نشریه علمی - پژوهشی علوم و فنون نقشه برداری، دوره دوم، شماره ۴، اردیبهشت ماه ۱۳۹۲

بررسی عملکرد روش‌های کشف مشاهدات اشتباه در سری‌های زمانی

صفورا زمین‌پرداز [*۱]، محمدعلی شریفی [۲]، علیرضا امیری سیمکویی [۳]

[۱]دانشجوی کارشناسی ارشد ژئودزی - گروه مهندسی نقشه‌برداری - پردیس دانشکده‌های فنی - دانشگاه تهران
s.zaminpardaz@ut.ac.ir

[۲]استادیار گروه مهندسی نقشه‌برداری - پردیس دانشکده‌های فنی - دانشگاه تهران
sharifi@ut.ac.ir

[۳]استادیار گروه مهندسی نقشه‌برداری - دانشکده فنی مهندسی - دانشگاه اصفهان
ar.amirisimkooei@surv.ui.ac.ir

(تاریخ دریافت اسفند ۱۳۹۰، تاریخ تصویب آبان ۱۳۹۱)

چکیده

هدف این مقاله، مقایسه‌ی روش‌های مختلف کشف مشاهدات اشتباه، در سری‌های زمانی می‌باشد. به این منظور، سه روش آنالیز موجک، جستجوی باردا و آستانه‌گذاری مورد بررسی قرار می‌گیرند. به منظور مقایسه‌ی عملکرد این سه روش در کشف مشاهدات اشتباه، از سری‌های زمانی شبیه‌سازی شده‌ی ۴ ماهه، بر اساس داده‌های جزر و مدی استفاده گردیده است. زمانی که مدل تابعی مشاهدات معلوم باشد، روش جستجوی باردا، نسبت به دو روش دیگر، عملکرد قوی تری دارد، زیرا بدون وابستگی به نوع اشتباهات، دارای نرخ موفقیت تقریبا ۱۰۰٪ و شکست تقریبا ۰/۶۴٪ می‌باشد. زمانی که مدل تابعی مشاهدات معلوم نباشد، آنالیز موجک نسبت به روش آستانه‌گذاری، عملکرد بهتری دارد.

واژگان کلیدی: کشف مشاهدات اشتباه، آنالیز موجک، جستجوی باردا، آستانه‌گذاری، سری‌های زمانی.

* نویسنده رابط

۱- مقدمه

در بحث آنالیز داده‌ها، چه در کاربردهای ژئودتیکی و چه در کاربردهای عمومی، تعداد زیادی از مشاهدات، ثبت و یا نمونه‌برداری شده که منجر به تشکیل یک مجموعه-داده می‌گردند. یکی از مراحل بسیار مهم به منظور داشتن یک آنالیز منطقی، کشف *اشتباهات*[۱] است. وجود مشاهدات اشتباه در داده‌ها، منجر به بروز خطا در مدل، برآورد اریب پارامترها، پیش‌بینی نادرست و نتایج غلط خواهد شد.

مشاهدات اشتباه، به مشاهداتی گفته می شود که نسبت به مابقی مشاهدات موجود در مجموعه داده‌ها، ناسازگارند [۲]. البته همیشه مقادیر بزرگ در مجموعه-داده‌ها، به معنای مقادیر اشتباه نیست. اگر این مقادیر نسبت به مقادیر اندازه‌گیری‌های همسایه‌ی خود، خیلی بزرگتر باشند، ممکن است نسبت به دیگر مقادیر موجود در کل مجموعه داده‌ها نیز، بزرگتر و ناسازگار بوده و به عنوان مشاهده‌ی اشتباه شناخته می شوند [۱۴].

علل بروز اشتباهات، عوامل متعددی از جمله: خرابی دستگاه مشاهداتی، قرائت اشتباه، خطای محاسباتی و ... می‌تواند باشد. مشاهده‌ی اشتباه باید از مجموعه‌داده‌ها حذف گردد.

به طور کلی، اشتباهات را می‌توان در سه گروه تقسیم‌بندی نمود [۱۴]:

i. اشتباه جمع‌شونده[۲]: اشتباه جمع‌شونده، از نظر ظاهری، به صورت منفرد (تکی) روی سیگنال (سری زمانی) قرار می‌گیرد. این نوع اشتباه، فقط مقدار مشاهده‌ی انجام شده در زمان بروز انحراف (زمان رخداد اشتباه) را تحت تاثیر قرار می‌دهد.

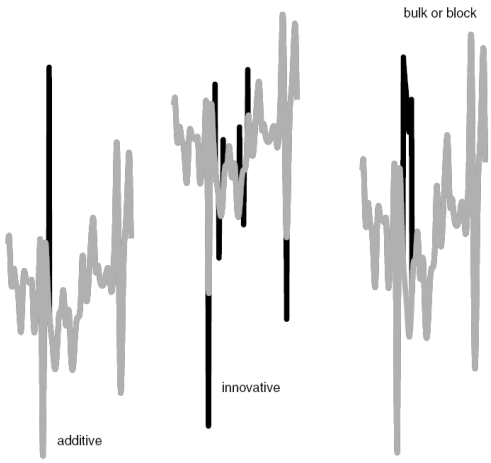
ii. اشتباه نو[۳]: این نوع اشتباه، علاوه بر اینکه مقدار مشاهده‌ی انجام شده در زمان بروز انحراف را تحت تاثیر قرار می‌دهد، مشاهدات بعد از زمان بروز انحراف را نیز، منحرف می‌کند. اشتباه نو، اغلب زمانی رخ می‌دهد که خود سیگنال نیز دارای مقادیر بزرگی است و به همین علت، کشف آن کمی مشکل است.

iii. اشتباه توده‌ای یا بلوکی[۴]: این نوع اشتباه معمولاً به علت نقص موقتی وسیله اندازه‌گیری رخ می-دهد.

در شکل ۱، نمونه‌ای از اشتباه جمع‌شونده، نو و توده-ای، نشان داده شده است. لازم به ذکر است که علاوه بر دسته‌بندی یاد شده، دسته‌بندی‌های دیگری نیز برای اشتباهات وجود دارد.

روش‌های بسیار متعددی برای مقابله با اشتباهات وجود دارد که بسته به نوع و کاربرد داده‌ها، می‌توان از آن‌ها استفاده کرد. بنابراین، آنالیز و کشف اشتباهات، یک مسئله‌ی نسبی است که به ویژگی‌های سیگنال و مدل-های مورد استفاده برای کشف اشتباه وابسته است.

برای کشف بهتر اشتباهات، ترجیح بر این است که نسبت سیگنال به نویز، کاهش یابد. زیرا در این صورت، اشتباهات در داخل مجموعه‌داده‌ها، واضح‌تر دیده می-شوند.



شکل ۱- انواع اشتباهات: اشتباهات جمع‌شونده، نو و توده‌ای با رنگ مشکی روی سیگنال با رنگ خاکستری به ترتیب از چپ به راست

۲- آنالیز موجک[۵]

به منظور کشف اشتباهات، اجزایی از سیگنال که دارای فرکانس بالا می‌باشند، مورد بررسی قرار گرفته تا اپک-هایی که در آن‌ها، تغییرات سیگنال، با دامنه‌ی بالا رخ می‌دهد، یافت شوند. بدین منظور، به یک ابزار آنالیز سیگنال نیازمندیم که سیگنال را هم نسبت به زمان و

[۱] outliers
[۲] additive outlier
[۳] innovative outlier
[۴] bulk or block outlier
[۵] wavelet

هم نسبت به فرکانس، آنالیز کند. چنین ابزاری، توسط آنالیز موجک فراهم می‌شود. در این روش، نیازی به دانستن مدل مشاهدات و حتی تابع توزیع مشاهدات نیست. در واقع، ایده‌ی اصلی آنالیز موجک، کاهش میزان نسبت سیگنال به نویز می‌باشد. کشف مشاهدات اشتباه با استفاده از آنالیز موجک، مسئله‌ای است که در سال-های اخیر بسیار زیاد مورد توجه قرار گرفته است [۴، ۵،۱۱، ۱۴، ۱۹].

در ادامه، ابتدا به توضیح تبدیل موجک و سپس به توضیح چگونگی کشف اشتباه با استفاده از این روش می‌پردازیم.

