مروری بر داده های خطا و عوامل ایجاد آنها در داده‌های حاصل از کمی‌سازی تغییرات خط ساحلی

کیومرث محمودی[۱]، عباس مرادی[۲]، مصباح سایبانی[۳]

kumarsmahmoodi@aut.ac.ir

۱- دانشجوی کارشناسی ارشد مهندسی سواحل، دانشکده مهندسی دریا، دانشگاه صنعتی امیرکبیر

۲- استادیار، دانشکده علوم و فنون دریائی، دانشگاه هرمزگان

۳- استادیار، دانشکده مهندسی دریا، دانشگاه صنعتی امیرکبیر

چکیده

داده‌های حاصل از کمی‌سازی تغییرات خط ساحلی در طی زمان، به عنوان مبنای بسیاری از برنامه‌های مدیریت یکپارچه مناطق ساحلی به شمار می‌آیند. جهت حصول نتایج مطلوب از این داده‌ها، آنها بایدنمایندهواقعیمشاهداتبودهوغیر واقعیااشتباهنباشند.عوامل متعددی ممکن است در داده‌های تغییرات خط ساحلی سبب بروز داده مشکوک به خطا شوند.در این تحقیق به معرفی داده‌های مشکوک به خطا، عوامل ایجاد آنها در داده‌های تغییرات خط ساحلی، روش‌های شناسایی و نحوه برخورد جهت کاهش نقش آنها در دقت محاسبات، پرداخته شده است.

واژگان کلیدی

خط ساحلی، کمی‌سازی خط ساحلی، داده مشکوک به خطا، شناسایی داده پرت

تاریخ دریافت مقاله :	۹۳/۲/۷
تاریخ پذیرش مقاله :	۹۳/۸/۲۵

۱- مقدمه

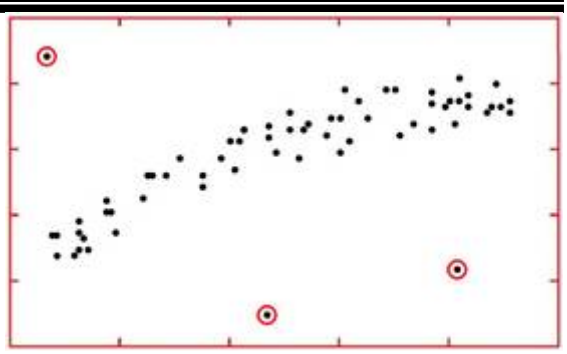
دربسیاری ازمطالعات آزمایشگاهی شمار زیادی از متغیرها ثبت و یا نمونه برداری می‌شوند. این مشاهدات وقتی کارآمدند که نماینده واقعی موضوع مورد مطالعه باشند. برای بدست آوردن نتایج مناسب و مطلوب از داده های جمع آوری شده، داده ها باید نماینده واقعی مشاهدات بوده و غیر واقعی االشتباه نباشند. همچنین از داده‌های آزمایشگاهی در صحت سنجی مدل‌های عددی و ریاضی استفاده می‌شود، از اینرو برداشت داده‌های آزمایشگاهی به صورت دقیق و باکمترین خطا اهمیت می‌یابد [۱، ۲]. برخی ازمشاهدات حاصل از یک آزمایش ممکن است براثرتغییر در رفتار سیستم، خطاهای انسانی، خطاهای دستگاهی، انحرافات طبیعی در جمعیت نمونه‌ها و یا براثر عواملی گذرا که موجب بد عملکردن سیستم می‌شوند به صورت متناقض و غیرمتعارف با سایر مشاهدات مجموعه داده ایجاد شوند. در یک تعریف عام به این نوع از مشاهدات، مشاهدات پرت[1] گفته می‌شود. در علوم مختلف و با توجه به نوع کاربری، به مشاهدات پرت نام‌های مختلفی نظیر داده خطا، نویز، عیب، نفوذ و غیره نسبت داده می‌شود. در این تحقیق این نوع از مشاهدات، مشاهدات مشکوک به خطا نامیده شده است. مشاهدات خطا نتایج حاصل از تجزیه و تحلیل داده‌ها را تحت تاثیر قرار داده و منجربه نتیجه‌گیری نادرست از مشاهدات می‌شوند. ازاینرو باید قبل از تحلیل و نتیجه‌گیری از داده‌ها، داده ای مشکوک به خطا راشناسایی و نسبت به رفع آنها اقدامکرد.

عوامل متعددی ممکن است در داده‌های تغییرات خط ساحلی سبب بروز داده مشکوک به خطا شوند. در این تحقیق به معرفی داده‌های مشکوک به خطا، عوامل ایجاد آنها در داده‌های تغییرات خط ساحلی، روش‌های شناسایی و نحوه برخورد جهت کاهش نقش آنها در دقت محاسبات، پرداخته شده است.

