدل توزیعی دارند علامت‌گذاری می‌شوند (Barnett et al., 1994, Billor et al., 2000, Eskin, 2000). روش‌های مبتنی بر فاصله با اندازه‌گیری فاصله بین نقاط داده‌های پرت را شناسایی می‌کنند (Aggarwal and Yu, 2001, Breunig, 2000, Knorr and Ng, 1998). الگوریتم‌های مبتنی بر فاصله که تا کنون ارائه شده‌اند یا با اندازه‌گیری فاصله بین نقاط، و یا با تخمین چگالی همسایه‌های محلی نقاط داده‌های پرت را شناسایی می‌کنند (Aggarwal and Yu, 2001, Papadimitriou et al., 2003). علاوه بر این از تکنیک‌های مبتنی بر خوشه بندی[2]نیز در این خصوص استفاده شده است. در این روش‌ها اگر داده‌ای به هیچ یک از خوشه‌ها تعلق نداشته باشد (Yu et al., 2002) و یا اندازه یک خوشه به طور قابل توجهی کوچک‌تر از سایر خوشه‌ها باشد، می‌تواند کاندیدای داده پرت باشد (Skin et al., 2002). در روش‌های مستندی، پروفیلی از رفتار نرمال داده‌ها با استفاده از تکنیک‌های مختلف داده‌کاوی ایجاد می‌شود، که در این صورت انحراف از آنها به عنوان رفتار غیر نرمال قلمداد می‌شود. در نهایت روش‌های مبتنی بر مدل، معمولاً ابتدا رفتار نرمال نمونه را با استفاده از برخی مدل‌های پیش‌بینی کننده (به عنوان مثال شبکه‌های عصبی تکرار شونده[3](Hawkins et al., 2002) و یا ماشین-های بردار پشتیبان بدون نظارت[4](Skin et al., 2002; Lazarevic et al., 2003)مشخص می‌کنند، آنگاه داده‌هایی که انحراف آن‌ها از مدل آموزش دیده زیاد باشد را به عنوان داده پرت در نظر می‌گیرند. Vaghefi et al. (2018a) به شناسایی داده های پرت در آزمایشات الگوی جریان در قوس ۹۰ درجه ملایم با استفاده از روش‌های: box plot, histograms, linear regression, k-nearest neighbors, local outlier factor, k-medoids clustering, multilayer perceptron, self-organizing map پرداختند. آن‌ها از فرکانس ۵۰ هرتز برای برداشت داده‌های ADV استفاده کردند. علاوه بر این Vaghefi et al. (2018b) به بررسی کارایی تعدادی از روش‌های داده‌کاوی در آزمایشات الگوی جریان در کانال قوسی ۱۸۰ درجه تند پرداختند. آن‌ها با استفاده از فرکانس ۲۵ هرتز، کارایی روش‌های: Z-score test, sum of sine curve fitting, Mahalanobis distance, hierarchical clustering, LSC-Mine, Self-organizing map, Fuzzy C-Means Clustering, voting را برای شناسایی داده های پرت با هم مقایسه کردند.

با وجود اهمیت زیاد صحت داده‌های آزمایشگاهی مربوط به الگوی جریان، تاکنون بررسی و شناسایی داده‌های پرت موجود در آزمایشات الگوی جریان در قوس‌های تند که از آشفتگی زیادی برخوردار است، کمتر مورد توجه قرار گرفته است. به همین دلیل در این تحقیق با استفاده از روش‌هایی جدید، به شناسایی داده‌های پرتی که از سرعت سنج Vectrino در دو آزمایش الگوی جریان با و بدون وجود آبشکن T شکل کوتاه در قوس ۱۸۰ درجه تند برداشت شده، پرداخته شده است. روش‌های داده‌کاوی مورد بحث در این مقاله عبارتند از: MAD test, K-Means clustering, Local Density Factor, Voting method. تمامی این روش‌ها جز دسته روش‌های آماری و مبتنی بر فاصله هستند.

1. Profiling methods
2. Clustering base methods
3. Replicator neural networks
4. Unsupervised support vector methods

## مواد و روش‌ها

### معرفی روش‌های شناسایی داده پرت

**- آزمون MAD**

با استفاده از آزمون[5] MAD می‌توان داده‌های پرت را در مجموعه داده‌های تک متغیره شناسایی کرد. ضریب MAD با استفاده از رابطه زیر قابل محاسبه است:

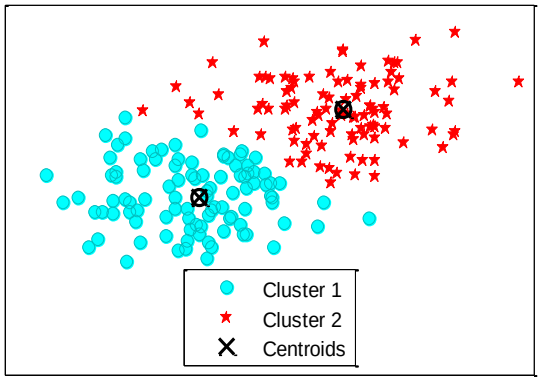
$$MAD = Median\{|x_i - \tilde{x}|\} \quad (۱)$$

در رابطه فوق $x_i$ عضو $i$ام مجموعه داده $(i=1,2,3,...,n)$ و $\tilde{x}$ میانه مجموعه داده است. معمولاً داده‌هایی که خارج از بازه $\tilde{x} \pm 3MAD$ قرار می‌گیرند، می‌توانند کاندیدای داده پرت باشند (Barnett and Lewis, 1994). البته مقدار ۳ وابسته به ماهیت داده‌های ورودی است و برای عملکرد بهتر روش در شناسایی داده‌های پرت می‌تواند کمتر و یا بیشتر انتخاب شود.

**- روش خوشه‌بندی K-Means**

هدف از خوشه‌بندی یک مجموعه داده این است که داده‌های موجود به چند گروه تقسیم شوند، به طوری که در این تقسیم‌بندی داده‌های گروه‌های مختلف حداکثر تفاوت ممکن را به هم داشته باشند و داده‌های موجود در یک گروه بسیار به هم شبیه باشند. به عنوان مثل شکل ۲ بیانگر یک مجموعه داده است که در آن داده‌ها در دو خوشه تقسیم‌بندی شده‌اند. مرکزیت هر خوشه با یک علامت × در شکل مشخص شده است.

یکی از مهمترین الگوریتم‌هایی که از آن برای خوشه‌بندی داده استفاده می‌شود الگوریتم K-Means است (Ng and Han, 1994, Ester et al., 1996). این روش با وجود سادگی، یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر (مانند خوشه‌بندی فازی) محسوب می‌شود. دلایل محبوبیت K-Means پیاده‌سازی آسان، سادگی، کارایی و موفقیت تجربی آن است.




**شکل (۲): یک مجموعه داده دلخواه با تعداد ۲ خوشه**

فرض کنید $X = \{x_i\} i = 1,2,...n$ یک مجموعه داده متشکل از $n$ نمونه با بعد $dim$ باشد که قرار است به مجموعه‌ای $K$ خوشه‌ای، خوشه‌بندی شود $C = \{c_k, k = 1,...,k\}$. الگوریتم K-Means یک بخش خاص را پیدا کرده، به‌گونه‌ای که مربع خطای بین میانه تجربی یک خوشه و نقاط واقع شده در آن خوشه کمینه شود. اگر $\mu_k$ میانه خوشه $c_k$ باشد، مربع خطای بین $\mu_k$ و نقاط $x_i$ در خوشه $c_k$ به صورت زیر تعریف می‌شود:

$$J(c_k) = \sum \|x_i - \mu_k\|^2 \quad (۲)$$

در این روش ابتدا به تعداد خوشه‌های مورد نیاز نقاطی به صورت تصادفی انتخاب می‌شود. سپس داده‌ها، با توجه به میزان نزدیکی (شباهت) به یکی از خوشه‌ها نسبت داده می‌شوند و بدین ترتیب خوشه‌های جدیدی حاصل می‌شود. با تکرار همین روال می‌توان در هر تکرار با میانگین‌گیری از داده‌ها مراکز جدیدی برای آنها محاسبه کرد و مجدداً داده‌ها را به خوشه‌های جدید نسبت داد. این روند تا زمانی ادامه پیدا می‌کند که دیگر تغییری در داده‌ها حاصل نشود. تابع هدف در K-Means، به حداقل رساندن مجموع مربعات خطا است:

5. Median of absolute deviations

$$J(c)=\sum_{k=1}^{k}\sum_{x_{i\in c_k}}\left\|x_i-\mu_k\right\|^2 \qquad (۳)$$

مراحل اصلی الگوریتم K-Means به شرح زیر است (Anil, 2009):

۱- در ابتدا $K$ نقطه به عنوان نقاط مراکز خوشه‌ها انتخاب می‌شوند.