۲-۱- تبدیل موجک

موجک، یک خانواده از توابع پایه است که برای تقریب بسیاری از توابع و بدست آوردن ویژگی‌هایی نظیر ناپیوستگی و سایر تغییرات ناگهانی، مناسب می‌باشد.

فرض کنید که داده‌های $\underline{Z} = (Z_1, Z_2, ..., Z_n)$، مشاهده شده‌اند که در آن $Z_i = f(t_i)$، $t_i = \frac{i}{n}$، $i = 1, ..., n$ و $n = 2^J$ است. تبدیل گسسته‌ی موجک، برای تجزیه‌ی بردار $\underline{Z}$ به بردارهایی از ضرایب موجک یعنی $D(0), ..., D(J-2), D(J-1)$ و $C(0)$، از تبدیل قائم بهره می‌گیرد. هر مجموعه از ضرایب، شامل 2^j عضو است که در آن، $j = 0, 1, ..., J-1$ می‌باشد. لازم به ذکر است که اگر تعداد داده‌ها، توانی از ۲ نباشد، روش‌هایی برای حل این مشکل ارائه شده است [۶ ، ۱۶].

تعداد موجک‌های مادر مورد استفاده برای تبدیل موجک، بسیار زیاد است. در این تحقیق، استفاده از موجک *هار*[۱]، مدنظر است. ضرایب موجک، با استفاده از الگوریتم‌های فیلترینگ سریع که بر اساس یک جفت فیلتر بالاگذر (***H***) و پایین‌گذر (***G***) هستند، قابل دستیابی می‌باشند. فیلترهای ***H*** و ***G***، به ترتیب به وسیله‌ی ضرایب $h(k)_{k\in Z}$ و $g(k)_{k\in Z}$ که "ضرایب فیلتر" نام دارند، تعریف می‌شوند. این فیلترها، هم بر خود و هم بر یکدیگر عمود هستند. فیلترهای ***H*** و ***G*** به ترتیب، به صورت روابط ۱ و ۲ ارائه می‌گردند:

$$\begin{cases} HZ_k = h(0)Z_k + h(1)Z_{k-1} \\ h(0) = \frac{1}{2}\ ,\ h(1) = -\frac{1}{2} \end{cases} \qquad (۱)$$

$$\begin{cases} GZ_k = g(0)Z_k + g(1)Z_{k-1} \\ g(0) = \frac{1}{2}\ ,\ g(1) = \frac{1}{2} \end{cases} \qquad (۲)$$

اگر داده‌ها، دارای سایز $n = 2^J$ باشند، آنگاه تبدیل گسسته‌ی موجک، نیازمند J سطح تجزیه‌سازی است. در مرحله‌ی اول، داده‌ها به $D(J-1)$ و $C(J-1)$ تجزیه می-شوند که به ترتیب شامل اجزای با فرکانس بالا و فرکانس پایین هستند. در مرحله‌ی دوم، تجزیه روی داده‌های $C(J-1)$ انجام می‌پذیرد و دو مجموعه‌ی جدید $D(J-2)$ و $C(J-2)$ را تولید می‌کند؛ این تجزیه‌سازی تا مرحله‌ی J ادامه می‌یابد، جایی که $C(0)$ دیگر قابل تجزیه شدن نیست. برای آشنایی بیشتر با موجک‌ها و ویژگی‌های آن‌ها، می‌توان به منابعی نظیر [۷، ۱۵] مراجعه نمود.

۲-۲- نحوه‌ی تشخیص اشتباهات با استفاده از آنالیز موجک

این روش در سال ۲۰۱۱ ارائه گردید [۵]. ضرایب موجک، یعنی $D(0), ..., D(J-2), D(J-1)$ ، شامل اجزای با فرکانس بالا هستند به طوری که نسبت به رفتارهای *غیر نرم*[۲] داده‌ها مانند نویز، پرش و *سیگنال‌های نوک تیز* با طول زمانی کوتاه[۳]، حساسیت بالایی دارند [۵، ۱۴]. از آن جایی که اشتباهات در سری‌های زمانی اغلب به صورت برآمدگی[۴] و پرش[۵] در داده‌های مشاهداتی بروز می‌کنند،

[۱] Haar wavelet

[۲] non-smooth
[۳] data spike
[۴] bump
[۵] jump

بنابراین، با استفاده از ضرایب موجک می‌توان به کشف این اشتباهات پرداخت.

تنها فرض این روش، استقلال تقریبی و نرمال بودن ضرایب موجک می‌باشد. در حقیقت، تبدیل موجک، دارای این ویژگی می‌باشد که اگر روی داده‌های وابسته به هم اعمال گردد، ضرایب موجک تولید شده، دارای وابستگی بسیار کمتری نسبت به داده‌های اصلی بوده و این وابستگی با افزایش مراحل تجزیه‌سازی، بیشتر کاهش می‌یابد [۵، ۱۳]. از طرفی، ضرایب موجک بدست آمده از داده‌هایی با توزیع نرمال، خود نیز دارای توزیع نرمال می‌باشند [۱۶]. این روش، با قرار دادن یک حد آستانه برای ضرایب موجک در هر یک از مراحل تجزیه سازی، به کشف اشتباهات می‌پردازد. حدآستانه‌ی مورد استفاده در مرحله‌ی jام عبارتست از:

$\tau = \sigma_j (2\log(n))^{\frac{1}{2}}$ که در آن σ_j، انحراف معیار ضرایب موجک در مرحله‌ی j ام می‌باشد [۹، ۱۹]. انتخاب این حدآستانه، از این موضوع برمی‌آید که برای داده‌های $Z_1, Z_2, ..., Z_n \sim N(0, \sigma_i^2)$، داریم:

$$\lim_{n \to \infty} P\left\{ \max_{1 \le i \le n} \left| Z_i / \sigma_i \right| > \sqrt{2\log(n)} \right\} \to 0 \qquad (۳)$$

رابطه‌ی ۳، هم برای داده‌های مستقل و هم برای داده‌های وابسته، برقرار است. احتمال موجود در این رابطه، دارای کران بالای $(4\pi\log(n))^{-\frac{1}{2}}$ می‌باشد [۱۳].

در این روش، در هر مرحله از تجزیه‌سازی، برای محاسبه‌ی انحراف معیار ضرایب موجک، از میانه استفاده می شود؛ یعنی انحراف معیار در مرحله‌ی j ام، عبارتست از:

$$\sigma_j = AD(D(J\text{-}j)) = \frac{1}{n_j} \sum_{k=1}^{n_j} \left| d_k(j) - M_j \right| \qquad (۴)$$

که در آن، $d_k(j)$، ضریب موجک در مرحله‌ی j ام از تجزیه‌سازی و در محل ***k***، M_j، میانه‌ی ضرایب مرحله‌ی j ام و n_j تعداد ضرایب مرحله‌ی j ام می‌باشد. همچنین ***AD***، بیانگر میانگین قدرمطلق انحراف از میانه می‌باشد.

بعد از محاسبه‌ی انحراف معیار و در نتیجه حد آستانه در هر مرحله، نوبت به قیاس ضرایب موجک و حد آستانه در هر مرحله می‌رسد. آن دسته از ضرایبی که قدر مطلقشان از حد آستانه بیشتر گردد مشخص شده و مشاهدات مربوط به آن‌ها، به عنوان مشاهدات اشتباه شناخته می‌شوند. در این روش، تجزیه‌سازی تا چند مرحله‌ی اول مورد نیاز است. زیرا اشتباهات در همان مراحل اولیه، خود را نشان می‌دهند. لازم به ذکر است که برای کشف پرش‌ها و برآمدگی‌ها، از دو مرحله‌ی آخر تجزیه‌ی موجک نیز استفاده شده است [۱۹].

اگر مدل صحیح داده‌های سری زمانی معلوم باشد، بهتر است که از این اطلاعات برای کشف اشتباهات استفاده شود. به این ترتیب که بعد از برازش مدل به داده‌ها، از بردار باقیمانده‌ها به عنوان سری مشاهداتی استفاده شده و مطابق روند ذکر شده، اشتباهات استخراج گردند.