۲- تعریف داده مشکوک به خطا

همانطور که در مقدمه گفته شد، در این تحقیق داده‌های پرت به عنوان داده‌های مشکوک به خطا در نظر گرفته شده‌اند. تا کنون از داده پرت تعاریف متعددی ارائه شده است، با این حال ارائه یک تعریف دقیق از داده پرت وابسته به ماهیت داده‌ها و روش بکار گرفته شده در شناسایی آنها است. تعریف ریاضیاتی ثابتی برای آنکه چه چیزی داده پرت را به وجودمی‌آوردوجود ندارد، تشخیص اینکه یک مشاهده، پرت است یا نه صرفا یک فرایند کیفی و ابتکاری است که به تجربه شخصی فرد و کاربرد داده‌ها وابسته است. در آمار، داده پرت به مشاهده‌ای گفته می‌شود که به صورت عددی از سایر داده‌ها دور باشد.[۲] مشاهده‌ای را پرت در نظر می‌گیرد که به طور مشخصی از سایر اعضای نمونه دور افتاده باشد. [۳] داده پرت را مشاهده‌ای در مجموعه داده‌ها تعریف می‌کند که به صورت متناقض با سایر داده‌ها ظاهر می‌شود. [۴] به مشاهده‌ای پرت می‌گوید که به صورت محسوسی با سایر عضوهای نمونه متفاوت باشد. از دیدگاه [۵] داده پرت مشاهده،نمونه یا داده‌ای است که به طور قابل توجهی غیر مشابه یا متناقض با سایر داده‌ها باشد، به‌گونه‌ای که این سوء ظن ایجاد شود که با یک روند متفاوت تولید شده است. در این تحقیق، از تعریف [۵] به عنوان مبنای شناسایی داده پرت استفاده شده است. به عنوان مثال شکل ۱ بیانگر یک مجموعه داده با تعداد سه داده پرت است. این داده‌ها با یک علامت دایره به دور آنها در شکل مشخص شده‌اند. همانطور که از این شکل مشخص است این نقاط به صورت محسوسی متناقض با سایر اعضای نمونه بوده و از الگوی کلی داده‌ها دور افتاده‌اند و به نظر می‌رسد که با یک روند متفاوت ایجاد شده‌اند.

[1] Outlier



شکل (۱)یک نمونه با تعداد ۳ داده پرت

۳- شناسایی داده‌های مشکوک به خطا

شناسایی داده‌های پرت[1] (داده‌های مشکوک به خطا)، شاخه‌ای کاربردی و بسیار مهم از داده‌کاوی[2] است که به شناسایی الگوهای متناقض با جمعیت نرمال نمونه، در یک جامعه آماری می‌پردازد.روش‌های اصلی تشخیص داده‌های پرت در گذشته موردی و ابتکاری بودند،اما امروزه از تکنیک‌های روشمند که با توسعه علم کامپیوتر و آمار محقق شده است، بهره برده می‌شود. تا کنون دسته-بندی‌های مختلفی از روش‌های شناسایی داده‌های پرت ارائه شده است. در یک دسته بندی، روش‌های شناسایی داده‌های پرت را می‌توان به روش‌های تک متغیره[3] و روش‌های چند متغیره[4] تقسیم‌بندی کرد. بسیاری از روش‌هایی که قبلاً برای شناسایی داده‌های پرت ارائه شده‌اند، با مجموعه داده‌های تک متغیره بکار می‌روند. این روش‌ها به یک تخمین کلی از پراکندگی و یا به مقدار داده‌ها وابسته‌اند. یکی از اصلی‌ترین عیوب این روش‌ها این است که عملکرد آنها به اندازه نمونه بستگی دارد. روش‌های چند متغیره قابل بکارگیری با مجموعه داده‌هایی هستند که دارای یک و یا چندین متغیر می‌باشند. در یک دسته‌بندی دیگر، روش‌های شناسایی داده‌های پرت به روش‌های آماری و روش‌های غیر آماری طبقه‌بندی می‌شوند. روش‌های آماری توزیع مشخصی از مشاهدات را در نظر می‌گیرند [۲] و یا وابسته به تخمین‌های آماری از توزیع پارامترهای ناشناخته هستند [۳]. این روش‌ها مشاهداتی را پرت در نظر می‌گیرند که به نظر می‌رسد از مدل دور افتاده‌اند و اغلب برای مجموعه داده‌های با بعد بالا و همچنین برای یک مجموعه داده بدون آگاهی قبلی از توزیع داده‌ها مناسب نیستند [۴]. روش‌های غیر آماری از الگوریتم‌های نوین کامپیوتری، نظیر الگوریتم‌های یادگیرنده برای شناسایی داده‌های پرت استفاده می‌کنند. اغلب توانایی این روش‌ها از روش‌های آماری بیشتر بوده و می‌توانند با مجموعه داده‌های چند بعدی بکار روند.

در یک دسته‌بندی دیگر، تمامی روش‌های شناسایی داده‌های پرت را در یک دسته به نام روش‌های داده‌کاوی قرار می‌دهند. تکنیک‌های داده‌کاوی که برای شناسایی داده‌های پرت توسعه یافته‌اند، شامل هر دو روش یادگیری نظارت شده[5] و یادگیری نظارت نشده[6] می‌شوند. روش‌های مبتنی بر یادگیری نظارت شده با ایجاد یک مدل پیش‌بینی کننده بر اساس داده‌های علامت‌گذاری شده (داده‌هایی که نرمال و یا غیر نرمال بودن آنها از پیش مشخص شده باشد)، داده‌های پرت را شناسایی می‌کنند [۵، ۶]. این روش‌ها دارای ایراداتی هستند از جمله اینکه:

۱) نیازمند داده‌های آموزشی[7] (داده‌های علامت‌گذاری شده) می‌باشند. این عامل به خصوص برای مجموعه داده‌های واقعی از لحاظ زمانی بسیار هزینه‌بر است.