۲- هر نمونه به خوشه‌ای که مرکز آن خوشه کمترین فاصله تا آن داده را دارا است، نسبت داده می‌شود.

۳- پس از تعلق تمام داده‌ها به یکی از خوشه‌ها، برای هر خوشه یک نقطه جدید به عنوان مرکز محاسبه می‌شود (میانگین نقاط متعلق به هر خوشه).

۴- مراحل ۲ و ۳ تا زمانی که دیگر هیچ تغییری در مراکز خوشه‌ها حاصل نشود، تکرار می‌شوند.

اجرای الگوریتم K-Means روی یک مجموعه داده نیازمند تعریف مقدار سه پارامتر توسط کاربر است: تعداد خوشه‌ها (K)، مقداردهی اولیه خوشه، و تابع اندازه‌گیری فاصله. مهمترین پارامتر، انتخاب صحیح مقدار K است. قانون مشخصی برای انتخاب صحیح مقدار این پارامتر وجود ندارد و وابسته به داده-های ورودی است. معمولاً الگوریتم به ازای مقادیر مختلف K اجرا شده، و مناسب‌ترین مقدار انتخاب می‌شود. مقداردهی اولیه متفاوت می‌تواند منجر به خوشه‌بندی نهایی مختلف شود. در این تحقیق برای اندازه‌گیری فاصله از مربع تابع اقلیدسی استفاده شده است. در روش K-Means، داده‌هایی به عنوان پرت انتخاب می‌شوند که به هیچ خوشه‌ای تعلق نداشته باشند و یا اینکه اگر تعداد داده‌های واقع شده در یک خوشه، به طور قابل توجهی کمتر از سایر خوشه‌ها باشد، تمامی داده‌های واقع شده در آن خوشه به عنوان کاندیدای پرت انتخاب می‌شوند.

**- روش ضریب چگالی محلی**

یکی از روش‌های شناسایی داده‌های پرت، استفاده از توابع چگالی کرنل است. روش‌های مبتنی بر چگالی به طور معمول با اندازه‌گیری چگالی محلی داده‌ها، داده‌های پرت موجود در نمونه را شناسایی می‌کنند. هرچه چگالی محلی یک داده کمتر باشد، میزان تناقض آن با سایر اعضای نمونه نیز بیشتر می‌شود. چگالی محلی یک داده بیانگر میزان تراکم داده‌های واقع شده در همسایگی آن است. در این تحقیق از روش ضریب چگالی محلی[1](LDF) که یکی از روش‌های مبتنی بر چگالی است، برای شناسایی داده‌های پرت در مجموعه داده‌های گردآوری شده از سرعت جریان، استفاده شده است (Latecki, 2007).

اساس عملکرد روش ضریب چگالی محلی در شناسایی داده‌های پرت، مقایسه چگالی محلی هر داده با میانگین چگالی محلی همسایه‌هایش است. در این روش، تابع تخمین چگالی محلی[2](LDE) از رابطه زیر محاسبه می‌شود:

$$LDE(x_j)\alpha\frac{1}{m}\sum_{x_i\in mNN(x_j)}\frac{1}{(2\pi)^{\frac{\dim}{2}h(x_i)^{\dim}}}\exp\left(-\frac{rd_k(x_j,x_i)^2}{2h(x_i)^2}\right)$$
$$=\frac{1}{m}\sum_{x_i\in mNN(x_j)}\frac{1}{(2\pi)^{\frac{\dim}{2}[h,d_k(x_i)]^{\dim}}}\exp\left(-\frac{rd_k(x_j,x_i)^2}{2(h,d_k(x_i))^2}\right) \qquad (۴)$$

در رابطه (۴)، $mNN(x_j)$ بیانگر تعداد $m$ تا نزدیک‌ترین همسایه نمونه $x_j$ است. $m$ یک عدد صحیح مثبت است که توسط کاربر تعیین می‌شود. $x_i$ و $x_j$ نمونه‌هایی از مجموعه داده هستند، $dim$ یک عدد صحیح مثبت است که تعداد مشخصه‌های (متغیرهای) مجموعه داده را تعیین می‌کند. $h$ یک عدد حقیقی است که پهنای باند[3]نامیده شده و با توجه به نوع مجموعه داده ورودی، توسط کاربر انتخاب می‌شود. $d_k(x_i)$ بیانگر فاصله $k$امین نزدیک‌ترین همسایه داده $x_i$ از آن است. این فاصله می‌تواند

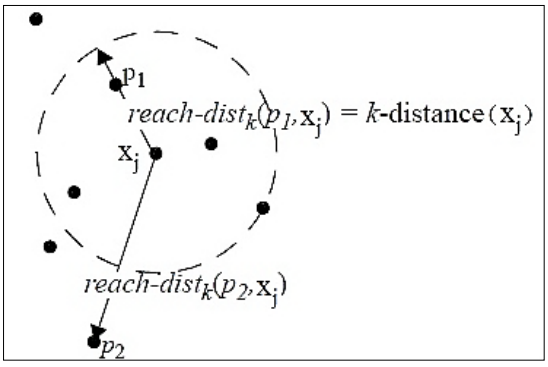
1. Local Density Factor (LDF)
2. Local Density Estimate (LDE)
3. Bandwidth

توسط یک تابع اندازه‌گیری فاصله مثل تابع اقلیدسی محاسبه شود. فاصله اقلیدسی از رابطه زیر قابل محاسبه است:

$$d(x_j, x_i) = \|x_j, x_i\|^2 \quad (۵)$$

$d(x_j, x_i)$ بیانگر فاصله دسترسی[4] نمونه $x_i$ نسبت $x_i$ به $x_j$ است. علت استفاده از این فاصله در رابطه بالا، افزایش قدرت تابع تخمین چگالی محلی است. این فاصله برگرفته از روش ضریب داده پرت محلی[5] (LOF) (محمودی و همکاران، ۱۳۹۲) است.

شکل ۳ بیانگر مفهوم فاصله دسترسی به ازای $k = 4$ است. اگر داده $p$ در فاصله زیادی نسبت به $x_j$ واقع شده باشد (یعنی $p_2$ در شکل)، آنگاه فاصله دسترسی آنها برابر فاصله خالص بین آن دو است. حال اگر این دو داده در فاصله کمی نسبت به یکدیگر قرار گرفته باشند (یعنی $p_1$ در شکل)، آنگاه فاصله دسترسی برابر فاصله $p_1$ امین همسایه $x_i$ از آن خواهد بود.