۳- جستجوی باردا[۱]

ایده‌ی اولیه‌ی این روش، توسط باردا در سال ۱۹۶۸ ارائه گردید [۱]. این روش مبتنی بر سرشکنی کمترین مربعات بوده و به صورت مرحله به مرحله به کشف و شناسایی اشتباهات می‌پردازد. در واقع، در هر مرحله، بزرگترین مشاهده یعنی مشاهده‌ای که بیشترین مقدار قدرمطلق آماره را به خود می‌گیرد، مورد آزمون کشف اشتباه قرار می‌گیرد؛ اگر این مشاهده، در آزمون آماری رد شود، به عنوان مشاهده‌ی اشتباه شناسایی و حذف می‌گردد و روند گفته شده، برای کشف اشتباه بعدی تکرار می‌شود. اگر مشاهده‌ی مذکور، در آزمون آماری پذیرفته گردد، آنگاه به‌عنوان اشتباه در نظر گرفته نشده و روند کشف اشتباه متوقف می‌گردد. زمانی که مدل تابعی مشاهدات، یک مدل خطی باشد، آنگاه اجرای این روش را می‌توان در سه قدم، خلاصه نمود [۱۷]:

i. کشف[۲]: بررسی صحت کلی مدل و مشاهدات از طریق آزمون فاکتور وریانس وزن واحد (در این مرحله، کلیه‌ی مشاهدات به صورت یکجا، تحت فرض صفر H_0، یعنی عدم وجود اشتباه در

[۱] Baarda
[۲] detection

مشاهدات، مورد آزمون قرار می‌گیرند. این آزمون بیانگر این موضوع است که آیا در مجموعه‌داده‌ها، اشتباه وجود دارد یا نه؟)؛

ii. شناسایی[۱]: مشخص کردن خطای مدل که سبب رد شدن فرض صفر H_0 می‌گردد (در این مرحله، تک تک مشاهدات برای تشخیص اشتباهات، مورد آزمون قرار می‌گیرند)؛

iii. سازگاری[۲]: تصحیح و یا حذف مشاهداتی که به عنوان اشتباه تشخیص داده شده‌اند (این مرحله به منظور داشتن یک مجموعه‌داده‌ی منطقی یعنی یک مجموعه داده‌ی بدون اشتباه با داده-های سازگار، انجام می‌پذیرد).

این سه مرحله، می‌توانند تا زمانی که مشاهدات باقیمانده، تحت فرض صفر H_0 پذیرفته شوند، ادامه یافته و یک روند تکراری را برای کشف مشاهدات اشتباه ایجاد کنند [۱۷].

به منظور کشف اشتباهات در یک مجموعه‌داده با استفاده از روش باردا، ابتدا می‌بایست یک سرشکنی کمترین مربعات روی مجموعه‌داده‌ها اعمال گردد. اگر معادلات مدل مشاهدات خطی، به صورت زیر باشد:

$$\underline{y}=A\underline{x}+\underline{e} \qquad (۵)$$

که در آن، $\underline{y}$، بردار m بعدی از مشاهدات، A، ماتریس طرح $m\times n$ بعدی، $\underline{x}$، بردار n بعدی از مجهولات و $\underline{e}$، بردار m بعدی از باقیمانده‌ها است، آنگاه برآورد کمترین مربعات از مجهولات بر اساس رابطه‌ی ۶ قابل حصول است:

$$\hat{\underline{x}}=(A^T Q_y^{-1} A)^{-1} A^T Q_y^{-1} \underline{y} \qquad (۶)$$

در رابطه‌ی ۶، Q_y، ماتریس کوریانس مشاهدات است.

به این ترتیب، بردار باقیمانده‌ی مشاهدات، با استفاده از رابطه‌ی ۷ و رابطه‌ی ۸ قابل برآورد است:

$$\begin{cases} \hat{\underline{e}}=P_A^{\perp}\underline{y} & (۷) \\ P_A^{\perp}=I-A(A^T Q_y^{-1} A)^{-1} A^T Q_y^{-1} & (۸) \end{cases}$$

ماتریس کوریانس بردار باقیمانده‌های برآوردشده نیز از طریق رابطه‌ی ۹، بدست می آید:

$$Q_{\hat{e}}=Q_y-A(A^T Q_y^{-1} A)^{-1} A^T = P_A^{\perp} Q_y \qquad (۹)$$

در روش باردا، آماره‌ای که برای کشف اشتباه در نظر گرفته می‌شود، بر اساس بردار باقیمانده‌هاست. در واقع باقیمانده‌ها نسبت به انحراف معیارشان مقایسه گشته و آماره‌ی w-test را تشکیل می‌دهند [۱۷]:

$$w_i=\frac{\underline{c}_i^T Q_y^{-1}\hat{\underline{e}}}{\sqrt{\underline{c}_i^T Q_y^{-1} Q_{\hat{\underline{e}}} Q_y^{-1}\underline{c}_i}} \qquad (۱۰)$$

که در آن، $\underline{c}_i$ ، یک بردار ستونی است که تمام مؤلفه‌های آن به جز مؤلفه ی i ام که یک است، صفر می‌باشد. اگر ماتریس Q_y، یک ماتریس قطری باشد، آنگاه رابطه‌ی ۱۰ به رابطه‌ی ۱۱ تبدیل می‌شود:

$$w_i=\frac{\hat{e}_i}{\sigma_{\hat{e}_i}} \qquad (۱۱)$$

که در آن داریم:

$$\sigma_{\hat{e}_i}=\sqrt{(Q_{\hat{\underline{e}}})_{ii}} \qquad (۱۲)$$

اگر مشاهدات، دارای تابع توزیع نرمال باشند، آنگاه باقیمانده‌ها نیز، دارای توزیع نرمال هستند. به این ترتیب، در صورتی‌که مقیاس ماتریس کوفاکتور مشاهدات معلوم باشد، آنگاه آماره‌ی موجود در رابطه‌ی ۱۰ دارای توزیع نرمال خواهد بود که در اینصورت فرض صفر H_0

[۱] identification
[۲] adaptation

و فرض مخالف H_a برای این آزمون آماری عبارتند از [۱۸]:

$$H_0: w\sim N(0,1) \quad , \quad H_a: w\sim N(\nabla w,1)$$

که در آن، ∇w، پارامتر عدم مرکزیت می‌باشد. در صورتی که مقیاس ماتریس کوفاکتور مشاهدات معلوم نباشد، آنگاه آماره‌ی مذکور دارای توزیع t استیودنت خواهد بود که در اینصورت فرض صفر H_0 و فرض مخالف H_a برای این آزمون آماری عبارتند از [۱۸]:

$$H_0: w\sim t(m\text{-}n\text{-}1,0) \quad , \quad H_a: w\sim t(m\text{-}n\text{-}1,\nabla w)$$

که در آن نیز، ∇w، پارامتر عدم مرکزیت می‌باشد. پس از آنکه w_i به ازای تمام مشاهدات بدست آمد، فقط بزرگترین مقدار قدرمطلق w_i مورد آزمون آماری قرار می‌گیرد.

برای درک بهتر این روش، یک مجموعه‌ی m تایی از مشاهدات را در نظر بگیرید. بدون آنکه به کلی بودن مسئله، لطمه‌ای وارد شود، فرض کنید که مشاهده‌ی mام به عنوان مشاهده‌ی اشتباه شناخته شده است. این مشاهده بایستی که از لیست مشاهدات و ماتریس‌های مورد استفاده برای سرشکنی، حذف گردد. یعنی بردار باقیمانده‌های جدید، باید با حذف مشاهده‌ی m ام برآورد گردد.