۲) عدم توانایی در شناسایی داده‌های پرت (نامتعارف) جدید. این ضعف به این دلیل است که مدل بر اساس داده‌های علامت‌گذاری شده آموزش دیده و قادر به شناسایی انواع جدید نیست.

روش‌هایی که از تکنیک‌های نظارت نشده استفاده می‌کنند نیازمند داده‌های آموزشی نیستند و با استفاده از یک سری اندازه‌گیری‌های از پیش تعیین شده، داده‌های پرت را شناسایی می‌کنند. این روش‌ها داده‌هایی را که از الگوی نرمال نمونه طبعیت نکنند به عنوان داده مشکوک به پرت

[1] Outlier Detection
[2] Data Mining
[3] Univariate Methods
[4] Multivariate Methods
[5] Supervised Learning
[6] Unsupervised Learning
[7] Training Data

شناسایی می‌کنند. عملکرد موفقیت‌آمیز این روش‌ها وابسته به نوع داده‌های ورودی، نوع الگوریتم، تجربه شخصی کاربر و انتخاب صحیح پارامترهای آنها است. الگوریتم‌هایی که از تکنیک‌های نظارت نشده استفاده می‌کنند دارای نقاط ضعفی نیز هستند، از جمله اینکه ممکن است خطای عملکرد آنها زیاد باشد، یعنی اینکه ممکن است برخی از داده‌های نرمال را به عنوان پرت تشخیص دهند و یا بالعکس.

در یک دسته‌بندی می‌توان روش‌های شناسایی داده‌های پرت را در چهار دسته طبقه‌بندی کرد:

۱. روش‌های آماری
۲. روش‌های مبتنی بر فاصله
۳. روش‌های مستندی[1]
۴. روش‌های مبتنی بر مدل

در تکنیک‌های آماری [۲، ۷، ۸] داده‌های نمونه معمولاً با استفاده از یک توزیع احتمالی مدل شده و داده‌ها بر حسب رابطه‌ای که با مدل توزیعی دارند، علامت‌گذاری می‌شوند. روش‌های مبتنی بر فاصله [۹-۱۱] با اندازه‌گیری فاصله بین نقاط، داده‌های پرت را شناسایی می‌کنند. الگوریتم‌های مبتنی بر فاصله که تا کنون ارائه شده‌اند یا با اندازه‌گیری فاصله بین نقاط [۹، ۱۱] و یا با تخمین چگالی همسایه‌های محلی نقاط [۶، ۱۰] داده‌های پرت را شناسایی می‌کنند. 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 [۱۲] و یا اندازه یک خوشه به طور قابل توجهی کوچک‌تر از سایر خوشه‌ها باشد، می‌تواند کاندیدای داده پرت باشند [۱۳]. 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] [۱۴] و یا ماشین-ماشین‌های بردار پشتیبان بدون نظارت[4] [۱۳، ۱۵]) مشخص می‌کنند، آنگاه داده‌هایی که انحراف آنها از مدل آموزش دیده زیاد باشد را به عنوان داده پرت در نظر می‌گیرند.

۴- نحوه برخورد با داده‌های مشکوک به خطا شناسایی شده

پس از شناسایی داده‌های مشکوک به خطا، باید در خصوص آنها تصمیم‌گیری شود. اگر تعداد آنها در نمونه مورد بررسی اندک باشد، می‌توان آنها را از نمونه حذف کرد. در غیر این صورت می‌توان آنها را با استفاده از روش‌های تصحیح داده، تصحیح نمود. باید توجه داشت که همواره داده‌های مشکوک به خطا بیانگر خطا و یا اختلال در سیستم نیستند، گاه ممکن است براثر رفتار یکتای سیستم در شرایط خاص به وجود آمده باشند و چه بسا ممکن است حاوی اطلاعات مهمی از رفتار سیستم باشند که تا کنون ناشناخته بوده‌اند. از اینرو باید پس از شناسایی این داده‌ها علت وقوع آنها را بررسی کرده و با درک نحوه اثرگذاری آنها در نمونه، با آنها به طور صحیح برخورد کرد.

۵- عوامل ایجاد داده‌های خطا در داده‌های تغییرات خط ساحلی

یکی از مهمترین مواردی که در یک فرآیند اندازه‌گیری مورد توجه است، خطا در اندازه‌گیری‌ها است. در یک مطالعه فیزیکی، در تمام مراحل طراحی، نصب، اجرا، اندازه‌گیری و تحلیل داده‌ها امکان ایجاد خطا وجود دارد. خطاهای ایجاد شده سبب کاهش میزان صحت و اعتماد پذیری نتایج می‌شوند. در داده‌های تغییرات خط ساحلی نیز عوامل متعددی ممکن است سبب بروز داده‌های خطا

[1] Profling Methods
[2] Clustering Base Methods
[3] Replicator Neural Networks
[4] Unsupervised Support Vector methods

شوند. در این قسمت به شماری از این عوامل اشاره شده است.

با توجه به تنوع خطاهای آزمایشگاهی، خطاها از منظرهای مختلف دسته‌بندی می‌شوند. در یک دسته‌بندی می‌توان خطاها را به سه دسته خطاهای سهوی، خطاهای سیستماتیک و خطاهای تصادفی طبقه‌بندی کرد:

۱. خطاهای سهوی[1] (اتفاقی): این خطاها بیشتر ناشی از اشتباهات انسانی یا دستگاهی است که باعث ایجاد خطا در اندازه‌گیری‌ها می‌شود. معمولاً با تکرار اندازه‌گیری و دقت بیشتر می‌توان این خطاها را کاهش داد. خطاهای سهوی هم از نظر مقدار و هم از نظر علامت دارای مقادیر ثابتی نیستند.