**شکل (۳): مفهوم فاصله دسترسی به ازای k=4**

**- روش رای‌گیری**

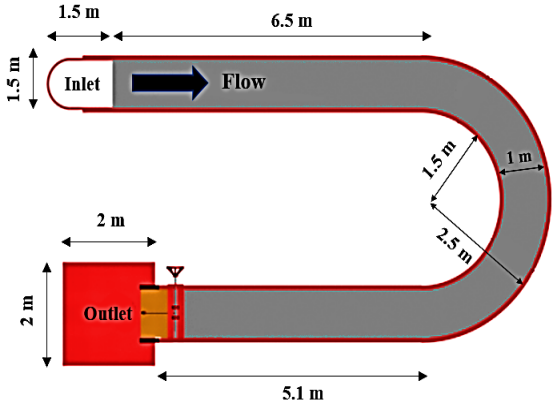
روش رای‌گیری یک روش جدید نیست، بلکه مبنای عملکرد آن در شناسایی داده‌های پرت استفاده از نتایج روش-های بیان شده است. در واقع در این روش داده‌هایی به عنوان کاندیدای پرت در نظر گرفته می‌شوند که توسط تمامی روش‌ها و یا اکثریت آنها به عنوان پرت تشخیص داده شده باشند. معمولاً داده‌هایی که توسط بیشتر روش‌ها به عنوان پرت تشخیص داده شوند، احتمال پرت بودن آنها بیشتر است. بنابراین می‌توان گفت استفاده از شیوه رای‌گیری منجر به نتایج دقیق‌تر و قابل اعتمادتر می‌شود.

**معرفی تجهیزات آزمایشگاهی**

**- کانال آزمایشگاهی**

آزمایشات تعیین الگوی جریان پیرامون آبشکن T شکل در کانال قوسی با زاویه مرکزی ۱۸۰ درجه در آزمایشگاه هیدرولیک دانشگاه خلیج فارس بوشهر انجام گرفته است. مطابق شکل ۴، این کانال از یک مسیر مستقیم به طول ۶/۵ متر در بالادست و همچنین مسیر مستقیم دیگری به طول ۵/۱ متر در پایین دست تشکیل شده است. ارتفاع این کانال ۷۰ سانتی‌متر و عرض آن نیز ۱ متر بوده و نسبت شعاع انحنا به عرض کانال برابر با ۲ است (R/B=2). کف کانال صلب بوده و به منظور تامین زبری کف، رسوباتی با قطر متوسط ۱ میلی‌متر به کف کانال چسبانده شده است. در کلیه آزمایشات، دبی و عمق آب ثابت می‌باشد. دبی با استفاده از دستگاه دبی سنج آلتراسونیک، ۹۵ لیتر بر ثانیه تنظیم شده است. همچنین عمق آب نیز بوسیله دریچه پروانه‌ای در انتهای مسیر مستقیم پایین دست به گونه‌ای تنظیم شده که در مسیر مستقیم بالادست عمق آب، ۲۰ سانتی‌متر باشد. بنابراین با توجه به عدد فرود ۰/۳۴ و عدد رینولدز ۶۷۸۵۷، در کلیه آزمایشات رژیم جریان به صورت کاملا آشفته برقرار است (Nortek, 2009;Vaghefi et al., 2016).

4. Reachability distance
5. Local Outlier Factor (LOF)



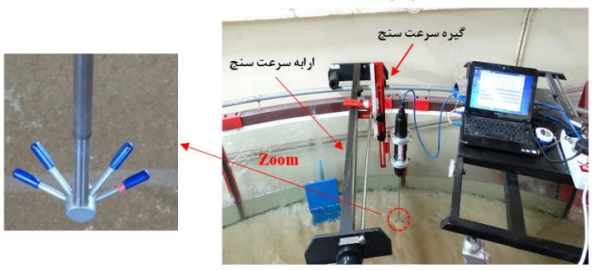

**شکل (۴): نمایش پلان کانال آزمایشگاهی و ابعاد قسمت‌های مختلف آن**

**- آبشکن**

آبشکن مورد استفاده از جنس پلکسی گلاس و دارای ارتفاع ۴۰ سانتی‌متر و ضخامت ۱ سانتی‌متر است. ابعاد هندسی آبشکن در پلان به صورت T شکل می‌باشد. این آبشکن به صورت غیر مستغرق، دماغه نیم دایره‌ای و منفرد در موقعیت ۹۰ درجه و به صورت عمود بر دیواره قوس خارجی کانال نصب شده است. ابعاد هندسی طول بال و جان آبشکن T شکل، برابر ۱۵ درصد عرض کانال است که با توجه به عمق آب و طول آبشکن جزء آبشکن‌های کوتاه است.

**- سرعت‌سنج Vectrino**

برای اندازه‌گیری مولفه‌های سه بعدی سرعت جریان در قوس با و بدون آبشکن T شکل از سرعت‌سنج Vectrino که یکی از پیشرفته‌ترین انواع سرعت‌سنج‌های ADV[1]می‌باشد، استفاده شده است. درشکل ۵، حسگر[2]این دستگاه در حالت جانب‌نگر[3]و نحوه برداشت داده پیرامون آبشکن و استقرار آن بر قوس نمایش داده شده است. برای برداشت داده‌ها، سرعت‌سنج به گونه‌ای تنظیم شد که به طور عمومی در هر نقطه، برداشت داده در مدت زمان ۶۰ ثانیه با فرکانس ۲۵ هرتز انجام گیرد. دوپارامتر سیگنال به نویز[4]و همبستگی[5]تعیین کننده کیفیت نسبی داده‌های ثبت شده بوده که مقادیر حداقل آن در انجام آزمایشات به ترتیب برابر با ۲۰ و ۷۰ درصد است (با استفاده از نرم افزارهای Vectrino و Explorer V) (Vaghefi et al., 2016).




**شکل(۵): استقرار سرعت سنج Vectrino در قوس و نمایش حسگر جانب‌نگر**

**- مش‌بندی کانال برای تعیین موقعیت نقاط برداشت شده**

در آزمایشات انجام شده به علت حساسیت و آشفتگی بیشتر جریان در محدوده استقرار آبشکن نسبت به حالت بدون وجود آبشکن، از مش‌بندی ریزتر در این محدوده و پایین‌دست آن استفاده شد. در شکل ۶ به عنوان نمونه، نحوه مش‌بندی در نظر گرفته شده برای آزمایش قوس توام با استقرار آبشکن T شکل در طول قوس نشان داده شده است.

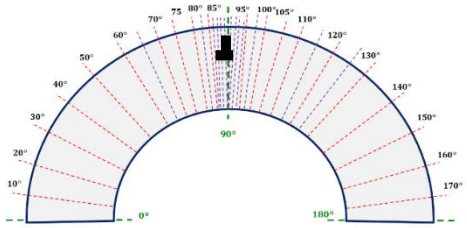
**- مشخصات نقاط مورد مطالعه**

در این تحقیق از میان نقاط برداشت شده، توانایی روش‌های شناسایی داده‌های پرت به صورت مطالعه موردی برای مختصات ۲ نقطه (مقادیر سرعت‌ها در جهت‌های U، V و W) مورد تجزیه و تحلیل قرار گرفته است. از این نقاط یک نقطه با وجود آبشکن و نقطه دیگر نیز بدون وجود آبشکن در

1. Acoustic Doppler Velocimeter
2. Probe
1. Side looking

2. Signal to Noise Ratio(SNR)
3. Coloration

قوس است. جزئیات مجموعه داده‌های مورد بررسی در جدول ۱، ارائه شده است. در ستون آخر جدول (سمت چپ) Z بیانگر ارتفاع از کف کنال، Θ زاویه با افق و d فاصله از دیواره خارجی قوس است.



