سیزدهمین کنفرانس
آمار ایران
۲-۴ شهریور ۱۳۹۵

شناسایی نقطه‌ی تغییر در سری‌های زمانی با استفاده از روش تحلیل مجموعه مقادیر تکین

مسعود یار محمدی[۱]، مهدی کلانتری[۲]
[۱،۲]گروه آمار، دانشگاه پیام نور، صندوق پستی ۴۶۹۷-۱۹۳۹۵، تهران، ایران

چکیده در تحلیل سری‌های زمانی اگر مشاهدات جمع آوری شده شامل بخش‌هایی با توزیع‌های متفاوت باشد، یعنی نقاطی وجود داشته باشد که وضعیت ظاهری سری به دلایل مختلف از جمله تغییر در توزیع داده‌ها دستخوش تغییر شود، به این نقاط از دیدگاه آماری نقاط تغییر گویند. بنابراین در نقاط تغییر، داده‌ها به بخش‌های همگن و متجانس مجزا تقسیم می‌گردند. در عمل تعداد نقاط تغییر و محل آنها نامعلوم بوده و توانایی کشف و شناسایی آنها از دیدگاه کاربردی به ویژه مدل بندی و پیش بینی سری زمانی از اهمیت به سزایی برخوردار می‌باشد. در این مقاله نحوه‌ی شناسایی نقاط تغییر در رکوردهای سری زمانی با استفاده از یک روش ناپارامتری به نام تحلیل مجموعه‌ی مقادیر تکین مورد بررسی قرار می‌گیرد. کارایی این روش با توجه به عدم نیاز به هیچگونه فرض اولیه در مورد سری زمانی، در مقایسه با سایر روش‌های متداول قابل توجه می‌باشد.

واژه‌های کلیدی: سری زمانی، نقطه‌ی تغییر، تحلیل مجموعه مقادیر تکین.
کد موضوع‌بندی ریاضی (۲۰۱۰): 37M10، 62M10، 91B84.

۱ مقدمه

فرآیندها غالباً طی یک مدت زمان نسبتاً طولانی در حالت پایدار و بدون تغییر خود به سر می‌برند. بدیهی است هیچ فرآیندی به طور دائمی پایدار نیست و تدریجاً تغییرات با دلیل (معمولاً به صورت اتفاقی) ظاهر شده و باعث می‌شوند تا فرآیند با خواسته‌های مورد نظر انطباق نداشته باشد. این نقاط را که باعث ایجاد تغییر در ساختار توزیع داده‌ها در یک فرآیند تصادفی یا سری زمانی می‌شود را نقطه‌ی تغییر می‌نامند. نقاط تغییر به دلایل مختلفی از جمله تغییر در مکانیسم تولید داده‌ها (توزیع احتمال داده‌ها) رخ می‌دهند. روش‌های شناسایی نقطه‌ی تغییر یکی از مباحث مهم در تحلیل سری‌های زمانی است و در سال‌های اخیر مورد توجه بسیاری از پژوهشگران قرار گرفته است. تحلیل نقطه‌ی تغییر در زمینه‌های مختلفی نظیر کنترل فرآیند، پیج (۱۹۵۴)، بررسی‌های هواشناسی، واتارد و جیل (۱۹۸۹)، اقتصاد،

[۱]مسعود یار محمدی، masyar@pnu.ac.ir

چن و گوپتا (۱۹۹۷) و تحلیل رکوردهای نوار مغزی،بروسکی و همکاران (۱۹۹۹) ، استفاده شده است.

چرنوف و زاکس (۱۹۶۴) یک آزمون بیزی برای تشخیص تغییرات میانگین در مشاهدات نرمال و شناسایی نقطه‌ی تغییر به کار بردند. روش‌های بیزی متعددی توسط چئون و کیم (۲۰۱۰) ، کیم و چئون (۲۰۱۰) برای شناسایی نقاط تغییر در مدل‌های مختلف از جمله نرمال، دوجمله‌ای، نمایی و پواسن به کار گرفته شده است. اولین روش ناپارامتری برای تشخیص نقطه‌ی تغییر توسط پاتاچاریا و فریرسون (۱۹۸۱) ارائه شده است. روش آنها مشابه روش چرنوف و زاکس (۱۹۶۴) می‌باشد، با این تفاوت که از رتبه‌ی داده‌ها به جای داده‌های اصلی استفاده کرده‌اند. در رابطه با شناسایی نقطه‌ی تغییر می‌توان گفت روش‌های ناپارامتری نسبت به روش‌های پارامتری به دلیل آنکه نیازمند هیچ گونه فرض اولیه در مورد داده‌ها نیستند، از کارایی خاصی برخوردارند. مهمترین روش در این زمینه بر اساس روش تحلیل مجموعه مقادیر تکین[1] (SSA) است.