اگر σ_m^2، واریانس، y_m، مقدار مشاهده و $\underline{a}_m$، سطری از A مربوط به اشتباه کشف شده باشد، آنگاه داریم:

$$A=\begin{bmatrix} A_{m-1} \\ \underline{a}_m \end{bmatrix} \qquad (۱۳)$$

$$Q_y=\begin{bmatrix} Q_{y_{m-1}} & 0 \\ 0 & \sigma_m^2 \end{bmatrix} \qquad (۱۴)$$

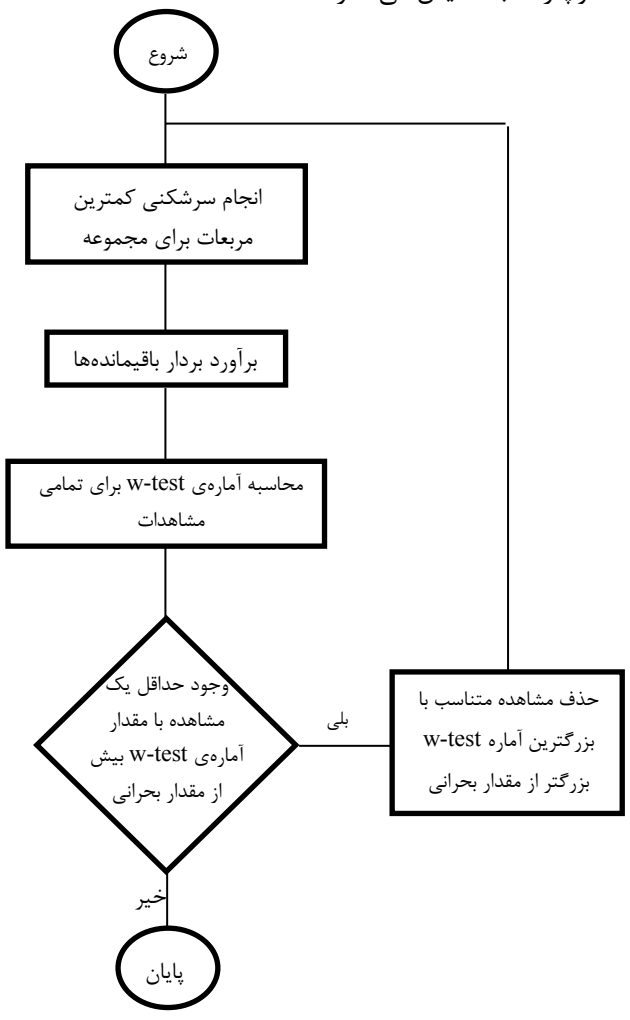
$$\underline{y}=\begin{bmatrix} \underline{y}_{m-1} \\ y_m \end{bmatrix} \qquad (۱۵)$$

با حذف اثر y_m از سرشکنی کمترین مربعات، برآورد جدیدی از بردار مجهولات و بردار باقیمانده‌ها بدست می‌آید که عبارتند از:

$$\hat{\underline{x}}_{m-1}=(A_{m-1}^T Q_{y_{m-1}}^{-1} A_{m-1})^{-1} A_{m-1}^T Q_{y_{m-1}}^{-1} \underline{y}_{m-1} \qquad (۱۶)$$

$$\hat{\underline{e}}_{m-1}=\underline{y}_{m-1}-A_{m-1}\hat{\underline{x}}_{m-1} \qquad (۱۷)$$

شکل ۲، روند اجرای این روش را با استفاده از یک فلوچارت، به نمایش می‌گذارد.



شکل۲- روند اجرایی روش باردا در کشف و حذف مشاهدات اشتباه

این روش، به علت ماهیت تکراری بودن و در واقع به علت تکرار کردن سرشکنی کمترین مربعات در هر مرحله، بار محاسباتی سنگینی را تحمیل می‌کند. البته تکنیک‌هایی وجود دارند که بار محاسباتی این روش را به طرز چشمگیری کاهش می‌دهند.

۴- آستانه‌گذاری

این روش، یک راه حل استاندارد برای کشف اشتباهات است و برای این کار از یک حدآستانه‌ی ماکزیمم استفاده می‌کند [۲]. این حدآستانه، معمولا با ویژگی‌های خود داده‌ها، در ارتباط است.

روش آستانه‌گذاری، برای مشاهدات، یک توزیع معلوم در نظر می‌گیرد که این توزیع، معمولا توزیع نرمال است. بنابراین، اگر یک مجموعه‌داده به صورت $y_i = \{y_1, y_2, ..., y_n\}$ داشته باشیم، آنگاه روش آستانه-گذاری، با فرض توزیع $N(\mu, \sigma^2)$ برای داده‌ها، به صورت زیر قابل تعریف است [۸]:

$$out(\alpha,\mu,\sigma)=\left\{i=1,2,...,n:\left|y_i-\mu\right|>z_{1-\alpha/2}\sigma\right\} \qquad (۱۸)$$

که در آن، $out(\alpha,\mu,\sigma)$، محدوده‌ی اشتباهات، σ,μ، به ترتیب میانگین و انحراف معیار (مجهول) داده‌ها و $z_{1-\alpha/2}$، *مقدار بحرانی*[۱] متناسب با سطح معنی‌دار $1-\alpha/2$ بوده که از جداول توزیع نرمال استاندارد، قابل استخراج می‌باشد. در واقع، این روش، متناسب با مقدار α، مشاهداتی را که از توزیع $N(\mu,\sigma^2)$ انحراف دارند، به عنوان مشاهدات اشتباه شناسایی می‌کند.

لازم به ذکر است که رابطه‌ی ۱۸ نه فقط برای توزیع نرمال که برای هر توزیع متقارن دیگری با تابع چگالی مثبت، قابل تعمیم است [۳]. همچنین می توان در این رابطه، به جای استفاده از میانگین، از میانه و یا هر مقدار پایدار دیگری استفاده نمود [۱۴].

اگر این محاسبات، برای قسمت کوچکتری از داده‌ها محدود در پنجره‌ای با بعد m بازنویسی شود، آنگاه محدوده‌ی اشتباهات برای آن‌ها عبارتست از:

$$out(\alpha,\bar{y}_m,s_m,m)=\left\{i=1,2,...,m:\left|y_i-\bar{y}_m\right|>z_{1-\alpha/2}s_m\right\} \qquad (۱۹)$$

که در آن $y_i = \{y_1, y_2, ..., y_m\}$، مشاهدات بردار $\underline{y} \in R^n$ بوده و $s_m, \bar{y}_m$، به ترتیب برآوردهایی از میانگین و انحراف معیار m داده‌ی مذکور می‌باشند. s_m را می توان از رابطه ۲۰ محاسبه نمود:

$$s_m=\left(\frac{1}{m-1}\sum_{i-1}^{m}(y_i-\bar{y}_m)^2\right)^{\frac{1}{2}} \qquad (۲۰)$$

اگر $y_i \in out(\alpha, \bar{y}_m, s_m, m)$، آنگاه y_i به عنوان مشاهده‌ی اشتباه شناخته می‌شود. بعد از اعمال روش آستانه‌گذاری برای این مجموعه‌ی کوچک از داده‌ها، پنجره‌ی مذکور به اندازه‌ی یک مشاهده، به سمت جلو حرکت کرده و محاسبات برای مجموعه‌ی جدید یعنی $y_i = \{y_2, y_3, ..., y_{m+1}\}$ تکرار می‌شود. حرکت این پنجره و انجام محاسبات، تا پایان مجموعه‌داده‌ها، ادامه می‌یابد.

حسن این روش، بار محاسباتی کم است در حالی که با استفاده از این روش، ممکن است بسیاری از مشاهدات، به غلط، به عنوان مشاهده‌ی اشتباه شناسایی شوند.

۵- آنالیز عددی

در این قسمت، به مقایسه‌ی عملکرد سه روش آستانه-گذاری، آنالیز موجک و جستجوی باردا در کشف مشاهدات اشتباه پرداخته خواهد شد. برای بررسی قدرت روش‌های نامبرده، باید از تعداد واقعی و مکان مشاهدات اشتباه، آگاهی داشته باشیم که در مورد داده-های واقعی، این امکان وجود ندارد. بنابراین، باید از داده‌های شبیه‌سازی شده استفاده نمود. بدین منظور یک سری زمانی از داده‌های جزر و مدی اخذ شده از *تایدگیج*[۲] واقع در ایستگاه *نیولین*[۳] انگلستان، در بازه‌ی زمانی ۱۹۹۳/۰۱/۰۶ ۱۳:۴۵′ – ۱۹۹۳/۰۵/۱۱ ۱۴:۳۰′، شامل ۱۲۰۰۰ داده انتخاب گردید. فاصله زمانی بین

[۱] critical value

[۲] tide gauge

[۳] Newlyn

داده‌ها، ۱۵ دقیقه بوده و به جز چند مورد محدود (۴ مورد)، گپی در آن ها وجود ندارد.