پارامترهای هر الگوریتم، فایل داده‌های فیلتر شده و فایل داده‌های کاندیدای پرت را در قالب Excel یا Notepad به صورت خودکار ذخیره کرده و در اختیار کاربر قرار می‌دهد. این برنامه دارای امکانات متنوعی بوده و به صورت رابط گرافیکی کاربرطراحی شده است. در شکل ۷ نمایی از محیط برنامه نشان داده شده است. ذکر این نکته ضروری است که صحت عملکرد روش‌های مورد استفاده در این تحقیق، وابسته به

**شکل (۶): نمونه‌ای از مش‌بندی کانال برای تعیین موقعیت نقاط برداشت شده با استفاده از Vectrino در طول کانال برای آزمایش با آبشکن T شکل**

## نتایج و بحث

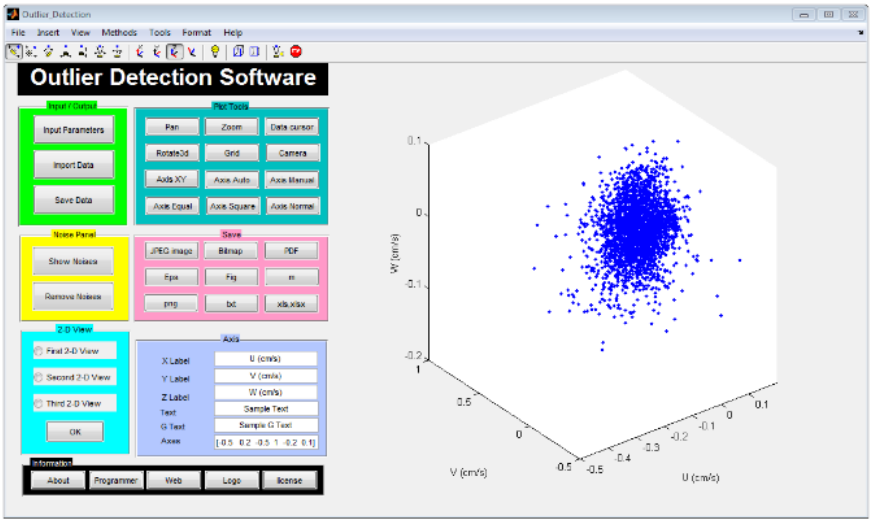
برای شناسایی داده‌های پرت بر اساس الگوریتم هر روش، یک برنامه کامپیوتری در نرم‌افزار MATLAB نوشته شده است. این برنامه، فایل داده‌های خام را در قالب فایل Excel یا Notepad از ورودی دریافت کرده، سپس پس از تنظیم

تعیین صحیح پارامترهای ورودی آنها توسط کاربر است. در واقع در روش‌های مورد بحث، مقادیر مختلف برای پارامترها مورد ارزیابی قرار گرفته و بهترین آنها که منجر به نتایج دقیق‌تر و به واقعیت نزدیک‌تر توسط الگوریتم شوند، به عنوان مقادیر نهایی انتخاب شده‌اند. به عنوان مثال یکی از پارامترهای بسیار مهم در صحت عملکرد روش‌ها، انتخاب صحیح مقدار پارامتر آستانه است

**جدول(۱): مشخصات نقاط مورد مطالعه در آزمایشات الگوی جریان در قوس**

| شماره | وضعیت | مولفه سرعت (cm/s) | مشخصات | تعداد | کمترین | بیشترین | میانگین | انحراف معیار |
|---|---|---|---|---|---|---|---|---|
| ۱ | قوس بدون آبشکن | U | Z=cm0.5 d=cm8 Θ=100° | ۱۵۳۳ | -۵۶/۲۶۰۰ | -۳۶/۸۲۰۰ | -۴۶/۸۲۴۴ | ۲/۵۲۷۰ |
| | | V | | | -۸/۶۱۰۰ | ۹/۸۰۰۰ | ۱/۱۴۲۷ | ۲/۶۷۵۳ |
| | | W | | | -۱۵/۸۹۰۰ | ۷/۸۱۰۰ | -۴/۱۱۴۴ | ۲/۸۲۷۹ |
| ۲ | قوس با وجود آبشکن | U | Z=cm17 d=cm20 Θ=130° | ۱۵۳۷ | -۸۴/۲۱۰۰ | ۷۴/۹۳۰۰ | ۳۲/۵۳۳۰ | ۱۷/۴۴۷۳ |
| | | V | | | -۲۵/۷۲۰۰ | ۱۸/۷۷۰۰ | ۱/۰۶۰۰ | ۸/۲۳۵۹ |
| | | W | | | -۱۵۵/۳۲۰۰ | ۱۶۴/۷۶۰۰ | ۲/۰۸۲۱ | ۲۰/۱۰۶۷ |

4. Graphical User Interface (GUI)




**شکل (۷): نمایی از محیط نرم‌افزار تهیه شده**

به طور کلی هرچه مقدار این پارامتر کمتر انتخاب شود، داده‌های بیشتری از حوزه نرمال خارج شده و توسط روش‌ها به عنوان پرت تشخیص داده می‌شوند. عکس این حالت نیز برقرار است؛ هرچه مقدار پارامتر آستانه بزرگتر انتخاب شود، شمار بیشتری از داده‌های پرت به عنوان نرمال انتخاب می‌شوند. در ادامه نتایج اجرای روش‌های مختلف بر روی مجموعه داده‌ها برای شناسایی داده‌های پرت ارائه شده است.

**- نتایج اجرای روش MAD**

اجرای روش آزمون MAD روی مجموعه داده‌ها نیازمند تعیین مقداری است که در ضریب MAD ضرب می‌شود. در این تحقیق بنابر آزمایش‌های صورت گرفته و آزمون و خطا، این مقدار برابر ۵ در نظر گرفته شده است. در جدول ۲ نتایج اجرای این روش روی مجموعه داده‌ها ارائه شده است. همانطور که در این جدول مشاهده می‌شود، اعمال این روش سبب شناسایی داده‌های پرت بیشتری در نقطه ۲ نسبت به نقطه ۱ شده است. بیشترین میزان داده‌های پرت شناسایی شده مربوط به مولفه عمقی سرعت جریان بوده که تاثیر زیادی بر محاسبات مربوط به قدرت جریان ثانویه دارد. نکته حایز اهمیت در این جدول این است که کمترین تعداد داده‌های پرت مربوط به مولفه‌های عرضی سرعت جریان در هر دو آزمایش می‌باشد.

**- نتایج اجرای روش ضریب چگالی محلی**

اعمال الگوریتم ضریب چگالی محلی برای شناسایی داده‌های پرت، نیازمند تعیین مقادیر پارامترهای آن ( $K$، $m$، $c$ و $h$ ) است. این مقادیر معمولاً با توجه به نوع مسئله انتخاب می‌شوند. با توجه به ماهیت داده‌ها و آزمون و خطا، در تمامی آزمون‌های انجام شده در این تحقیق $c = 0.1$، $h = 1$، $m =$ $k = 50$ و $100$ انتخاب شده است. در اینجا نیز مقادیر بالای $m$ و $k$ سبب افزایش میزان دقت روش شده، ولی از طرفی حجم محاسبات را افزایش داده و در نتیجه زمان اجرای روش نیز افزایش می‌یابد. با افزایش مقادیر $m$ و $k$ تا یک حد مشخص، دیگر تغییر قابل ملاحظه‌ای در نتایج حاصل نمی‌شود، در نتیجه نباید مقادیر این پارامترها را خیلی بالا در نظر گرفت. مقدار پارامتر آستانه نیز با توجه به ماهیت داده‌های ورودی و آزمون و خطا برابر ۶/۲ انتخاب شده است. از اینرو اگر ضریب چگالی محلی محاسبه شده برای یک داده از ۶/۲ بیشتر باشد، به عنوان کاندیدای داده پرت انتخاب می‌شود. در جدول ۳ نتایج اجرای این روش روی مجموعه داده‌ها ارائه شده است. از

تابع اقلیدسی برای اندازه‌گیری فاصله بین نقاط استفاده شده است.