SSA بر خلاف مدل‌های کلاسیک باکس- جنکینز روشی ناپارامتری در حوزه‌ی تحلیل سری‌های زمانی است و نیازی به برقراری شرط مانایی سری زمانی و نرمال بودن خطاها ندارد. از طرف دیگر، کم بودن تعداد مشاهدات محدودیت جدی برای آن ایجاد نمی‌کند. به دلیل دارا بودن چنین مزیت‌هایی، این روش کاربردهای وسیعی در بسیاری از شاخه‌های علوم یافته است. SSA سعی در تجزیه‌ی سری زمانی به اجزایی نظیر روند، مولفه‌های فصلی (با دوره تناوب‌های مختلف) و خطای تصادفی (نوفه) دارد. در بخش دوم، این روش به اختصار معرفی خواهد شد. برای کسب اطلاعات بیشتر در مورد SSA به گولیاندینا و همکاران (۲۰۰۱) ، گولیاندینا و ژیگلجاوسکی (۲۰۱۳) و حسنی (۲۰۰۷) مراجعه کنید. در بخش سوم، الگوریتمی برای شناسایی نقاط تغییر در سری زمانی ارائه خواهد شد که مبتنی بر روش SSA است و نیازمند هیچ گونه فرض اولیه در مورد مشاهدات سری زمانی نیست. در بخش چهارم نیز نحوه‌ی به کارگیری الگوریتم ارائه شده را به کمک یک سری زمانی شبیه سازی شده نشان خواهیم داد.

۲ تحلیل مجموعه مقادیر تکین

روش SSA از دو مرحله تشکیل شده است: تجزیه و بازسازی. هر کدام از این مراحل نیز شامل دو گام هستند. فرض کنید $x_1, x_2, \dots, x_N$ یک سری زمانی مشاهده شده به طول N بوده و M نیز عددی صحیح، که طول پنجره نامیده می‌شود، باشد به طوری که $1 < M < N$. مراحل SSA عبارتند از:

۱.۲ تجزیه

این مرحله شامل دو گام است: نشاندن[2] و تجزیه‌ی مقدار تکین[3] (SVD) .

۱.۱.۲ نشاندن

در این گام، ابتدا سری زمانی را به K زیر سری تبدیل می‌کنیم وقتی که $K = N - M + 1$. زیر سری iام به ازای $i = 1, \dots, K$ به صورت $X_i = (x_i, \dots, x_{i+M-1})^T$ تعریف می‌شود. دقت کنید که زیر سری iام یک بردار ستونی با M مولفه است و گاهی اوقات بردار M-تاخیری نامیده می‌شود. سپس این بردارهای M-تاخیری را به صورت ستونی در کنار هم می‌نشانیم تا تشکیل یک

---

[1] Singular Spectrum Analysis (SSA)

[2] Embedding

[3] Singular Value Decomposition (SVD)

ماتریس $M \times K$ دهند. این ماتریس که با $\mathbf{X}$ نشان داده می‌شود، ماتریس مسیر[۴] می‌نامیم. داریم:

$$\mathbf{X} = [X_1 : \ldots : X_K] = (x_{ij})_{i,j=1}^{M,K} = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_K \\ x_2 & x_3 & x_4 & \cdots & x_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_M & x_{M+1} & x_{M+2} & \cdots & x_N \end{pmatrix}$$

دقت کنید که عناصر روی قطرهای فرعی ماتریس $\mathbf{X}$ با هم برابرند. چنین ماتریسی را ماتریس هنکل[۵] می‌نامند. از طرفی واضح است که با در اختیار داشتن ماتریس $\mathbf{X}$، می‌توان سری زمانی را به دست آورد. در واقع با کنار هم قرار دادن ستون اول و سطر آخر (یا سطر اول و ستون آخر) ماتریس $\mathbf{X}$، سری زمانی حاصل می‌شود. بنابراین یک تناظر یک به یک بین ماتریس مسیر (که هنکل نیز می‌باشد) و سری زمانی وجود دارد.