داده‌های شبیه‌سازی شده، طی مراحل زیر بدست آمده‌اند:

ابتدا از طریق سرشکنی کمترین مربعات، بردار مشاهدات سرشکن شده $\underline{\hat{y}}$، برآورد گردید؛ بدین منظور می‌بایست مدلی به مشاهدات، برازش داده می‌شد. با استفاده از ۱۴۵ فرکانس اصلی شناخته شده برای جزر و مد و با استفاده از توابع هارمونیک، مدل مشاهدات عبارتست از [۱۰]:

$$E(\underline{y}(t_i)) = y_0 + rt_i + \sum_{k=1}^{145} a_k \cos\omega_k t_i + b_k \sin\omega_k t_i \quad (۲۱)$$
$$i=1,2,\dots,n$$

در رابطه‌ی ۲۱، t_i از سری زمانی، $\underline{y}(t_i)$، مشاهده‌ی لحظه‌ی n، تعداد مشاهدات سری زمانی، ω_k، فرکانس اصلی k ام، a_k, b_k، ضرایب توابع هارمونیک، r, y_0، به ترتیب، محل تقاطع و شیب ترند خطی می‌باشند.

مطابق با رابطه‌ی ۲۱، ماتریس طرح A را تشکیل داده و سرشکنی را انجام دادیم:

$$\underline{\hat{x}}=(A^T A)^{-1} A^T \underline{y} \quad (۲۲)$$

$\underline{\hat{x}}$، بردار مجهولات می‌باشد که شامل ضرایب توابع هارمونیک، a_k, b_k و محل تقاطع و شیب، r, y_0، است. حال نوبت به محاسبه ی $\underline{\hat{y}}$ می‌رسد:

$$\underline{\hat{y}}=A\underline{\hat{x}} \quad (۲۳)$$

سپس یک نویز سفید با توزیع نرمال $N(0,0.05)$ ساخته شد و به $\underline{\hat{y}}$ اضافه گردید:

$$\underline{y'} = \underline{\hat{y}} + WN \quad (۲۴)$$

در نهایت، اشتباهات نیز به داده های $\underline{y'}$ اضافه گردید.

برای یک مقایسه‌ی همه جانبه بین سه روش نامبرده چه از نظر حساسیت آن‌ها به اندازه اشتباهات و چه از نظر حساسیت آن‌ها به نوع اشتباهات، ۱۰ دسته‌ی مختلف از اشتباهات به شرح زیر ساخته شد که هر دسته به طور مجزا به داده‌های موجود در بردار $\underline{y'}$، اعمال گردیده و بررسی شدند:

۱- ۱۰۰ عدد اشتباه جمع‌شونده که به صورت کاملا تصادفی توزیع شده‌اند و اندازه‌ی آن‌ها در بازه‌ی [۳۰ ۲۰] سانتی‌متر قرار می‌گیرد؛
۲- ۸۰ عدد اشتباه جمع‌شونده که به صورت کاملا تصادفی توزیع شده‌اند و یک اشتباه توده‌ای با طول ۵ ساعت (۲۰ مشاهده) که اندازه‌ی آن‌ها در بازه‌ی [۳۰ ۲۰] سانتی‌متر قرار می‌گیرد؛
۳- ۱۰۰ عدد اشتباه جمع‌شونده که به صورت کاملا تصادفی توزیع شده‌اند و اندازه‌ی آن‌ها در بازه‌ی [۵۰ ۴۰] سانتی‌متر قرار می‌گیرد؛
۴- ۸۰ عدد اشتباه جمع‌شونده که به صورت کاملا تصادفی توزیع شده‌اند و یک اشتباه توده‌ای با طول ۵ ساعت (۲۰ مشاهده) که اندازه‌ی آن‌ها در بازه‌ی [۵۰ ۴۰] سانتی‌متر قرار می‌گیرد؛
۵- ۱۰۰ عدد اشتباه جمع‌شونده که به صورت کاملا تصادفی توزیع شده‌اند و اندازه‌ی آن‌ها در بازه‌ی [۷۰ ۶۰] سانتی‌متر قرار می‌گیرد؛
۶- ۸۰ عدد اشتباه جمع‌شونده که به صورت کاملا تصادفی توزیع شده‌اند و یک اشتباه توده‌ای با طول ۵ ساعت (۲۰ مشاهده) که اندازه‌ی آن‌ها در بازه‌ی [۷۰ ۶۰] سانتی‌متر قرار می‌گیرد؛
۷- ۱۰۰ عدد اشتباه جمع‌شونده که به صورت کاملا تصادفی توزیع شده‌اند و اندازه‌ی آن‌ها در بازه‌ی [۹۰ ۸۰] سانتی‌متر قرار می‌گیرد؛
۸- ۸۰ عدد اشتباه جمع‌شونده که به صورت کاملا تصادفی توزیع شده‌اند و یک اشتباه توده‌ای با طول ۵ ساعت (۲۰ مشاهده) که اندازه‌ی آن‌ها در بازه‌ی [۹۰ ۸۰] سانتی‌متر قرار می‌گیرد.
۹- ۱۰۰ عدد اشتباه جمع‌شونده که به صورت کاملا تصادفی توزیع شده‌اند و اندازه‌ی آن‌ها در بازه‌ی [۱۰۰ ۹۰] سانتی‌متر قرار می‌گیرد؛
۱۰- ۸۰ عدد اشتباه جمع‌شونده که به صورت کاملا تصادفی توزیع شده‌اند و یک اشتباه توده‌ای با

طول ۵ ساعت (۲۰ مشاهده) که اندازه‌ی آن‌ها در بازه‌ی [۱۰۰ ۹۰] سانتی‌متر قرار می‌گیرد.

علت انتخاب این محدوده از اشتباهات این است که بعد از برازش مدل ذکر شده به مشاهدات مورد استفاده، بردار باقیمانده‌ها در بازه ۳۰- تا ۴۰ سانتی‌متر قرار می‌گرفتند. البته انتخاب محدوده های ۴۰ تا ۵۰ سانتی‌متر، ۶۰ تا ۷۰ سانتی‌متر،۸۰ تا ۹۰ سانتی‌متر و ۹۰ تا ۱۰۰ سانتی‌متر به منظور بررسی رفتار روش‌های کشف مشاهدات اشتباه به ازای افزایش بزرگی اشتباهات بوده است.

شبیه‌سازی داده‌ها، به ازای هر یک از ۱۰دسته اشتباهات، ۵۰ بار صورت گرفت. در ادامه، نتایج حاصل از کشف این اشتباهات، با استفاده از سه روش مذکور، توضیح داده می‌شوند. به منظور مقایسه‌ی عملکرد سه روش، از دو معیار *نرخ موفقیت*[1] و *نرخ شکست*[2] استفاده شده است. نرخ موفقیت عبارتست از نسبت تعداد مشاهداتی که به درستی به عنوان مشاهدات اشتباه کشف شده‌اند، به تعداد کل مشاهدات اشتباه و نرخ شکست عبارتست از نسبت تعداد مشاهداتی که به غلط به عنوان مشاهدات اشتباه کشف شده‌اند، به تعداد کل مشاهدات.

در تمامی دسته‌ها، روش آستانه‌گذاری ارائه‌شده در رابطه ۱۹ با ضریب معنی‌دار $\alpha = ۰/۰۵$ و طول پنجره‌ی $m=۸۰$ به‌کار رفته است؛ طول پنجره با سعی و خطا بدست آمده است. همچنین در روش جستجوی باردا، ضریب معنی‌دار $\alpha = ۰/۰۰۵$ مورد استفاده قرار گرفته است. در روش موجک، به‌منظور به‌کارگیری تمام قدرت موجک، تجزیه‌سازی تا مرحله‌ی آخر صورت گرفته است. در بکارگیری روش باردا به‌علت نامعلوم بودن مقیاس ماتریس کوفاکتور مشاهدات، آماره w-test با توزیع $\boldsymbol{t}$ استیودنت مورد آزمون قرار گرفته است. با توجه به اینکه مدل تابعی مشاهدات در دست است، روش‌های موجک و آستانه‌گذاری، علاوه‌بر بررسی داده‌ها، امکان بررسی بردار باقیمانده‌ها (بدست‌آمده از برازش مدل به بردار مشاهدات) به منظور کشف مشاهدات اشتباه را نیز دارند. بنابراین، در هر یک از دسته‌ها، روش‌های موجک و آستانه‌گذاری، هم روی بردار داده‌ها و هم روی بردار باقیمانده‌ها اعمال شده‌اند. نتایج آنالیز عددی هریک از ۱۰ دسته‌ی بالا، در ۱۰جدول، گردآوری شده‌اند.