**جدول (۲): داده‌های پرت شناسایی شده در مجموعه داده‌ها با استفاده از روش آزمون MAD**

| شماره | مولفه سرعت | شماره داده پرت | | | | | | | | | تعداد |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ۱ | U | ۲۵۰ | ۲۶۷ | ۵۶۹ | ۵۷۸ | ۵۷۹ | ۷۵۲ | ۸۲۰ | ۱۰۹۹ | ۱۱۰۰ | ۱۳ |
| | | | | | ۱۳۵۷ | ۱۴۸۴ | ۱۴۹۱ | ۱۵۰۳ | | | |
| | V | | | | | ۷۴۶ | | | | | ۱ |
| | W | | ۲۰۶ | ۲۶۹ | ۷۱۴ | ۸۹۴ | ۱۰۷۶ | ۱۱۷۱ | | | ۶ |
| ۲ | U | ۳۲ | ۸۶ | ۸۷ | ۱۰۸ | ۱۰۹ | ۱۱۶ | ۱۳۳ | ۱۴۲ | ۱۵۷ | ۴۲ |
| | | | ۱۷۷ | ۱۸۷ | ۴۱۲ | ۵۰۷ | ۵۱۹ | ۶۳۱ | ۶۳۲ | ۶۶۴ | |
| | | | ۷۷۱ | ۸۳۱ | ۸۵۸ | ۸۶۷ | ۸۸۹ | ۹۳۶ | ۹۳۷ | ۹۴۲ | |
| | | | ۹۹۹ | ۱۰۰۰ | ۱۱۹۵ | ۱۲۱۳ | ۱۲۲۱ | ۱۲۲۳ | ۱۲۲۴ | ۱۲۴۲ | |
| | | | ۱۲۷۸ | ۱۲۸۶ | ۱۳۷۱ | ۱۳۷۲ | ۱۴۱۵ | ۱۴۱۶ | ۱۴۲۰ | ۱۴۳۵ | |
| | | | | | | ۱۵۱۱ | | | | | |
| | V | | | | | --- | | | | | ۰ |
| | W | ۳۰ | ۳۲ | ۴۴ | ۸۶ | ۸۷ | ۱۰۸ | ۱۰۹ | ۱۱۶ | ۱۳۳ | ۴۵ |
| | | | ۱۸۷ | ۲۱۹ | ۲۲۰ | ۲۹۵ | ۳۶۵ | ۴۱۲ | ۵۱۹ | ۵۵۳ | |
| | | | ۵۷۳ | ۵۷۷ | ۶۳۱ | ۶۳۲ | ۶۶۴ | ۷۷۱ | ۸۰۴ | ۸۳۱ | |
| | | | ۸۶۷ | ۹۳۷ | ۹۴۲ | ۱۰۰۰ | ۱۰۰۲ | ۱۰۱۴ | ۱۱۹۵ | ۱۲۲۱ | |
| | | | ۱۲۴۲ | ۱۲۸۶ | ۱۲۸۷ | ۱۳۳۹ | ۱۳۷۱ | ۱۳۷۵ | ۱۴۰۳ | ۱۴۱۶ | |
| | | | | | ۱۴۲۰ | ۱۴۳۵ | ۱۴۷۳ | ۱۵۱۱ | | | |

مقایسه جدول‌های ۲ و ۳ نشان دهنده روند مشابه نتایج حاصل از روش‌های ضریب چگالی محلی و آزمون MAD می‌باشد. اعمال روش ضریب چگالی محلی علاوه بر اینکه کمترین تعداد داده‌های پرت را در راستای عرضی نتیجه می‌دهد، همچنین سبب می‌شود که بیشترین تعداد داده‌های پرت برای هر دو آزمایش در راستای طولی باشد. علاوه بر این، مجددا مجموع داده‌های پرت شناسایی شده در آزمایشی که توام با آبشکن است، به دلیل افزایش آشفتگی جریان و بی‌نظمی‌های حاصل از آن، بیشتر گزارش شده است.

**- نتایج اجرای روش K-Means**

اجرای الگوریتم K-Means روی مجموعه داده‌ها نیازمند تعریف مقدار سه پارامتر توسط کاربر است: تعداد خوشه‌ها ($K$)، مقداردهی اولیه خوشه، و تابع اندازه‌گیری فاصله. در این تحقیق برای تمامی مجموعه داده‌ها تعداد خوشه‌ها برابر ۷۰ در نظر گرفته شده است.

**جدول (۳): داده‌های پرت شناسایی شده در مجموعه داده‌ها با استفاده از روش ضریب چگالی محلی**

| شماره | مولفه سرعت | شماره داده پرت | تعداد |
|---|---|---|---|
| ۱ | U | ۲۴۸ ۲۵۰ ۲۶۷ ۲۶۹ ۵۶۹ ۵۷۸ ۵۷۹ ۷۵۲ ۸۲۰<br>۱۰۹۹ ۱۱۰۰ ۱۳۵۲ ۱۳۵۳ ۱۳۵۷ ۱۴۳۸ ۱۴۳۹ | ۱۹ |
| | V | ۵۶ ۳۶۹ ۳۷۰ ۷۱۷ ۷۴۵ ۷۴۶ ۱۱۵۶ ۱۱۵۷ | ۸ |
| | W | ۲۶۹ ۱۱۷۱ | ۲ |
| ۲ | U | ۱۵ ۳۲ ۸۶ ۸۷ ۱۰۹ ۱۱۶ ۱۳۳ ۱۴۲ ۱۵۷<br>۱۸۷ ۴۱۲ ۵۱۹ ۵۷۸ ۶۳۱ ۶۳۲ ۶۶۴ ۷۷۱<br>۸۳۱ ۸۵۸ ۸۶۷ ۸۸۹ ۹۳۶ ۹۳۷ ۹۴۲ ۹۹۹<br>۱۰۰۰ ۱۱۹۵ ۱۲۱۳ ۱۲۲۱ ۱۲۲۳ ۱۲۲۴ ۱۲۴۲ ۱۲۸۶<br>۱۳۷۱ ۱۳۷۲ ۱۴۱۵ ۱۴۱۶ ۱۴۲۰ ۱۴۹۵ ۱۵۱۱ | ۴۰ |
| | V | --- | ۰ |
| | W | ۳۰ ۳۲ ۸۶ ۸۷ ۱۰۸ ۱۰۹ ۱۱۶ ۱۳۳ ۱۸۷<br>۲۱۹ ۲۲۰ ۲۹۵ ۳۶۵ ۴۱۲ ۵۱۹ ۵۵۳ ۵۷۳<br>۶۳۱ ۶۳۲ ۶۶۴ ۷۷۱ ۸۳۱ ۸۶۷ ۹۳۷ ۹۴۲<br>۱۰۰۰ ۱۰۰۲ ۱۱۹۵ ۱۲۲۱ ۱۲۴۲ ۱۲۸۶ ۱۲۸۷ ۱۳۷۱<br>۱۳۷۵ ۱۴۰۳ ۱۴۱۶ ۱۴۲۰ ۱۴۳۵ ۱۵۱۱ | ۳۹ |

اگر تعداد داده‌های واقع شده در یک خوشه از یک پارامتر آستانه $t$ کمتر باشد، آنگاه تمامی داده‌های واقع شده در آن خوشه به عنوان کاندیدای داده پرت در نظر گرفته می‌شوند. مقدار پارامتر آستانه در تمامی مجموعه داده‌ها برابر ۵ انتخاب شده است، یعنی اگر تعداد داده‌های واقع شده در یک خوشه از ۵ عدد کمتر باشد، داده‌های واقع شده در آن خوشه کاندیدای پرت هستند. در اینجا مقداردهی اولیه خوشه‌ها به صورت تصادفی انجام شده است. همچنین مربع تابع اقلیدسی به عنوان تابع اندازه‌گیری فاصله انتخاب شده است. در جدول ۳ نتایج اجرای این روش روی مجموعه داده‌ها نشان داده شده است. مقایسه نتایج ارائه شده در این جدول با جدول‌های ۲ و ۳ بیانگر تفاوت‌های بسیار زیادی در شناسایی داده های پرت است. با توجه به جدول ۴، مشاهده می‌شود که با اعمال روش خوشه بندی K-Means داده‌های بسیار کمتری به عنوان داده‌های پرت برای هر دو آزمایش انتخاب شده‌اند.

