پیش‌بینی سری‌های زمانی قضایی با استفاده از روش تحلیل مجموعه‌ی مقادیر تکین

فاطمه سادات آل‌حسینی
گروه آمار، دانشکده علوم ریاضی، دانشگاه شهید بهشتی

چکیده

تعیین روند شاخص‌های مهم آماری یک سازمان در هر مقطع زمانی، نقشی کلیدی در ارزیابی عملکرد گذشته، تحلیل وضعیت موجود، برنامه‌ریزی در سطوح مختلف مدیریتی و تعیین خط‌مشی آتی آن سازمان خواهد داشت. از این بین، تحلیل داده‌های قضایی به دلیل دربرداشتن طیف وسیعی از شاخص‌ها و اهمیت آن در حفظ و ارتقای امنیت جامعه، از اهمیت بالایی برخوردار است. داده‌های دستگاه قضا، طیف وسیعی از داده‌های ثبتی را در شعب تحت پوشش شامل می‌شود که مدل‌بندی آنها به شکل یک سری زمانی و در ادامه، پیش‌بینی هر یک می‌تواند بستر مناسبی را به منظور تحلیل و اتخاذ تصمیم‌گیری‌های موثر در جهت کنترل جرم به وجود آورد. از سوی دیگر، این داده‌ها به دلیل تاثیرپذیری شدید و مستقیم از وضعیت اقتصادی، سیاسی و معیشتی افراد جامعه، در اغلب موارد، از پیش فرض‌های رایجی مانند ایستایی و نرمال بودن در مدل‌های پارامتری سری‌های زمانی پیروی نکرده و درنتیجه، امکان مدل‌بندی و در ادامه، پیش‌بینی آن‌ها به کمک این مدل‌ها وجود ندارد. تحلیل مجموعه‌ی مقادیر تکین، یک روش ناپارامتری نسبتاً جدید در تحلیل سری‌های زمانی است که به پیش‌فرض‌های رایجی که برخی از آنها ذکر شد، وابسته نبوده و تاکنون، توانایی خود را به خوبی در مدل‌بندی و پیش‌بینی انواع مختلفی از سری‌های زمانی نشان داده است. از این رو، در این مقاله، به بررسی امکان استفاده از این روش در مدل‌بندی و پیش‌بینی سری‌های زمانی قضایی خواهیم پرداخت.

واژه‌های کلیدی: اتهام، تجزیه‌ی ویژه مقدار، روش تحلیل مجموعه‌ی مقادیر تکین.

۱ مقدمه

نتایج مطالعه‌ی روند تغییرات شاخص‌های اصلی هر سازمان در طول زمان، به فراخور موضوع آنها، می‌تواند در تحلیل روند گذشته، بررسی میزان اثرگذاری سیاست‌های پیشین، تعیین دوره‌های زمانی پرخطر (دوره‌هایی که سازمان با حجم انبوهی از پیشامدهای نامطلوب مواجه می‌شود) و به منظور کسب آمادگی لازم و اتخاذ تدابیر کارساز در آینده مورد استفاده قرار گیرد.

تحلیل داده‌های قضایی به دلیل دربرداشتن طیف وسیعی از شاخص‌ها و اهمیت آن در حفظ و ارتقای امنیت جامعه از اهمیت بالایی برخوردار است. داده‌های دستگاه قضا، طیف وسیعی از انواع داده‌های تثبیت شده در شعب تحت پوشش را (مجتمع‌های قضایی و دادسراها) به شکل ماهانه، فصلی و سالانه شامل می‌شود که برای مثال، مدل‌بندی و در ادامه، پیش‌بینی سری ورودی هر یک می‌تواند بستر مناسبی را در جهت اتخاذ تصمیم‌گیری‌های راهبردی آتی به منظور کنترل سطح آنها در جامعه ایجاد کند.

در اینجا، ارایه‌ی تعریفی کوتاه از دو مفهوم کلیدی در واژگان تخصصی قضایی خالی از لطف نیست. پرونده‌های ورودی به دستگاه قضا در یکی از دو گروه طرح دعاوی (خواسته) یا وقوع جرایم (اتهام) قرار می‌گیرند. خواسته، آن چیزی است که خواهان (کسی که برای ادعای خود در دادگاه اقامه‌ی دعوا می‌کند) از دادگاه درخواست کرده و خوانده را (کسی که خواهان به ضرر او در دادگاه اقامه‌ی دعوا می‌کند) به انجام آن یا پرداخت وجه محکوم می‌کند؛ درحالی که اتهام، جرمی است که به کسی نسبت داده می‌شود و در دادسراها مورد بررسی قرار می‌گیرند. برای مثال، «تهدید» یک اتهام و «مطالبه‌ی مهریه» یک خواسته است. بدیهی است که تعداد اتهام و مقدار ورودی آن در هر بازه‌ی زمانی نسبت به خواسته بیشتر خواهد بود.