لازم به ذکر است که در بکارگیری تمامی روش‌های مورد استفاده برای کشف مشاهدات اشتباه در این مقاله، اساس سعی و خطا و همچنین انتخاب پارامترهای موثر، اخذ بهترین نتایج از این روش‌ها بوده است. به عبارت دیگر، از آن‌جایی که داده‌های مورد استفاده، داده‌های شبیه‌سازی شده می‌باشند و مکان بروز اشتباهات مشخص است، پارامترهای روش‌های آستانه‌گذاری، موجک و جستجوی باردا به نحوی انتخاب گردیدند که بتوان از آن‌ها بهترین نتایج را گرفت و به این ترتیب بهترین نتایج آن‌ها را با هم مقایسه کرد. همچنین، به این نکته باید توجه نمود که روش‌های باردا و آستانه گذاری به مقیاس کوفاکتور مشاهدات حساسند و این مقیاس می‌تواند عملکرد این روش‌ها را در کشف مشاهدات اشتباه، تحت تاثیر قرار دهد که بررسی این موضوع از حوصله این بحث خارج است.

جدول ۱ شامل نرخ موفقیت متوسط و نرخ شکست متوسط روش‌های آستانه‌گذاری، موجک و جستجوی باردا، در صورت وجود ۱۰۰ عدد اشتباه جمع‌شونده با بزرگی ۲۰ تا ۳۰ سانتی‌متر و با توزیع تصادفی می‌باشد. جدول ۲ شامل نرخ موفقیت متوسط و نرخ شکست متوسط روش‌های آستانه‌گذاری، موجک و جستجوی باردا، در صورت وجود ۸۰ عدد اشتباه جمع‌شونده و ۲۰ عدد اشتباه توده‌ای با بزرگی ۲۰ تا ۳۰ سانتی‌متر و با توزیع تصادفی می‌باشد. جدول ۳ شامل نرخ موفقیت متوسط و نرخ شکست متوسط روش‌های آستانه‌گذاری، موجک و جستجوی باردا، در صورت وجود ۱۰۰ عدد اشتباه جمع‌شونده با بزرگی ۴۰ تا ۵۰ سانتی‌متر و با توزیع تصادفی می‌باشد. جدول ۴ شامل نرخ موفقیت متوسط و نرخ شکست متوسط روش‌های آستانه‌گذاری، موجک و جستجوی باردا، در صورت وجود ۸۰ عدد اشتباه جمع‌شونده و ۲۰ عدد اشتباه توده‌ای با بزرگی ۴۰ تا ۵۰ سانتی‌متر و با توزیع تصادفی می‌باشد. جدول ۵ شامل نرخ موفقیت متوسط و نرخ شکست متوسط روش‌های آستانه‌گذاری، موجک و جستجوی باردا، در صورت وجود ۱۰۰ عدد اشتباه جمع‌شونده با بزرگی ۶۰ تا ۷۰ سانتی‌متر و با توزیع تصادفی می‌باشد. جدول ۶ شامل

[1] outlier rate of success (ORS)
[2] outlier rate of failure (ORF)

نرخ موفقیت متوسط و نرخ شکست متوسط روش‌های آستانه‌گذاری، موجک و جستجوی باردا، در صورت وجود ۸۰ عدد اشتباه جمع‌شونده و ۲۰ عدد اشتباه توده‌ای با بزرگی ۶۰ تا ۷۰ سانتی‌متر و با توزیع تصادفی می‌باشد. جدول ۷ شامل نرخ موفقیت متوسط و نرخ شکست متوسط روش‌های آستانه‌گذاری، موجک و جستجوی باردا، در صورت وجود ۱۰۰ عدد اشتباه جمع‌شونده با بزرگی ۸۰ تا ۹۰ سانتی‌متر و با توزیع تصادفی می‌باشد. جدول ۸ شامل نرخ موفقیت متوسط و نرخ شکست متوسط روش‌های آستانه‌گذاری، موجک و جستجوی باردا، در صورت وجود ۸۰ عدد اشتباه جمع‌شونده و ۲۰ عدد اشتباه توده‌ای با بزرگی ۸۰ تا ۹۰ سانتی‌متر و با توزیع تصادفی می‌باشد. جدول ۹ شامل نرخ موفقیت متوسط و نرخ شکست متوسط روش‌های آستانه‌گذاری، موجک و جستجوی باردا، با بزرگی ۹۰ تا ۱۰۰ سانتی‌متر و با توزیع تصادفی می‌باشد. جدول ۱۰ شامل نرخ موفقیت متوسط و نرخ شکست متوسط روش‌های آستانه‌گذاری، موجک و جستجوی باردا، در صورت وجود ۸۰ عدد اشتباه جمع‌شونده و ۲۰ عدد اشتباه توده‌ای با بزرگی ۹۰ تا ۱۰۰ سانتی‌متر و با توزیع تصادفی می‌باشد.

جدول ۱- نتایج روش‌های کشف مشاهدات اشتباه، در صورت وجود ۱۰۰ عدد اشتباه جمع‌شونده با بزرگی ۲۰ تا ۳۰ سانتی‌متر و با توزیع تصادفی

نرخ شکست	نرخ موفقیت	روش
٪۰/۰۲	٪۵/۹۶	آستانه گذاری (با بردار داده ها)
٪۳۴/۸۹	٪۱۰۰	آستانه گذاری (با بردار باقیمانده ها)
٪۰/۲۲	٪۸/۵۰	موجک (با بردار داده ها)
٪۱/۳۵	٪۹۴/۸۲	موجک (با بردار باقیمانده ها)
٪۰/۶۴	٪۹۸/۳۲	جستجوی باردا

جدول ۲- نتایج روش‌های کشف مشاهدات اشتباه، در صورت وجود ۸۰ عدد اشتباه جمع‌شونده و ۲۰ عدد اشتباه توده ای با بزرگی ۲۰ تا ۳۰ سانتی‌متر و با توزیع تصادفی

نرخ شکست	نرخ موفقیت	روش
٪۰/۰۲	٪۴/۹۸	آستانه گذاری (با بردار داده ها)
٪۳۴/۵۷	٪۱۰۰	آستانه گذاری (با بردار باقیمانده ها)
٪۰/۲۱	٪۶/۶۶	موجک (با بردار داده ها)
٪۱/۴۶	٪۷۹/۲۸	موجک (با بردار باقیمانده ها)
٪۰/۶۴	٪۹۸/۲۴	جستجوی باردا

جدول ۳- نتایج روش‌های کشف مشاهدات اشتباه، در صورت وجود ۱۰۰ عدد اشتباه جمع‌شونده با بزرگی ۴۰ تا ۵۰ سانتی‌متر و با توزیع تصادفی

نرخ شکست	نرخ موفقیت	روش
٪۰/۰۲	٪۱۵/۶۲	آستانه گذاری (با بردار داده ها)
٪۳۴/۵۷	٪۱۰۰	آستانه گذاری (با بردار باقیمانده ها)
٪۰/۲۹	٪۲۵/۵۰	موجک (با بردار داده ها)
٪۰/۸۱	٪۹۹/۵۰	موجک (با بردار باقیمانده ها)
٪۰/۶۴	٪۱۰۰	جستجوی باردا

جدول ۴- نتایج روش‌های کشف مشاهدات اشتباه، در صورت وجود ۸۰ عدد اشتباه جمع‌شونده و ۲۰ عدد اشتباه توده ای با بزرگی ۴۰ تا ۵۰ سانتی‌متر و با توزیع تصادفی

نرخ شکست	نرخ موفقیت	روش
٪۰/۰۴	٪۱۳/۴۶	آستانه گذاری (با بردار داده ها)
٪۳۴/۰۲	٪۱۰۰	آستانه گذاری (با بردار باقیمانده ها)
٪۰/۲۷	٪۲۱/۲۲	موجک (با بردار داده ها)
٪۰/۹۹	٪۸۲/۹۲	موجک (با بردار باقیمانده ها)
٪۰/۶۴	٪۱۰۰	جستجوی باردا

جدول ۵- نتایج روش‌های کشف مشاهدات اشتباه، در صورت وجود ۱۰۰ عدد اشتباه جمع‌شونده با بزرگی ۶۰ تا ۷۰ سانتی‌متر و با توزیع تصادفی

نرخ شکست	نرخ موفقیت	روش
٪۰/۰۲	٪۲۱/۹۴	آستانه گذاری (با بردار داده ها)
٪۳۳/۷۲	٪۱۰۰	آستانه گذاری (با بردار باقیمانده ها)
٪۰/۳۸	٪۴۴/۲۶	موجک