**جدول ۴: داده‌های پرت شناسایی شده در مجموعه داده‌ها با استفاده از روش خوشه‌بندی K-Means**

| شماره | مولفه سرعت | شماره داده پرت | | | | | تعداد |
|---|---|---|---|---|---|---|---|
| ۱ | U | --- | | | | | ۰ |
| | V | ۱۱۵۶ | ۵۶ | | | | ۲ |
| | W | --- | | | | | ۰ |
| ۲ | U | ۱۴۲۰ | ۱۲۱۳ | ۴۱۲ | | | ۳ |
| | V | --- | | | | | ۰ |
| | W | ۱۵۱۱ | ۴۱۲ | ۱۴۰۳ | ۱۰۰۲ | ۵۷۳ | ۵ |

**- نتایج اجرای روش رای گیری**

با توجه به نتایج حاصل از روش‌های MAD، خوشه‌بندی K-Means و ضریب چگالی محلی مشاهده شد که بعضی از

بنابراین بر اساس نتایج بدست آمده، تعیین داده پرت واقعی دشوار به نظر می‌آید. به همین دلیل استفاده از روش رای‌گیری، راه اصولی برای تعیین داده‌های پرت نهایی می‌باشد. بر اساس روش رای‌گیری، آن دسته از داده‌هایی که بیشتر روش‌ها آنها را به عنوان پرت انتخاب کرده‌اند، دارای پتانسیل بیشتری از لحاظ پرت بودن هستند. بنابراین استفاده از این روش می‌تواند به مراتب دقت نتایج حاصله را افزایش دهد. در این تحقیق داده‌هایی که توسط ۲ روش و یا بیشتر به عنوان پرت شناسایی شده‌اند، به عنوان کاندیدای نهایی داده پرت انتخاب شده‌اند. برای جستجوی نقاط و تعیین تعداد تکرارهای هر داده از الگوریتم جستجوی دودویی[1] استفاده شده است(Cormen, 1990). در جدول ۵ داده‌های پرت شناسایی شده در مجموعه داده‌ها با استفاده از روش رای‌گیری ارائه شده است. با توجه به اینکه روش رای‌گیری از اشتراک نتایج ارائه شده توسط سایر روش‌ها استفاده می‌کند، بنابراین می‌توان گفت نتایج ارائه شده توسط آن از دقت و اعتبار بیشتری برخوردار است. به همین دلیل در این تحقیق نتایج این روش (جدول ۵) به عنوان مبنای بررسی‌ها انتخاب شده است.

با توجه به این جدول، نقطه ۱ دارای داده پرت (در مجموع ۱۵ داده) کمتری نسبت به نقطه ۲ (در مجموع ۷۶ داده) است. روش‌ها یک داده را به عنوان پرت شناسایی کرده‌اند در حالی که روشی دیگر آن را نرمال به شمار آورده است.

این موضوع به دلیل تاثیر آبشکن بر تغییر الگوی جریان در پایین دست آن (یعنی در محدوده قرارگیری نقطه شماره ۲) است. داده‌های پرت شناسایی شده در مجموعه داده‌ها با استفاده از روش رای‌گیری در شکل ۸ ارائه شده است. همانطور که در این شکل ملاحظه می‌شود در آزمایش قوس بدون وجود آبشکن (نقطه شماره ۱)، بیشترین تعداد داده‌های پرت در راستای طولی گزارش شده در حالی که پس از استقرار آبشکن در قوس علاوه بر اینکه تعداد داده‌های پرت در راستای طولی همچنان زیاد است، اما راستای عمقی جریان داده‌های بیشتری را به عنوان کاندیدای داده پرت، نشان داده است.

پس از شناسایی داده‌های پرت می‌توان آنها را با توجه به روش‌های موجود تصحیح و یا حذف کرد. اگر تعداد این داده‌ها اندک باشد، می‌توان آنها را از مجموعه داده‌ها حذف کرد، ولی در صورتی که تعداد آنها زیاد باشد می‌توان داده‌ها را تصحیح کرد و یا اینکه عمل اندازه‌گیری را از ابتدا انجام داد. باید به این نکته مهم توجه داشت که همواره داده‌هایی که توسط روش‌ها به عنوان کاندیدای پرت انتخاب می‌شوند، بیانگر بروز خطا و یا عیب در اندازه‌گیری‌ها نیست و چه بسا ممکن است بر اثر تغییر در شرایط طبیعی سیستم (به عنوان مثال تغییر شرایط حاکم بر جریان، نوسانات جزئی برق و تأثیر آن بر دبی تولید شده