۲. خطاهای سیستماتیک[2]: این خطاها بر اثر عوامل ناشناخته حین اندازه‌گیری ایجاد شده و می‌توانند بوسیله دستگاه‌های اندازه‌گیری و یا در اثر شرایط اندازه‌گیری به وجود آیند. خطاهای سیستماتیک یک کمیت هم از نظر مقدار و هم از نظر علامت دارای مقادیری ثابتی‌اند که مقدار آنها را می‌توان با استفاده از روش‌های محاسباتی تحلیلی و یا تئوریک برآورد نمود.

۳. خطاهای تصادفی[3]: این خطاها ممکن است از یک توزیع آماری مشخص پیروی کنند که اثر آنهامی‌تواند با افزایش تعداد اندازه‌گیری کاهش یابد. خطاهای تصادفی می‌تواند بر اثر عواملی نظیر نوسانات قرائت که ممکن است بر اثر عوامل متعددی مثل نوسانات الکتریکی دستگاه‌های اندازه‌گیری، لغزش‌های جزئی در وسیله مورد نظر، لغزش‌های مکانیکی و غیره به وجود آید.

در یک دسته‌بندی خطاهای موجود در داده‌های تغییرات خط ساحلی به صورت زیر قابل طبقه‌بندی‌اند [۱۶]:

۱. خطاهای مفهومی
۲. خطاهای داده‌های منبع
۳. خطاهای کدگذاری داده
۴. خطاهای تبدیل و ویرایش داده
۵. خطاهای تحلیل و پردازش داده
۶. خطاهای خروجی داده

خطاهای مفهومی:

این خطاها از درک ما از واقعیت داده‌های تغییرات خط ساحلی و چگونگی مدل‌سازی آن ایجاد می‌شوند. آگاهی و فهم واقعیت داده‌های تغییرات خط ساحلی از فردی به فرد دیگر تغییر می‌کند، که این امر بر داده‌ها اثرگذار است. ماهیت فیزیکی تغییرات خط ساحلی بسیار پیچیده است، متخصصین علوم ساحلی معمولاً با ساده-سازی واقعیت سعی در مدل‌سازی و کمی‌سازی تغییرات خط ساحلی می‌کنند، که در این صورت هر ساده‌سازی ممکن است سبب بروز داده‌های خطا شود.

خطاهای داده‌های منبع:

داده‌های توصیفی و مکانی گردآوری شده از منابع داده-های تغییرات خط ساحلی ممکن است دارای خطاهایی باشند. این خطاها در داده‌های نقشه‌برداری، ناشی از خطاهای مشاهده‌ای، خطاهای ابزاری و خطاهای فردی هستند؛ در حالی که داده‌های گردآوری شده از طریق سنجش از دور و عکس‌برداری هوایی خطاهایی ناشی از ارجاع مکانی نادرست و اشتباه در طبقه‌بندی و تفسیر دارند. تغییرات زمانی در عوارض نیز خطاهایی ناشی از زمان و تاریخ کسب داده‌ها ایجاد می‌کنند. نقشه‌ها و نمودارها نیز مثل داده‌های توصیفی دارای خطاهایی هستند که ناشی از ضعف تجهیزات و یا خطاهای انسانی هستند. مثلاً فرایند کارتوگرافی مورد استفاده در تولید نقشه، سبب ایجاد خطاهای نامحسوس در تولید نقشه می-شود.

خطاهای کدگذاری داده:

فرایند انتقال داده‌های گردآوری شده از طریق منابع داده (مثل نقشه‌ها، سنجش از دور و پژوهش‌های زمینی) در قالب GIS به کدگذاری داده معروف است. کدگذاری داده ممکن است بزرگ‌ترین منبع تولید خطا باشد. رقومی‌سازی نقشه یکی از فرایندهای کدگذاری داده در GIS است. این

[1] Spurious Errors
[2] Systematic Errors
[3] Random Errors

کار می‌تواند با استفاده از سخت‌افزار رایانه‌ای مناسب به شکل دستی و یا به صورت خودکار انجام شود. با وجود قابلیت استفاده سخت‌افزار برای رقومی‌سازی خودکار، بیشتر کار رقومی‌سازی به صورت دستی با دخالت، قضاوت و محدودیت‌های بشر، که یکی از منابع اصلی خطا در GIS است، انجام می‌شود. به عنوان مثال انتقال خط ساحلی از روی نقشه به یک تصویر رقومی، از طریق نمونه‌برداری انجام می‌شود که این عمل به دلیل محدودیت‌های بشر و محدودیت دستگاه‌های بکار گرفته شده ممکن است به درستی و با دقت کافی انجام نشود. همچنین خطاها ممکن است در نتیجه ثبت نادرست انواع نقشه‌ها و تصاویر پیش از شروع رقومی‌سازی از طریق دستی و خودکار ایجاد شوند. اسکنرهای رستر که در رقومی‌سازی خودکار استفاده می‌شوند، دارای مشکلاتی نظیر قدرت تفکیک می‌باشند.