۲.۱.۲ تجزیه‌ی مقدار تکین

در این گام، تجزیه‌ی مقدار تکین ماتریس مسیر را به دست می‌آوریم. فرض کنید $\lambda_1, \ldots, \lambda_M$ مقادیر ویژه‌ی ماتریس $\mathbf{X}\mathbf{X}^T$ بوده که به صورت نزولی (از بزرگ به کوچک) مرتب شده‌اند $(\lambda_1 \geq \cdots \geq \lambda_M \geq 0)$ و $U_1, \ldots, U_M$ نیز بردارهای یکامتعامد ویژه‌ی متناظر با مقادیر ویژه‌ی $\lambda_1, \ldots, \lambda_M$ باشند. تعریف می‌کنیم $d = \max\{i, \lambda_i > 0\}$ (تعداد مقادیر ویژه‌ی مثبت) و $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$ در این صورت SVD ماتریس $\mathbf{X}$ به صورت $\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_d$ نوشته می‌شود وقتی که $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$ به ازای $i = 1, \ldots, d$. سه تایی $(\sqrt{\lambda_i}, U_i, V_i)$ را سه تایی ویژه[۶] (ET) می‌نامند.

۲.۲ بازسازی

این مرحله شامل دو گام است: گروه بندی و میانگین‌گیری قطری.

۱.۲.۲ گروه بندی

هدف از اجرای این مرحله، یافتن سه تایی‌های ویژه‌ی مربوط به هر یک از اجزای سری زمانی نظیر روند، مولفه‌های فصلی، نوفه و غیره است. در این گام، پس از اینکه SVD ماتریس مسیر به دست آمد، مجموعه‌ی اندیس‌های $\{1, \ldots, d\}$ را به m زیر مجموعه‌ی $I_1, \ldots, I_m$ افراز می‌کنیم. در واقع این افراز کردن معادل با این است که اجزای سری زمانی را به کمک سه تایی‌های ویژه تشخیص داده و از هم تفکیک کنیم. فرض کنید $I = \{i_1, \ldots, i_p\}$، در این صورت ماتریس $\mathbf{X}_I$ متناظر با گروه I به صورت $\mathbf{X}_I = \mathbf{X}_{i_1} + \cdots + \mathbf{X}_{i_p}$ تعریف می‌شود. مثلاً اگر $I = \{1, 2, 7\}$ آنگاه، $\mathbf{X}_I = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_7$. بدین ترتیب می‌توان $\mathbf{X}_I$ را به ازای $I = I_1, \ldots, I_m$ به دست آورد. در این صورت از SVD ماتریس $\mathbf{X}$ نتیجه میگیریم که $\mathbf{X} = \mathbf{X}_{I_1} + \cdots + \mathbf{X}_{I_m}$. شیوه‌ی انتخاب مجموعه‌های $I_1, \ldots, I_m$، گروه بندی سه تایی ویژه نامیده می‌شود.

---

[۴] Trajectory Matrix

[۵] Hankel Matrix

[۶] Eigen Triple

۲.۲.۴ میانگین‌گیری قطری

هدف اصلی در این گام، تبدیل هر ماتریس $\mathbf{X}_{I_j}$ ، $j = 1, \ldots, m$ ، از گام گروه بندی به یک سری زمانی به طول N است. همان طور که در گام نشاندن اشاره شد، با در اختیار داشتن یک ماتریس هنکل می‌توان سری زمانی متناظر با آن را به دست آورد؛ ولی ماتریس‌های $\mathbf{X}_{I_j}$ که در مرحله‌ی گروه بندی به دست می‌آیند دارای خاصیت هنکلی نیستند. . هنکل سازی ماتریس $\mathbf{X}_{I_j}$ به وسیله‌ی میانگین‌گیری روی عناصر قطرهای فرعی انجام می‌شود، بدین معنی که همه‌ی عناصر روی یک قطر فرعی را با میانگین عناصر همین قطر فرعی جایگزین می‌کنیم. بدین ترتیب ماتریس $\mathbf{X}_{I_j}$ به یک ماتریس هنکل تبدیل می‌شود. پس از این تبدیل، زیر سری بازسازی شده‌ی $\{\tilde{x}_1^{(j)}, \ldots, \tilde{x}_N^{(j)}\}$ به طول N به دست می‌آید. بنابراین با توجه به SVD ماتریس $\mathbf{X}$، سری زمانی اصلی به صورت مجموع m زیر سری بازسازی شده به شکل زیر بازسازی می‌شود:

$$x_t = \sum_{j=1}^{m} \tilde{x}_t^{(j)}, \quad t = 1, 2, \ldots, N.$$

۳ تشخیص نقطه تغییر بر اساس SSA

موسکوینا و ژیگلجاوسکی (۲۰۰۳) الگوریتمی را برای تشخیص نقطه‌ی تغییر ارائه داده‌اند که مبتنی بر روش SSA است. این الگوریتم شامل سه گام است. فرض کنید $x_1, x_2, \ldots$ یک سری زمانی بوده و N, M, l, p و q اعداد صحیحی باشند به طوری که $M \leq N/2$ و $l < M$ و $0 \leq p < q$. به ازای هر $n = 0, 1, \ldots$ گام‌های زیر را اجرا می‌کنیم:

گام ۱ در این گام سه مرحله‌ی اول روش SSA ، یعنی نشاندن، SVD و گروه بندی را روی سری زمانی در بازه‌ی $[n+1, n+N]$ اجرا می‌کنیم. به عبارت دیگر مراحل زیر را به ترتیب طی می‌کنیم:

۱- یک ماتریس مسیر به نام ماتریس پایه به صورت زیر می‌سازیم:

$$\mathbf{X}_B^{(n)} = \begin{pmatrix} x_{n+1} & x_{n+2} & x_{n+3} & \cdots & x_{n+K} \\ x_{n+2} & x_{n+3} & x_{n+4} & \cdots & x_{n+K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n+M} & x_{n+M+1} & x_{n+M+2} & \cdots & x_{n+N} \end{pmatrix}$$

وقتی که $K = N - M + 1$. توجه کنید که ستون‌های ماتریس پایه[7] را بردارهای $X_j^{(n)} = (x_{n+j}, \ldots, x_{n+j+M-1})^T$ $j = 1, \ldots K$ تشکیل می‌دهند.

۲- SVD ماتریس $\mathbf{R}_n = \mathbf{X}_B^{(n)} (\mathbf{X}_B^{(n)})^T$ را به دست می‌آوریم. در این مرحله M بردار ویژه حاصل می‌شود.

۳- اولین l بردار ویژه از M بردار ویژه‌ی مرحله قبل را انتخاب می‌کنیم.

[7] Base Matrix

گام ۲ در این گام یک ماتریس به نام ماتریس آزمون[۸] با بعد $M \times Q$ می‌سازیم. داریم:

$$\mathbf{X}_T^{(n)} = \begin{pmatrix} x_{n+p+1} & x_{n+p+2} & x_{n+p+3} & \cdots & x_{n+q} \\ x_{n+p+2} & x_{n+p+3} & x_{n+p+4} & \cdots & x_{n+q+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n+p+M} & x_{n+p+M+1} & x_{n+p+M+2} & \cdots & x_{n+q+M-1} \end{pmatrix}$$

وقتی که $q = p + Q$. توجه کنید که ستون‌های این ماتریس را بردارهای $X_j^{(n)}, j = p+1, \dots p+Q$ تشکیل می‌دهند.