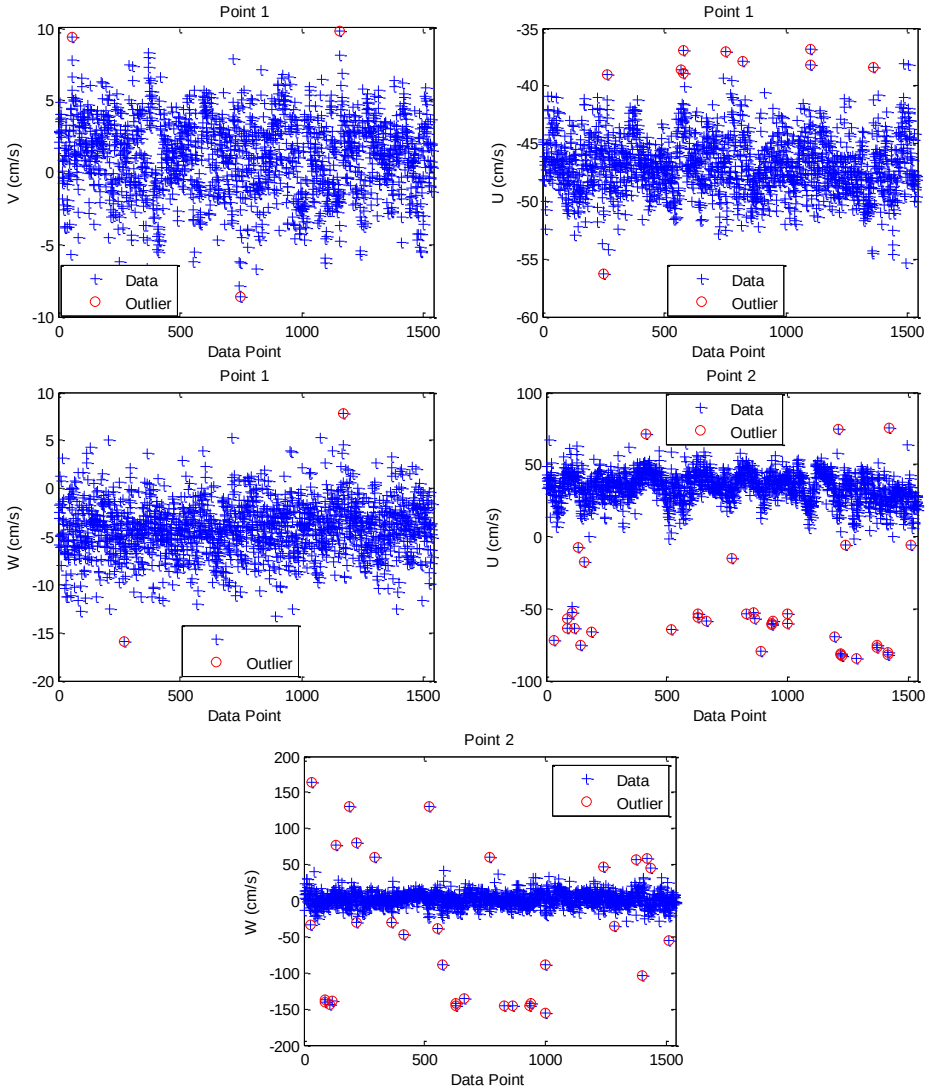
1. Binary search algorithm

توسط پمپ، خطاهای مشاهداتی، خطاهای سهوی[1] خطاهای سیستماتیک[2] خطاهای تصادفی[3] و سایر عوامل نظیر اشکالات موجود در دستگاه‌های اندازه‌گیری، عدم کالیبره بودن دستگاه-ها، عوامل انسانی مثل خطای دید، عدم تجربه و مهارت کافی کاربر در استفاده از دستگاه‌های اندازه‌گیری و...) به وجود آمده باشند. همچنین ممکن است این داده‌ها بیانگر رفتاری از سیستم مورد مطالعه باشند که تا کنون ناشناخته بوده‌اند و در نتیجه شناسایی آنها اطلاعات بسیار مهمی را در خصوص ماهیت مسئله در اختیار قرار دهد که این خود منجر به شناخت هرچه بهتر آن می‌شود. بنابراین باید پس از شناسایی داده‌های پرت، از لحاظ مختلف علل وقوع آنها را سنجید و مناسب‌ترین راه حل برخورد با آنها را انتخاب کرد؛ که این مهم در این تحقیق مد نظر قرار گرفته است. همچنین ذکر این نکته نیز ضروری است که همواره یک روش نسبت به سایر روش‌ها برتری ندارد و چه بسا ممکن است یک روش برای یک نوع داده خاص بسیار کارآمد باشد، در حالی که برای یک نوع داده دیگر کارآیی مطلوبی نداشته باشد. از اینرو پیشنهاد می‌شود هنگام کار با داده‌های مختلف، از روند مورد استفاده در این تحقیق استفاده شده و داده‌هایی به عنوان کاندیدای پرت نهایی انتخاب شوند که توسط بیشتر روش‌ها به عنوان پرت انتخاب شده‌اند.

**جدول ۵: داده‌های پرت شناسایی شده در مجموعه داده‌ها با استفاده از روش رای‌گیری**

| شماره | مولفه سرعت | شماره داده پرت | | | | | | | | تعداد |
|---|---|---|---|---|---|---|---|---|---|---|
| ۱ | U | ۲۵۰ | ۲۶۷ | ۵۶۹ | ۵۷۸ | ۵۷۹ | ۷۵۲ | ۸۲۰ | ۱۰۹۹ | ۱۰ |
| | | | | | | ۱۱۰۰ | ۱۳۵۷ | | | |
| | V | | | ۵۶ | | ۷۴۶ | ۱۱۵۶ | | | ۳ |
| | W | | | | ۲۶۹ | ۱۱۷۱ | | | | ۲ |
| | U | ۳۲ | ۸۶ | ۸۷ | ۱۰۹ | ۱۱۶ | ۱۳۳ | ۱۴۲ | ۱۵۷ | ۳۷ |
| | | | ۱۸۷ | ۴۱۲ | ۵۱۹ | ۶۳۱ | ۶۳۲ | ۶۶۴ | ۷۷۱ | |
| | | | ۸۳۱ | ۸۵۸ | ۸۶۷ | ۸۸۹ | ۹۳۶ | ۹۳۷ | ۹۴۲ | |
| | | | ۹۹۹ | ۱۰۰۰ | ۱۱۹۵ | ۱۲۱۳ | ۱۲۲۱ | ۱۲۲۳ | ۱۲۲۴ | |
| | | | ۱۲۴۲ | ۱۲۸۶ | ۱۳۷۱ | ۱۳۷۲ | ۱۴۱۵ | ۱۴۱۶ | ۱۴۲۰ | |
| | | | | | | ۱۵۱۱ | | | | |
| ۲ | V | | | | | --- | | | | ۰ |
| | W | ۳۰ ۳۲ | ۸۶ | ۸۷ | ۱۰۸ | ۱۰۹ | ۱۱۶ | ۱۳۳ | ۱۸۷ | ۳۹ |
| | | | ۲۱۹ | ۲۲۰ | ۲۹۵ | ۳۶۵ | ۴۱۲ | ۵۱۹ | ۵۵۳ | |
| | | | ۵۷۳ | ۶۳۱ | ۶۳۲ | ۶۶۴ | ۷۷۱ | ۸۳۱ | ۸۶۷ | |
| | | | ۹۳۷ | ۹۴۲ | ۱۰۰۰ | ۱۰۰۲ | ۱۱۹۵ | ۱۲۲۱ | ۱۲۴۲ | |
| | | | ۱۲۸۶ | ۱۲۸۷ | ۱۳۷۱ | ۱۳۷۵ | ۱۴۰۳ | ۱۴۱۶ | ۱۴۲۰ | |
| | | | | | | ۱۴۳۵ | ۱۵۱۱ | | | |

1. Spurious errors
2. Systematic errors
3. Random errors




**شکل ۸: نمایش داده‌های پرت شناسایی شده پس از اعمال روش رای‌گیری بر مجموعه داده‌های سرعت در جهات مختلف**

## نتیجه‌گیری

عوامل متعددی ممکن است سبب بروز داده‌های پرت در اندازه‌گیری‌های آزمایشگاهی شود. داده‌های پرت ممکن است بر اثر بروز خطا در اندازه‌گیری‌ها و یا تغییر ماهیت جریان به وجود آیند. به همین دلیل شناسایی این داده‌ها از جنبه‌های مختلف حائز اهمیت بوده و می‌تواند سبب افزایش دقت نتایج حاصل شده از داده‌های آزمایشگاهی شود. در این تحقیق با استفاده از آزمون MAD، خوشه‌بندی K-Means، ضریب چگالی محلی و روش رای‌گیری به شناسایی داده‌های پرت موجود در آزمایشات الگوی جریان در یک کانال با قوس ۱۸۰ درجه و با و بدون استقرار آبشکن T شکل پرداخته شده است. مقایسه روش های مختلف بیانگر این است که بهترین کارائی مربوط به روش رای-گیری بوده است. چون این روش از اشتراک نتایج سایر روش‌ها استفاده می‌کند؛ از اینرو نتایج ارائه شده توسط این روش قابل اعتمادتر است. مؤلفین این مقاله پیشنهاد می‌کنند قبل از تجزیه و تحلیل داده‌های گردآوری شده حاصل از آزمایشات جریان، از روند مورد بحث در این تحقیق جهت شناسایی داده-های پرت استفاده شود.

## منابع

محمودی، ک. و م. سایبانی. ۱۳۹۲. سلامت سنجی سیستم‌های دریایی به روش تشخیص ناهنجاری در یادگیری ماشین. اولین همایش ملی فناوری‌های نوین دریایی، نوشهر، ایران.

محمودی، ک.، م. واقفی، ع. مرادی و م. سایبانی. ۱۳۹۲. شناسایی خطاهای موجود در برداشت داده های مربوط به تعیین الگوی جریان و آبشستگی با استفاده از روش ضریب داده پرت محلی. پنجمین همایش ملی صنایع فراساحل(OIC2013)، تهران، ایران.

واقفی، م.، م. قدسیان و س. ع. ا. صالحی نیشابوری. ۱۳۸۷. مطالعه آزمایشگاهی الگوی جریان سه بعدی و آبشستگی در قوس ۹۰ درجه. مجله هیدرولیک، سال سوم، شماره ۳، ص ۴-۱۳.

واقفی، م.، م. قدسیان و س. ع. ا. صالحی نیشابوری. ۱۳۸۷. بررسی آزمایشگاهی الگوی جریان کمی اطراف آبشکن T شکل در قوس ۹۰ درجه با بستر صلب. مجله مهندسی عمران و محیط زیست، سال سی و هفتم، شماره ۳، ص ۸۱-۸۷.