خطاهای تبدیل و ویرایش:

از آنجایی که ورودی داده‌ها از طریق رقومی‌سازی دستی و خودکار بدون خطا نیست، همواره به ویرایش و پاکسازی نیاز خواهد بود. تشخیص دقیق جای خطاها و از بین بردن آنها دشوار است، اما بسیاری از آنها را می‌توان با بررسی دقیق داده‌ها رفع کرد. هنگام کار با GIS رستری با استفاده از روش‌های خودکار برای پاکسازی، مشکلات متفاوتی ایجاد می‌شود. در GIS رستری، نویز (طبقه‌بندی اشتباه پیکسل‌ها) می‌تواند منظم و یا تصادفی باشد. تعیین نویز منظم آسان، ولی تعیین نویز تصادفی مشکل است. خطاهای نویز را می‌توان با بکارگیری فیلترهای مناسب برطرف کرد. این فیلترها یک و یا تعدادی از فیلترها را با روند کلی نمونه مطابقت داده و از این طریق نویز در پیکسل‌ها را تشخیص می‌دهند. در اینجا نیز انتخاب صحیح فیلتر ضروری است. چون انتخاب نامناسب فیلتر ممکن است تغییرات اصلی در داده‌ها را حذف کرده و یا خود سبب تولید مقدار زیادی نویز شود.

پس از ویرایش و پاکسازی داده‌ها، تبدیل داده‌های برداری به داده‌های رستری و برعکس، خود ممکن است سبب تولید داده خطا شود. مثلاً هنگام تبدیل داده‌های برداری به داده‌های رستری، اندازه سلول‌های رستر و نیز روش رستر کردن، نقش مهمی در ایجاد خطای موضعی و در برخی موارد عدم قطعیت داده‌های توصیفی دارند. همچنین تبدیل از بردار به رستر ممکن است به از دست دادن چند ضلعی‌های کوچک و نقشه رستر متفاوت منجر شود، که به علت قرار گیری نادرست شبکه برای رسترسازی با توجه به جهت‌گیری و مبدأ است. همچنین هنگامی که داده‌های ایجاد شده از طریق یک سامانه توسط سامانه دیگر خوانده می‌شوند، گاهی به تبدیل داده نیاز است. در چنین مواردی انتقال یک پایگاه داده از یک بسته نرم‌افزاری به یک بسته دیگر منجر به خطاهایی می‌شود که به طور معمول خطاهای مکانیکی[1] نامیده می‌شوند.

خطاهای پردازش و تحلیل:

خطاهای پردازش داده که می‌توانند باعث تولید خطا شوند عبارتند از: نامناسب بودن داده‌ها جهت تحلیل و پردازش، عدم همخوانی مجموعه‌های داده، عدم تناسب روش‌های بکارگرفته شده، طبقه‌بندی داده، تجمیع و تفکیک داده-های منطقه‌ای و ادغام داده‌ها از طریق روش‌های کلی همپوشانی. مثلاً یک خطای معمول در هنگام قرارگیری دو نقشه چندضلعی روی هم، شکل‌گیری چندضلعی‌های کوچکی در مرزهای دو نقشه ورودی است که به آن سلیورس[2] (درهم فرورفتگی) گفته می‌شود. وجود سلیورس به طور معمول ناشی از خطاهای رقومی‌سازی است. وقتی دو یا چند نقشه یک منطقه به طور جداگانه و در زمان-های متفاوت رقومی و یا اسکن می‌شوند و سپس روی هم قرار می‌گیرند، مرزهای رقومی شده همدیگر را قطع می‌کنند و سبب ایجاد سیلورس می‌شوند. همچنین گاهی خطاهای موقعیت ممکن است خطاهای توصیفی در عملیات روی هم قرار گرفتن نقشه‌ها را ایجاد کند. روی هم قرار گرفتن نقشه‌های تهیه شده از یک منطقه خاص

[1] Mechanical Errors
[2] Slivers

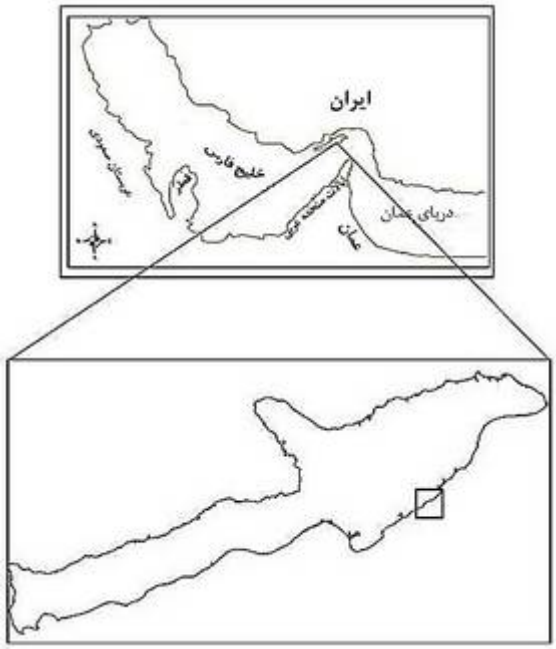
طی چندین سال، یکی از اعمال متداول جهت کمی‌سازی نرخ تغییرات خط ساحلی است. بنابراین ایجاد خطاهای سیلورس در این مورد بسیار شایع است.

خطاهای خروجی داده:

به دلیل نبود صحت در پایگاه داده‌های ایجاد شده و خطاهای ایجاد شده در بکارگیری و تحلیل داده‌ها، وجود خطا در تمامی خروجی‌ها اجتناب‌ناپذیر است. میزان این خطا به دقت و مراقبت در تمامی مراحل از ساخت و بکارگیری تا تحلیل پایگاه داده‌ها بستگی دارد.