گام ۳ در این گام آماره‌های مربوط به تشخیص نقطه‌ی تغییر محاسبه می‌شوند. این آماره‌ها عبارتند از:

- $D_{n,I,p,q} = \sum_{j=p+1}^{q}((X_j^{(n)})^T X_j^{(n)} - (X_j^{(n)})^T U U^T X_j^{(n)})$ وقتی که U یک ماتریس $M \times I$ بوده که ستون‌های آن را بردارهای ویژه‌ای که در مرحله‌ی سوم از گام اول انتخاب کردیم، تشکیل می‌دهند.
- $S_n = \tilde{D}_{n,I,p,q}/\mu_{n,I}$ وقتی که $\tilde{D}_{n,I,p,q} = \frac{1}{MQ} D_{n,I,p,q}$ ، $\mu_{n,I}$ برآوردگر $\tilde{D}_{j,I,p,q}$ در بازه‌ی زمانی $[j+1, j+m]$ است وقتی که فرض عدم وجود تغییر را بتوان پذیرفت.
- آماره‌ی جمع انباشته (CUSUM)

$$W_1 = S_1, \quad W_{n+1} = \max(0, W_n + S_{n+1} - S_n - k/\sqrt{MQ}), \quad n \geq 1$$

k یک مقدار ثابت نامنفی کوچکی است و موسکویتا (۲۰۰۱) مقدار آن را برابر $1/3\sqrt{MQ}$ پیشنهاد کرده است.

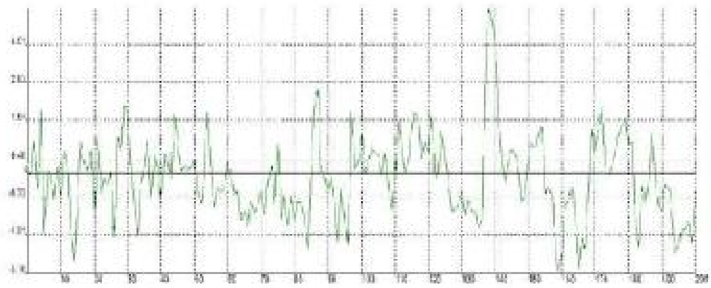
مقادیر بزرگ $D_{n,I,p,q}$، S_n و W_n نشان دهنده‌ی یک تغییر در ساختار سری زمانی است. این الگوریتم یک تغییر در زمان n را نشان می‌دهد هرگاه $W_n > h$ وقتی که $h = \frac{2z_\alpha}{MQ}\sqrt{\frac{1}{3}Q(3MQ - Q^2 + 1)}$ و z_α چندک $(1-\alpha)$ توزیع نرمال استاندارد است. برای کسب اطلاعات بیشتر در مورد نحوه‌ی تعیین پارامترهای M، I، p و q به موسکویتا و ژیگلجاوسکی (۲۰۰۳) مراجعه کنید.

۴ شبیه سازی

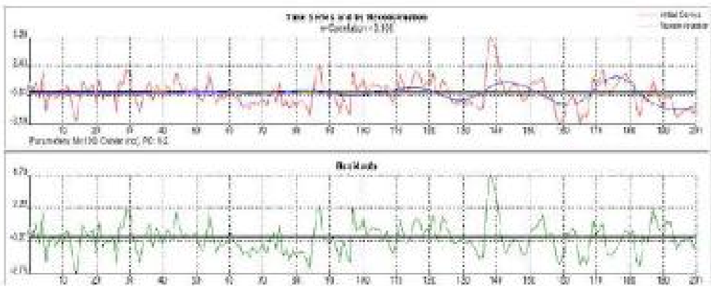
در این بخش الگوریتم ارائه شده را روی یک مثال شبیه سازی شده پیاده خواهیم کرد. ۲۰۰ مشاهده از فرآیند $AR(1)$ تولید شده است ($N = 200$). ۱۰۰ مشاهده‌ی اول با پارامتر $\phi = 0/2$ و ۱۰۰ مشاهده‌ی بعدی با پارامتر $\phi = 0/8$. توزیع نوفه‌ها هم نرمال استاندارد در نظر گرفته شده است. بدین ترتیب انتظار داریم که در زمان ۱۰۱ یک تغییر رخ دهد. شکل ۱ نمودار سری زمانی داده‌های شبیه سازی شده را نشان می‌دهد.