نرخ شکست	نرخ موفقیت	روش
		(با بردار داده ها)
٪۰/۴۷	٪۹۹/۶۶	موجک
		(با بردار باقیمانده ها)
٪۰/۶۴	٪۱۰۰	جستجوی باردا

جدول ۶- نتایج روش‌های کشف مشاهدات اشتباه، در صورت وجود ۸۰ عدد اشتباه جمع‌شونده و ۲۰ عدد اشتباه توده ای با بزرگی ۶۰ تا ۷۰ سانتی‌متر و با توزیع تصادفی

نرخ شکست	نرخ موفقیت	روش
٪۰/۰۷	٪۱۸/۸۴	آستانه گذاری
		(با بردار داده ها)
٪۳۳/۸۱	٪۱۰۰	آستانه گذاری
		(با بردار باقیمانده ها)
٪۰/۳۴	٪۳۶/۲۰	موجک
		(با بردار داده ها)
٪۰/۶۵	٪۸۲/۹۰	موجک
		(با بردار باقیمانده ها)
٪۰/۶۴	٪۱۰۰	جستجوی باردا

جدول ۷- نتایج روش‌های کشف مشاهدات اشتباه، در صورت وجود ۱۰۰ عدد اشتباه جمع‌شونده با بزرگی ۸۰ تا ۹۰ سانتی‌متر و با توزیع تصادفی

نرخ شکست	نرخ موفقیت	روش
٪۰/۰۲	٪۲۶/۷۴	آستانه گذاری
		(با بردار داده ها)
٪۳۳/۷۳	٪۱۰۰	آستانه گذاری
		(با بردار باقیمانده ها)
٪۰/۴۵	٪۵۵/۷۰	موجک
		(با بردار داده ها)
٪۰/۳۰	٪۹۹/۵۰	موجک
		(با بردار باقیمانده ها)
٪۰/۶۴	٪۱۰۰	جستجوی باردا

جدول ۸- نتایج روشهای کشف مشاهدات اشتباه، در صورت وجود ۸۰ عدد اشتباه جمع‌شونده و ۲۰ عدد اشتباه توده ای با بزرگی ۸۰ تا ۹۰ سانتی‌متر و با توزیع تصادفی

نرخ شکست	نرخ موفقیت	روش
٪۰/۰۶	٪۲۵/۸۰	آستانه گذاری
		(با بردار داده ها)
٪۳۳/۸۶	٪۱۰۰	آستانه گذاری
		(با بردار باقیمانده ها)
٪۰/۳۹	٪۴۵/۸۶	موجک
		(با بردار داده ها)
٪۰/۴۵	٪۸۲/۶۸	موجک

نرخ شکست	نرخ موفقیت	روش
		(با بردار باقیمانده ها)
٪۰/۶۴	٪۱۰۰	جستجوی باردا

جدول ۹- نتایج روش‌های کشف مشاهدات اشتباه، در صورت وجود ۱۰۰ عدد اشتباه جمع‌شونده با بزرگی ۹۰ تا ۱۰۰ سانتی‌متر و با توزیع تصادفی

نرخ شکست	نرخ موفقیت	روش
٪۰/۰۲	٪۲۹/۹۸	آستانه گذاری
		(با بردار داده ها)
٪۳۳/۱۲	٪۱۰۰	آستانه گذاری
		(با بردار باقیمانده ها)
٪۰/۴۴	٪۵۹/۰۸	موجک
		(با بردار داده ها)
٪۰/۲۵	٪۹۹/۶۴	موجک
		(با بردار باقیمانده ها)
٪۰/۶۴	٪۱۰۰	جستجوی باردا

جدول ۱۰- نتایج روش‌های کشف مشاهدات اشتباه، در صورت وجود ۸۰ عدد اشتباه جمع‌شونده و ۲۰ عدد اشتباه توده ای با بزرگی ۹۰ تا ۱۰۰ سانتی‌متر و با توزیع تصادفی

نرخ شکست	نرخ موفقیت	روش
٪۰/۰۹	٪۲۷/۲۳	آستانه گذاری
		(با بردار داده ها)
٪۳۳/۸۶	٪۱۰۰	آستانه گذاری
		(با بردار باقیمانده ها)
٪۰/۳۷	٪۴۸/۴۴	موجک
		(با بردار داده ها)
٪۰/۳۹	٪۸۲/۷۲	موجک
		(با بردار باقیمانده ها)
٪۰/۶۴	٪۱۰۰	جستجوی باردا

در این جا، با توجه به جداول تهیه شده، به چند نوع مقایسه می‌پردازیم.

۵-۱- مقایسه‌ی نتایج آنالیز بردار داده‌ها، در حضور اشتباه جمع‌شونده

با استناد به جداول ۱، ۳، ۵، ۷ و ۹ می‌توان گفت که در هریک از دسته اشتباهات بالا، نرخ موفقیت روش جستجوی باردا، نسبت به دو روش دیگر، بسیار چشمگیرتر است. همچنین، آنالیز موجک از نرخ موفقیت

بیشتری نسبت به روش آستانه‌گذاری برخوردار است. البته نرخ شکست در روش جستجوی باردا بیش از دو روش دیگر و در روش موجک بیش از روش آستانه‌گذاری می‌باشد که البته این اختلاف در مقابل اختلاف موجود در نرخ موفقیت، بسیار ناچیز است.

با افزایش بزرگی اشتباهات، نرخ شکست و موفقیت روش باردا، تقریبا ثابت مانده است. این در حالی است که، نرخ موفقیت روش موجک و آستانه‌گذاری، با افزایش بزرگی اشتباهات، افزایش می‌یابد. پس زمانی که بزرگی اشتباهات، کاهش یابد، آنالیز موجک و آستانه‌گذاری، موفقیت کمتری در یافتن اشتباهات دارند.

در آنالیز موجک، بیشتر اشتباهات، در مرحله‌ی اول تجزیه‌سازی، یافت می‌شد.

۵-۲- مقایسه‌ی نتایج آنالیز بردار باقیمانده‌ها، در حضور اشتباه جمع‌شونده

با استناد به جداول ۱، ۳، ۵، ۷ و ۹ می‌توان گفت که آنالیز بردار باقیمانده‌ها به جای بردار داده‌ها، نرخ موفقیت آنالیز موجک و آستانه‌گذاری را به طرز چشمگیری افزایش می‌دهد. این مسئله ناشی از این است که در حالت آنالیز بردار باقیمانده‌ها، نسبت سیگنال به نویز تا حدود زیادی کاهش یافته و این، امر بسیار مهمی در موفقیت بیشتر در کشف اشتباهات است. همچنین نرخ موفقیت این دو روش، به ازای افزایش بزرگی اشتباهات، همانند روش جستجوی باردا در قسمت قبل، تقریبا ثابت مانده است.

نرخ شکست روش آستانه‌گذاری، نسبت به حالت قبل، افزایش چشمگیری داشته است به گونه‌ای که، با استفاده از این روش، بخش زیادی از اطلاعات موجود در سری زمانی، از دست می‌رود. این افزایش را می‌توان ناشی از طول پنجره و آستانه مورد استفاده دانست؛ زیرا این حد-آستانه و طول پنجره به گونه‌ای انتخاب گردید که روش آستانه‌گذاری در حالت آنالیز بردار داده‌ها بهترین جواب را بدهد و بعد از آن برای حفظ شرایط یکسان به منظور مقایسه، برای آنالیز بردار باقیمانده‌ها نیز مورد استفاده قرار گرفت. این در حالی است که نرخ شکست آنالیز موجک، تفاوت چندانی نکرده و همچنین با افزایش بزرگی اشتباهات، کاهش می‌یابد.

در این جا نیز، در آنالیز موجک، بیشتر اشتباهات، در مراحل تجزیه‌سازی اولیه، ، یافت می‌شد. البته زمانی که بزرگی اشتباهات، کاهش می‌یافت، نیاز به تجزیه‌سازی در مراحل بالاتر، افزایش می‌یافت.