واقفی، م.، م. قدسیان و س. ع. ا. صالحی نیشابوری. ۱۳۸۸. بررسی آزمایشگاهی اثر شعاع انحنا و موقعیت استقرار آبشکن های T شکل در قوس ۹۰ درجه بر میزان آبشستگی اطراف آن‌ها. مجله هیدرولیک ایران، سال چهارم، شماره ۱، ص ۹۱-۱۰۷.

Aggarwal, C.C., and P. Yu. 2001. Outlier detection for high dimensional data. Proc. of the ACM SIGMOD International Conference on Management of Data, California, U.S.A.

Akbari, M., and M. Vaghefi. 2017. Experimental investigation on streamlines in a 180° sharp bend. Acta Scientiarum Technology, 39(4): 425-432.

Anil, K.J. 2009. Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters, 31(8): 51-666.

Barnett, V., and T. Lewis. 1994. Outliers in Statistical Data. John Wiley and Sons, New York, U.S.A.

Billor, N., A. Hadi, and P. Velleman. 2000. BACON: Blocked Adaptive Computationally Efficient Outlier Nominators. Computational Statistics and Data Analysis, 34: 279-298.

Breunig, M. M., H. P. Kriegel, R. T. Ng, and J. Sander. 2000. LOF: Identifying density based local outliers. Proc. of the ACM SIGMOD Conference, Dallas, U.S.A.

Cormen, T.H., C. E. Leiserson, and R. L. Rivest. 1990. Introduction to Algorithms. MIT Press and McGraw-Hill, New York, U.S.A.

Duan, J.G., L. He, X. Fuand, Q. Wang. 2011. Turbulent burst around experimental spur dike. International Journal of Sediment Research, 26: 471-476.

Eskin, E., 2000. Anomaly Detection over Noisy Data using Learned Probability Distributions. Proc. of the International Conference on Machine Learning, California, U.S.A.

Eskin, E., A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. 2002. A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data, in Applications of Data Mining in Computer Security. Kluwer Academic Publishers, Boston, U.S.A.

Ester, M., H. P. Kriegel, J. Sanderand, X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. KDD, California, U.S.A.

Fazli, M., M. Ghodsianand, S.A.A. Salehi Neyshabouri. 2009. Scour and flow field around a spur dike in a 90 bend. International Journal of Sediment Research, 23(1): 56-68.

Ghodsian, M., and M. Vaghefi. 2009. Experimental study on scour and flow field in a scour hole around a T-shapedspur dike in a 90° bend. International Journal of Sediment Research, 24: 145-158.

Giri, S., Y. Shimizuand, and B. Surajate. 2004. Laboratory measurement and numerical simulation of flow and turbulence in a meandering-like flume with spurs. Flow Measurement and Instrumentation, 15(5): 301-309.

Hawkins, S., H. He, G. Williamsand, and R. Baxter. 2012. Outlier Detection Using Replicator Neural Networks. Proc. of the 4th International Conference on Data Warehousing and Knowledge Discovery (DaWaK02), France.

Keshavarzi, A., B. Melville, and J. Ball. 2014. Three-dimensional analysis of coherent turbulent flow structure around a single circular bridge pier. Environmental Fluid Mechanics, 14(4): 821-847.

Knorr, E., and R. Ng. 1998. Algorithms for mining distance based outliers in large data sets. Proc. of the Very Large Databases (VLDB) Conference, New York, U.S.A.

Latecki, L. J., A. Lazarevic, and D. Pokrajac. 2007. Outlier Detection with Kernel Density Functions. Proc. of 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM), Germany.

Lazarevic, A., L. Ertoz, A. Ozgur, J. Srivastava, and V. Kumar. 2003. A comparative study of anomaly detection schemes in network intrusion detection. Proc. of the Third SIAM International Conference on Data Mining, California, U.S.A.

Ng, R., and J. Han. 1994. Efficient and Effective Clustering Methods for Spatial Data Mining. Proc. of VLDB'94, New York, U.S.A.

Nortek, A.S. 2009. Vectrino velocimeter user guide, Nortek AS, Vangkroken, Norway.

Papadimitriou, S., H. Kitagawa, P. B. Gibbons, and C. Faloutsos. 2003. LOCI: Fast Outlier Detection Using the Local Correlation Integral. Proc. of the 19th International Conference on Data Engineering (ICDE'03), India.

Sukhodolov, A. N. 2014. Hydrodynamics of groyne fields in a straight river reach: insight from field experiments. Journal of Hydraulic Research, 52(1): 105-120.

Tiwari, H., and N. Sharma. 2015. Turbulence study in the vicinity of piano key weir: relevance, instrumentation, parameters and methods. Applied Water Science, 7(2): 525-534.

Vaghefi, M., K. Mahmoodi, and M. Akbari. 2018a. Detection of Outlier in 3D Flow Velocity Collection in an Open-Channel Bend Using Various Data Mining Techniques. Iranian Journal of Science and Technology, Transactions of Civil Engineering, 1-18.

Vaghefi, M., K. Mahmoodi, and M. Akbari. 2018b. A comparison among data mining algorithms for outlier detection using ow pattern experiments. Scientia Iranica, 25(2): 590−605.

Vaghefi, M., M. Akbari, and A. R. Fiouz. 2016. An experimental study of mean and turbulent flow in a 180 degree sharp open channel bend: Secondary flow and bed shear stress. KSCE Journal of Civil Engineering, 20(4): 1582-1593.

Vaghefi, M., M. Ghodsianand, and M. Akbari. 2017. Experimental Investigation on 3D Flow around a Single T-Shaped Spur Dike in a Bend. Periodica Polytechnica Civil Engineering, 61(3): 462-470.

Vaghefi, M., M. Ghodsianand, and S. A. A. Salehi Neyshabouri. 2012. Experimental study on scour around a T-shaped spur dike in a channel bend. Journal of Hydraulic Engineering, 138(5): 471-474.

Yu, D., G. Sheikholeslamiand, and A. Zhang. 2002. Find Out: Finding Outliers in Very Large Datasets. The Knowledge and Information Systems (KAIS), 4: 35-47.

# Identification of Outlier Data in Flow Pattern Experiments in a Bend by Using Statistical Methods

**Mohammad. Vaghefi[1,*], Kumars. Mahmoodi , Maryam. Akbari[3]**

Various factors such as human or instrument errors, measurement conditions, and the nature of the flow under unique circumstances may lead to generation of data inconsistent with the normal pattern of the statistical population, and result in the assumption that they may have been generated through a different process. In a general definition, these data are called outlier data. Identification of outliers is significant in many aspects, and will thus result in an ever better and more precise understanding of flow pattern. The main purpose of this study was analysis and identification of outliers existing in flow pattern experiments in a bend channel with a central angle of 180 degrees and width of 1 meter in the presence and absence of a spur dike in the bend by employing statistical methods. The intended channel is located in the Hydraulic Laboratory of Persian Gulf University, and Vectrino velocimeter has been utilized for collection of 3D flow velocities. Median of Absolute Deviations (MAD), K-Means Clustering, Local Density Factor (LDF), and Voting were the methods employed for outlier detection in this study. The results of applying these methods on the collected experimental data suggested that most of the methods were efficient and appropriate. Eventually, the Voting method was used to achieve the optimum results in this paper. In this method, the data which have been identified as outlier by most of the methods are considered the final candidates as outlier.

**Key Words:** Statistical Methods, Outlier Data, Flow Pattern, 180 Degree Sharp Bend, Vectrino Velocimeter.

[1,*] Associate Professor of Hydraulic Structures, Department of Civil Engineering, Persian Gulf University, Bushehr, Iran (Corresponding Author); Email: Vaghefi@pgu.ac.ir
[2]Ph.D. student, Department of Marine Engineering, Amirkabir University of Technology, Tehran, Iran
[3]M.Sc. Graduated Student of Hydraulic Structures, Department of Civil Engineering, Persian Gulf University, Bushehr, Iran