پارامتر طول پنجره برابر $M = \frac{N}{2} = 100$ در نظر گرفته شده است. پس از اجرای مرحله‌ی SVD و به دست آوردن مقادیر ویژه و بردارهای ویژه، دو بردار ویژه‌ی اول برای بازسازی سری زمانی در نظر گرفته شد. شکل ۲ نمودار سری زمانی بازسازی شده به وسیله‌ی دو بردار ویژه‌ی اول به همراه مانده‌ها را نشان می‌دهد. با انتخاب مقادیر $I = 2$ ، $p = 27$ و $q = 75$ مقدار h برابر با ۰/۳۳۴ به دست می‌آید. در شکل ۳ نمودار آماره‌ی تغییر نشان داده شده است. این شکل به وضوح وجود یک تغییر در زمان ۱۰۱ را نشان می‌دهد.

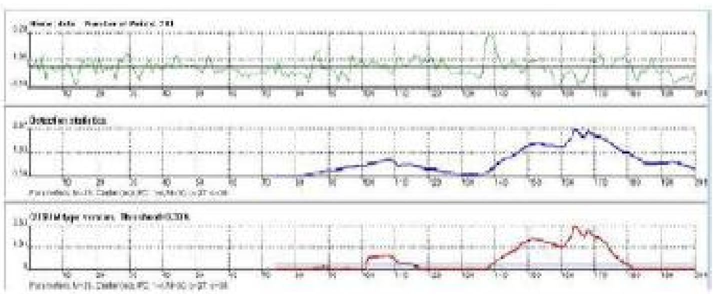
[۸] Test Matrix



شکل ۱: نمودار سری زمانی داده‌های شبیه سازی شده



شکل ۲: نمودار سری زمانی بازسازی شده



شکل ۳: نمودار آماری تغییر.

بحث و نتیجه‌گیری

شناسایی نقاط تغییر در سری‌های زمانی از دیدگاه مدل بندی و پیش بینی از اهمیت ویژه‌ای برخوردار است. در این مقاله کاربرد روش تحلیل مجموعه مقادیر تکین جهت شناسایی نقاط تغییر بیان گردید. در ادامه کارایی این روش برای شناسایی یک نقطه‌ی تغییر مشخص، در یک سری زمانی شبیه سازی شده مورد بحث و بررسی قرارگرفت. عدم نیاز این روش به فرض‌های مانایی سری زمانی، تعداد مشاهدات و نرمال بودن خطاها را می‌توان به عنوان مزیت‌های این روش محسوب نمود.

مراجع

Bhattacharya P. K. and Frierson J. R. (1981), A Nonparametric Control Chart for Detecting Small Disorders, *The Annals of Statistics*, 9, 3, 544-554.

Brodsky B.E., Darkhovsky, B.S., Kaplan A.Ya., Shishkin S.L. (1999), A nonparametric method for the segmentation of the EEG, *Computer Methods and Programs in Biomedicine*, 60, 93-106.

Chen J. and Gupta A.K. (1997), Testing and locating variance change points with application to stock prices, *Journal of the American Statistical Association*, 92, 739-747.

Cheon S. and Kim J. (2010), Multiple change-point detection of multivariate mean vectos with the bayesian approach. *Computational Statistics and Data Analysis*, 54, 406-415.

Chernoff H. and Zacks S. (1964), Estimating the current mean of a normal distribution which is subject to changes in time. *Annals of Mathematical Statistics*, 35, 999-1018.

Golyandina N., Nekrutkin V., Zhigljavsky A., (2001), *Analysis of time series structure: SSA and related techniques*, Boca Raton, Chapman and Hall/CRC.

Golyandina N. and Zhigljavsky A. (2013), *Singular Spectrum Analysis for Time Series*, London, Springer.

Hassani H. (2007), Singular Spectrum Analysis: Methodology and Comparison, *Journal of Data Science*, 5(2), 239–257.

Kim J. and Cheon S. (2010), Bayesian multiple change-point estimation with annealing stochastic approximation monte carlo, *Computational Statistics* 25, 215-239.

Moskvina V. and Zhigljavsky A. (2003), An algorithm based on singular spectrum analysis for change-point detection, *Communication in Statistics: Simulation and Computation*, 32(2), 319-352.

Moskvina V. (2001), Application of the Singular-Spectrum Analysis for Change-Point Detection in Time Series. Ph.D. thesis, School of Mathematics, Cardiff: Cardiff University.

Page E.S. (1954), Continuous inspection scheme, *Biometrika*, 1, 100-115.

Vautard R. and Ghil M. (1989), Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series, *Physica D*, 35, 395-424.