۵-۳- مقایسه‌ی نتایج آنالیز بردار داده‌ها، در حضور اشتباه جمع‌شونده و اشتباه توده‌ای

با استناد به جداول ۲، ۴، ۶، ۸ و ۱۰ می‌توان گفت که زمانی که اشتباهات توده‌ای و جمع‌شونده هر دو با هم حضور دارند، رفتار روش جستجوی باردا، تفاوتی با حالتی که فقط اشتباه جمع‌شونده حضور دارد، ندارد. یعنی در هر یک از دسته اشتباهات، نرخ شکست و موفقیت روش جستجوی باردا، در هر دو حالت، با تقریب بسیار خوبی، یکسان است. در واقع این روش، تمامی مشاهداتی را که از مدل ایده‌آل فاصله دارند را به عنوان مشاهده اشتباه در نظر می‌گیرد؛ بنابر این نوع اشتباه نباید تاثیر خاصی در قدرت این روش داشته باشد.

روش آستانه‌گذاری، در تشخیص اشتباهات توده‌ای، تقریبا ناموفق است هرچند که با افزایش بزرگی اشتباهات، موفقیت روش در تشخیص اشتباهات توده‌ای، کمی افزایش می‌یابد. همچنین نرخ موفقیت و شکست این روش، نسبت به حالتی که فقط اشتباه جمع‌شونده وجود دارد، به ترتیب، کاهش و افزایش می‌یابد. روش موجک، تقریبا در همه ی نمونه‌های در نظر گرفته شده، در همه‌ی سطوح بزرگی اشتباهات، فقط قادر به تشخیص ابتدا و انتهای اشتباه توده‌ای می‌باشد. البته زمانی که بزرگی اشتباه خیلی کم باشد (مانند دسته ی اول)، قادر به تشخیص اشتباه توده ای نمی باشد. همچنین نرخ موفقیت و شکست این روش، نسبت به حالتی که فقط اشتباه جمع‌شونده وجود دارد، کاهش می یابد. در روش موجک، به منظور کشف اشتباهات توده‌ای، به مراحل تجزیه سازی بالاتری نیاز بود. این مسئله توسط منبع [۵] نیز بیان شده است.

لازم به ذکر است که تمام نتایج اخذ شده در حالت حضور اشتباه جمع‌شونده، در این جا نیز برقرار است.

۵-۴- مقایسه‌ی نتایج آنالیز بردار باقیمانده‌ها، در حضور اشتباه جمع‌شونده و اشتباه توده‌ای

با استناد به جداول ۲، ۴، ۶، ۸ و ۱۰ می‌توان گفت که در این حالت، روش آستانه‌گذاری، اشتباه توده‌ای را به طور کامل تشخیص می‌دهد. روش موجک تقریبا در اکثر نمونه‌های در نظر گرفته شده، در همه‌ی سطوح بزرگی اشتباهات، فقط قادر به تشخیص ابتدا و انتهای اشتباه توده‌ای می‌باشد. همچنین نرخ موفقیت و شکست این روش، نسبت به حالتی که فقط اشتباه جمع‌شونده وجود دارد، به ترتیب کاهش و افزایش می‌یابد. در این جا نیز، در روش موجک، به منظور کشف اشتباهات توده‌ای، به مراحل تجزیه‌سازی بالاتری، نیاز بود.

لازم به ذکر است که تمام نتایج اخذ شده در حالت حضور اشتباه جمع‌شونده، در این جا نیز برقرار است.

۶- نتیجه گیری

در بحث آنالیز داده‌ها، یکی از مراحل بسیار مهم به منظور داشتن یک آنالیز منطقی، کشف اشتباهات است. وجود مشاهدات اشتباه در داده‌ها، منجر به بروز خطا در مدل، برآورد اریب پارامترها، پیش‌بینی نادرست و نتایج غلط خواهد شد. در این مقاله، سه روش از روش‌های کشف مشاهدات اشتباه، یعنی، آنالیز موجک، جستجوی باردا و آستانه‌گذاری مورد بررسی قرار گرفتند و رفتار این روش‌ها در مقابل اشتباهات جمع‌شونده و توده‌ای با بزرگی‌های گوناگون، در یک سری زمانی جزر و مدی شبیه‌سازی شده، مقایسه گردید.

نتیجه‌ای که از قیاس سه روش مذکور یافت شد این بود که اگر مدل تابعی مشاهدات، معلوم باشد، آنگاه روش جستجوی باردا، روش پیشنهادی مناسب می‌باشد. زیرا این روش فارغ از توده‌ای و یا جمع‌شونده بودن اشتباهات دارای نرخ موفقیت و شکست خوب و تقریبا پایداری می‌باشد. اگر مدل مشاهدات معلوم نباشد، استفاده از موجک نسبت به روش آستانه‌گذاری، ترجیح داده می‌شود. زیرا روش موجک دارای نرخ موفقیت بالاتری می‌باشد. از طرفی در روش آستانه‌گذاری، باتغییر حد آستانه‌ی مورد استفاده برای تشخیص اشتباهات، نرخ شکست به طرز چشمگیری تغییر می‌یابد.

سپاسگزاری

نویسندگان این مقاله، نهایت سپاس‌گذاری را از نظرات ارزشمند و سازنده‌ی داوران محترم دارند که باعث ارتقا سطح این مقاله و بهبود کیفیت ارائه‌ی آن گردید.

مراجع

[1] Baarda W. (1968) "A testing procedure for use in geodetic networks". Publ 2(5), Netherlands Geodetic Commission, Delft.

[2] Barnett V., Lewis T. (1994) "Outliers in statistical data", 3rd edn. John Wiley, Chichester.

[3] Ben-Gal I., Maimon O., Rockach L. (2005) Data Mining and Knowledge Discovery Handbook A Complete Guide for Practitioners and Researchers. Kluwer Academic Publishers, ISBN: 0-387-24435-2.

[4] Bruce A. G., Donoho L. G., Gao H. Y., Martin R. D. (1994) Denoising and robust nonlinear wavelet analysis. SPIE Proceedings, Wavelet Applications, vol 2242, Harald H. San (ed), The Internat. Society for Optical Engineering (SPIE) Orlando, FL pp 325–336.

[5] Bilen C., Huzurbazar S. (2002) Wavelet-based detection of outliers in time series. Journal of computational and graphical statistics, Vol. 11, No. 2, pp. 311-327.

[6] Cohen A., Daubechies I., Jawerth B., Vial P. (1993) Wavelets on the interval and fast wavelet transforms. Applied and Computational Harmonic Analysis 1: 54-81

[7] Daubechies I. (1992) Ten lectures on wavelets. Society for Industrial and Applied Mathematics, Philadelphia, PA.

[8] Davies L., Gather U. (1993) The identification of multiple outliers. Journal of the American Statistical Association, 88(423), 782-792

[9] Donoho D. L., Johnstone I. M. (1995) Ideal spatial adaptation via wavelet shrinkage. Biometrika 81:425-455.

[10] Foreman M. G. G., Neufeld E. M. (1991) Harmonic tidal analyses of long time series. Int. Hydrogr. Rev., LXVIII, 85-108.

[11] Götzelmann M., Keller W., Reubelt T. (2006) Gross error compensation for gravity field analysis based on kinematic orbit data. Journal of Geodesy, 80: 184–198, doi: 10.1007/s00190-006-0061-9.

[12] Hawkins D. (1980) Identification of Outliers. Chapman and Hall.

[13] Johnstone I. M., Silverman B. W. (1997) Wavelet threshold estimators for data with correlated noise. Journal of Royal Statistical Society , series B, 59, 319-351.

[14] Kern M., Preimesberger T., Allesch M., Pail R., Bouman J., Koop R. (2005) Outlier detection algorithms and their performance in GOCE gravity field processing. Journal of Geodesy, 78: 509–519, doi: 10.1007/s00190-004-0419-9.

[15] Mallat S. (1999) A wavelet tour of signal processing, 2nd edn. Academic Press, New York pp 167–174.

[16] Ogden R. T. (1997) On preconditioning for the discrete wavelet transform when the sample size is not a power of two. Communications in Statistics B: Simulation and Computation, to appear.

[17] Teunissen P. J. G. (2000) Testing Theory: An introduction. Website: http://www.vssd.nl: Delft University Press. Series on Mathematical Geodesy and Positioning

[18] Teunissen P. J. G., Simons D.G., Tiberius C.C.J.M (2005) Probability and observation theory. Lecture notes AE2-E01, Faculty of Aerospace Engineering, Delft University of Technology, the Netherlands.

[19] Wang Y. (1995) Jump and sharp cusp detection by wavelets. Biometrika 82:385–397