۴- بحث

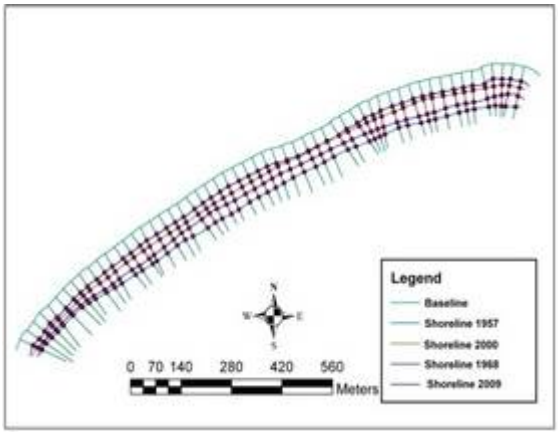
در این قسمت به عنوان نمونه داده‌های مشکوک به خطا (پرت) موجود در یک مجموعه داده مربوط به اندازه‌گیری نرخ تغییرات خط ساحلی شناسایی شده است.نمونه مورد بررسی حاصل کمی‌سازی نرخ تغییرات خط ساحلی منطقه سوزا (واقع در جزیره قشم) است.در شکل ۲ موقعیت این منطقه روی نقشه نشان داده شده است. همچنین در جدول ۱ مشخصات ناحیه ساحلی انتخاب شده برای بررسی نرخ تغییرات خط ساحلی آن، ارائه شده است.اندازه‌گیری نرخ تغییرات خط ساحلی به کمک ابزار DSAS[۱۷] و در محیط نرم‌افزار ArcGIS انجام شده است. منابع داده‌ای که خط ساحلی از آنها استخراج شده، عبارتند از: عکس‌های هوایی اسکن شده با کیفیت ۵۰۰DPI[1]، عکس‌های ماهواره‌ای با دقت بالا شامل GeoEye، QuickBird و panchromatic IRS.در شکل ۳ مقاطع عرضی ایجاد شده توسط DSAS، در محیط ArcGIS نشان داده شده است. فاصله بین تمامی مقاطع ۳۰ متر در نظر گرفته شده که در مجموع باعث تولید ۵۴ مقطع شده است. با استفاده از داده‌های موجود، در مجموع ۴ خط ساحلی استخراج شده که متعلق به سال‌های ۱۹۵۷، ۱۹۶۸، ۲۰۰۰ و ۲۰۰۹ میلادی است. پس از ایجاد مقاطع، نرخ تغییرات خط ساحلی حول مقاطع ایجاد شده نسبت به خط مبنا، با استفاده از روش‌های EPR[۱۸] و LRR[۱۹] محاسبه شد. در شکل ۴ نرخ‌های اندازه‌گیری شده نشان داده شده است.



شکل (۲) موقعیت منطقه سوزا روی نقشه

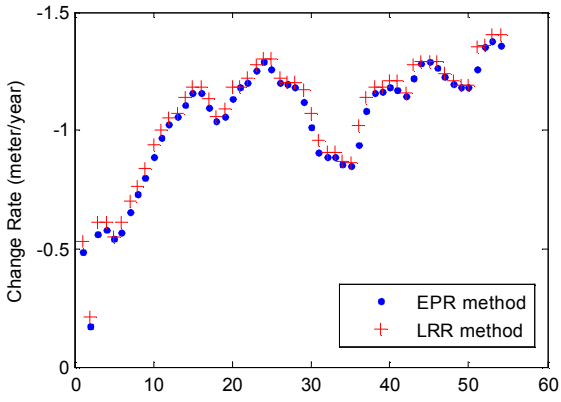
جدول (۱) مشخصات ناحیه ساحلی مورد بررسی

جنس	شنی
طول انتخابی	۱/۶۳ کیلومتر
محل	قسمت جنوبی جزیره قشم



شکل (۳) مقاطع عرضی ایجاد شده توسط DSAS، در ناحیه خط ساحلی سوزا

[1] Dot Per Inch (DPI)



شکل (۴) نرخ تغییرات خط ساحلی سوزا به روش EPR و LRR

در اینجا برای شناسایی داده‌های پرت موجود در مجموعه داده‌های EPR و LRR از آزمون chi-square score استفاده شده است. با استفاده از این آزمون می‌توان داده-های پرت را در مجموعه داده‌های تک متغیره شناسایی کرد. برای انجام این آزمون از رابطه زیر استفاده می‌شود:

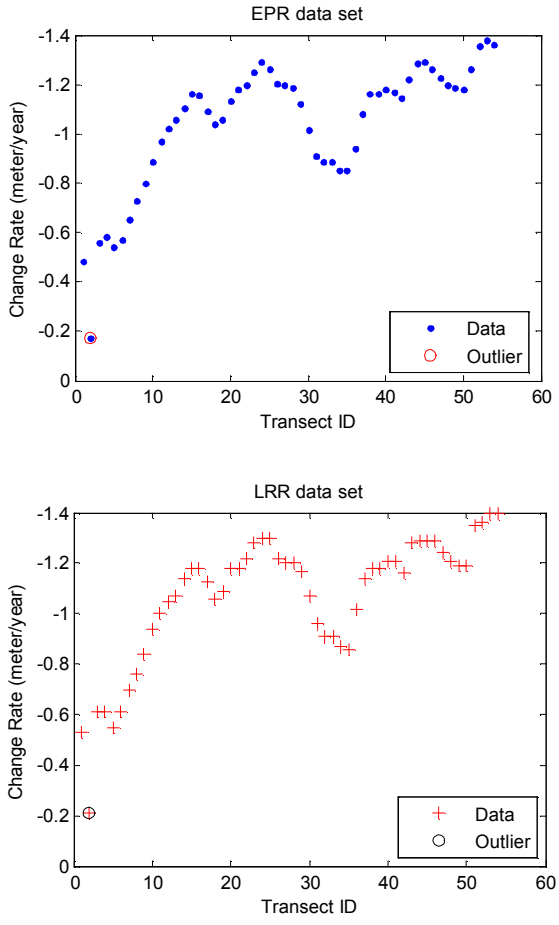
$$X \quad (۱)$$

که X_X عضو X ام مجموعه داده، $\bar{X}$ میانگین حسابی نمونه‌ها و XX انحراف معیار مجموعه داده است. در این روش داده‌ای به عنوان کاندیدای پرت در نظر گرفته می‌شود که مقدار $X^X{}_{XXXXX}(X)$ آن از حد آستانه تعریفی کاربر (X) بیشتر باشد [۲۰]. انتخاب مقدار پارامتر آستانه وابسته به ماهیت داده‌های ورودی بوده و معمولا با توجه به نوع داده‌ها و آزمون و خطا تعیین می‌شود.

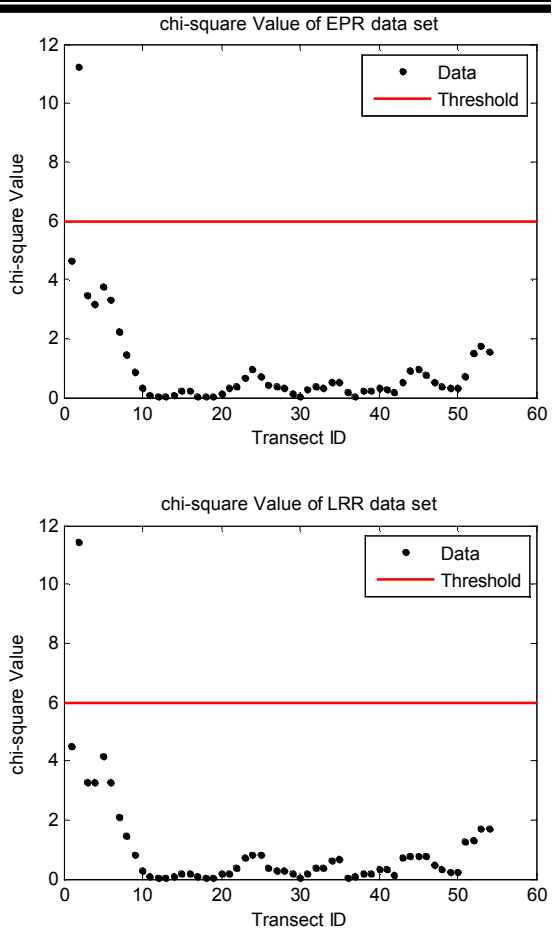
برای شناسایی داده‌های پرت موجود در مجموعه داده‌های EPR و LRR با استفاده از آزمون chi-square score، یک برنامه کامپیوتری با استفاده از نرم‌افزار MATLAB نوشته شده است. اجرای آزمون chi-square score روی مجموعه داده‌ها نیازمند تعیین مقدار پارامتر آستانه توسط کاربر است که در اینجا با توجه به ماهیت داده‌های ورودی و آزمون و خطا، این مقدار برابر ۳ در نظر گرفته شده است. در جدول ۲ حاصل اجرای این روش روی مجموعه داده‌ها ارائه شده است. در این جدول، مقدار و اندیس داده‌های پرت ذکر شده است. همچنین در شکل ۵ یک نمایش گرافیکی از داده‌های انتخاب شده به عنوان کاندیدای پرت نشان داده شده است. داده‌های پرت با یک علامت دایره به دور آنها در شکل مشخص شده‌اند. در شکل ۶ نیز chi-square score محاسبه شده برای تمامی نمونه‌های مجموعه داده‌هاارائه شده است. مقادیری که بالای خط آستانه (خط افقی) قرار گیرند، به عنوان کاندیدای پرت در نظر گرفته می‌شوند.

جدول (۲) داده‌های پرت شناسایی شده در مجموعه داده‌هایEPR و LRR

نمونه	اندیس	مقدار
EPR	۲	-۰/۱۷۵
LRR	۲	-۰/۲۱



شکل (۵) داده‌های پرت شناسایی شده در مجموعه داده‌های EPR و LRR



شکل (۶) مقدار chi-square score محاسبه شده برای مجموعه دادههای EPR و LRR

همانطور که از نتایج جدول ۲ مشخص است در هر دو مجموعه داده، داده با اندیس ۲ به عنوان کاندیدای پرت انتخاب شده است. از شکل ۵ نیز مشخص است که این داده به طرز مشکوکی از سایر اعضا دور افتاده است و این گمان می‌رود که با یک روند متفاوت از سایرین تولید شده است. در این نمونه پس از بررسی‌های انجام شده مشخص شد که علت ایجاد آن، مقطع عرضی نامناسب بوده است. در واقع مقطع ایجاد شده برای این قسمت توسط DSAS، به درستی عمود بر خطوط ساحلی نبوده است. پس از تصحیح مقطع مورد نظر و محاسبه نرخ تغییرات از ابتدا، خطای این داده بر طرف شد. مقدار جدید نرخ تغییر خط ساحلی حول این مقطع به روش EPR برابر ۰/۴۷۵-، و به روش LRR برابر ۰/۵۱- محاسبه شد، که اختلاف این مقادیر با همسایه‌هایشان بسیار ناچیز است.

۵- نتیجه‌گیری

در این تحقیق مبحث شناسایی داده‌های مشکوک به خطا و عوامل ایجاد این نوع از داده‌ها در داده‌های حاصل از کمی‌سازی تغییرات خط ساحلی معرفی شد. آگاهی از داده‌های خطا و عوامل ایجاد آنها سبب می‌شود تا حد امکان از ایجاد مقادیر اشتباه در اندازه‌گیری‌ها جلوگیری شود، که این خود سبب افزایش میزان دقت تجزیه و تحلیل داده‌ها و نتایج حاصل شده از آنها می‌شود.

۶- مراجع

[۱]محمودی، کیومرث؛ سایبانی، مصباح؛ مرادی، عباس؛ "شناسایی خطاهای موجود در برداشت داده‌های آزمایشگاهی، با استفاده از الگوریتم‌های تشخیص داده خطا"، *پانزدهمین کنفرانس دینامیک شاره‌ها FD2013*، بندرعباس، دانشگاه هرمزگان، ۲۷- ۲۹ آذر ۱۳۹۲.

[۲]محمودی، کیومرث؛ کتابداری، محمد جواد؛ سایبانی، مصباح؛ "شناسایی داده‌های مشکوک به خطا، در داده‌های تعیین الگوی جریان"، *پانزدهمین کنفرانس دینامیک شاره‌ها FD2013*، بندرعباس، دانشگاه هرمزگان، ۲۷- ۲۹ آذر ۱۳۹۲.

[3] Barnett, V; Lewis, T; Outliers in Statistical Data, New York, NY, John Wiley and Sons, 3rd edition, 1994.

[4] Johnson, R; Applied Multivariate Statistical Analysis, Prentice Hall, 1992.

[5] Grubbs, F. E; "Procedures for detecting outlying observations in samples", Technometrics, Vol. 11, pp. 1–21, 1969.

[6] Hawkins, D; Identification of Outliers, Chapman and Hall, London, 1980.

[7] Papadimitriou, S; Kitawaga, H; Gibbons, P.G; Faloutsos, C; "LOCI: Fast Outlier Detection Using the Local Correlation Integral", Intel research Laboratory Technical report no. IRP-TR-02-09, 2002.

[8] Billor, N; Hadi, A; Velleman, P; "BACON: Blocked Adaptive Computationally Efficient Outlier Nominators", Computational Statistics and Data Analysis, Vo. 34, pp. 279-298, 2000.

[9] Eskin, E; "Anomaly Detection over Noisy Data using Learned Probability Distributions", Proceedings of the Int. Conf. on Machine Learning, Stanford University, CA, June 2000.

[10] Aggarwal, C. C; Yu, P; "Outlier detection for high dimensional data", In Proceedings of the ACM SIGMOD International Conference on Management of Data, 2001.

[11] Zwass, V., "Management information systems". New York, Wm. C. Brown; 1992.

[12] Breunig, M. M; Kriegel, H.P; Ng, R.T; Sander, J; "LOF: Identifying Density Based Local Outliers", In Proceedings of the ACM SIGMOD Conference, Dallas, TX, May 2000.

[13] Knorr, E. M; Ng, R. T; "Algorithms for Mining Distance-based Outliers in Large Dataset", In Proc. of 24th International Conference on Very Large Data Bases (VLDB'98), New York, NY, pp 392-403, 1998.

[14] Yu, D; Sheikholeslami, G; Zhang, A; "FindOut: Finding Outliers in Very Large Datasets", The Knowledge and Information Systems (KAIS), Vol. 4, No. 4, October 2002.

[15] Eskin, E; Arnold, A; Prerau, M; Portnoy, L; Stolfo, S; "A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data", in Applications of Data Mining in Computer Security, Advances In Information Security, Jajodia, S; Barbara, D; Boston, Ed; Kluwer Academic Publishers, 2002.

[۱۶]چاندرا، آ.م؛ گوش، س.ک؛ ترجمه: علوی پناه، سید کاظم، لدنی، مسلم؛ سنجش از دور و سامانه اطلاعات جغرافیایی، چاپ اول، انتشارات دانشگاه تهران، سال انتشار ۱۳۸۹.

[17] Thieler, E.R;Himmelstoss, E.A;Zichichi, J.L; and Ergul, Ayhan. "Digital Shoreline Analysis System (DSAS) version 4.0 — An ArcGIS extension for calculating shoreline change", U.S. Geological Survey Open-File Report 2008-1278. 2009.

[18] Liu, J. K. "Developing Geographic Information System Applications in Analysis of Responses to Lake Erie Shoreline Changes", M.S. thesis, The Ohio State University, 1998.

[19] Galgano, F; and Douglas, B. "Shoreline Position Prediction: Methods and Errors", Environmental Geosciences, Vol. 7, No. 1. pp. 23-31. 2000.

[20] Hawkins, D; "Identification of Outliers", Chapman and Hall, 1980.