از سوی دیگر در اغلب موارد، این داده‌ها به دلیل تاثیرپذیری شدید و مستقیم از تغییرات اقتصادی، سیاسی و معیشتی افراد جامعه از الگوی مشخصی پیروی نکرده و تابع فرض‌های موجود در روش‌های پارامتری سری‌های زمانی نیستند. از این رو، به عنوان یک راهکار در مدل‌بندی و پیش‌بینی چنین داده‌هایی می‌توان از روش‌های ناپارامتری سری‌های زمانی بهره جست.

تحلیل مجموعه‌ی مقادیر تکین[1] (SSA)، یک روش ناپارامتری نسبتاً جدید در تحلیل سری‌های زمانی است که به کمک تجزیه‌ی ویژه‌مقدارها[2] به کاهش سطح نوفه[3]، مدل‌بندی و پیش‌بینی سری‌های زمانی می‌پردازد. سادگی، عدم وابستگی به پیش‌فرض‌های محدودکننده‌ی رایج در دیگر روش‌ها (همچون مانایی، خطی بودن سری و نرمال بودن باقی‌مانده‌ها)، کارایی مناسب در تحلیل سری‌های زمانی با طول کم و توانمندی این روش در بررسی مسئله‌های متعددی در این مبحث از جمله دلایل گسترش روزافزون این روش در علوم مختلف است؛ علاقه‌مندان به مطالعه‌ی نمونه‌ای از کاربردهای آن می‌توانند به نجاری [۴] یا آل‌حسینی [۱] مراجعه کنند.

در این مقاله به بررسی امکان استفاده از روش تحلیل مجموعه‌ی مقادیر تکین در مدل‌بندی و پیش‌بینی رفتار آتی سری‌های قضایی خواهیم پرداخت. یک مطالعه‌ی اولیه در این خصوص در آل‌حسینی [۲] ارایه شده است. در بخش آتی، مقدمه‌ی کوتاه‌ای از روش SSA ارایه شده است. در بخش سوم، پس از معرفی سری تحت بررسی، به چگونگی بازسازی و پیش‌بینی مقادیر آتی آن به روش SSA خواهیم پرداخت. بخش پایانی نیز به بحث در خصوص کیفیت نتایج و راهکارهای ممکن به منظور بهبود آنها اختصاص دارد.

۲ تحلیل مجموعه‌ی مقادیر تکین

یکی از مهمترین مسئله‌ها در تحلیل سری‌های زمانی یافتن روشی است که در کنار توانایی بالا، نسبت به تغییرهای

[1] Singular Spectrum Analysis

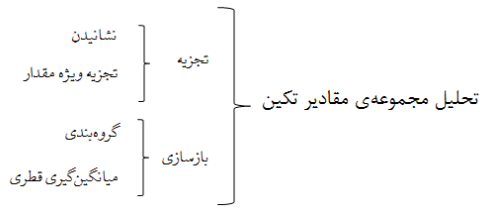
[2] Singular Value Decomposition

[3] Noise

تصادفی نامنظم (نوفه) استوار باشد. تلاش‌های بسیاری در این حوزه انجام گرفته است، با این حال بسیاری از این روش‌ها مبتنی بر فرض‌های محدودکننده‌ای همچون مانایی، خطی بودن سری و نرمال بودن باقی‌مانده‌ها است. یک راه حل برای رفع این مشکل، تلاش در جهت یافتن روشی ناپارامتری است که درکنار برخورداری از توانمندی بالا، نسبت به نوفه‌ی موجود در سری زمانی مورد مطالعه نیز استوار باشد. روشهای بسیاری به منظور کاهش سطح نوفه وجود دارد که در میان آن‌ها، روش‌های مبتنی بر تجزیه‌ی ویژه‌مقدار از توانمندی بیشتری برخوردار هستند. پیدایش روش SSA اغلب به مقاله‌های بروم‌هد و کینگ [۵، ۶] نسبت داده می‌شود. تحقیق‌های این دو نفر و جذابیت روش SSA منجر به کاربرد آن در حوزه‌های مختلفی از علوم همچون علوم فیزیکی، اقتصادی، زیستی و مهندسی شده است. در این خصوص، بیش از صدها مقاله به نگارش درآمده است که بر اساس نتایج آن‌ها، روش SSA در مدل‌بندی بسیاری از داده‌ها در مقایسه با سایر روش‌های موجود دارای توانمندی بسیار بالاتری است (علاقه‌مندان به مطالعه‌ی بیشتر در این خصوص می‌توانند به محمودوند [۳] مراجعه کنند).

۱.۲ ساختار روش

هدف اصلی SSA تجزیه‌ی سری اصلی به تعدادی زیر سری است، به طوری که هر زیر سری را بتوان به عنوان روند، دوره، مولفه‌های فصلی یا نوفه‌ی سری تحت مطالعه در نظر گرفت. ویرایش اصلی تحلیل مجموعه‌ی مقادیر تکین شامل دو مرحله است؛ هر یک از مرحله‌های فوق نیز خود شامل دو مرحله هستند. شکل ۱ مراحل اجرای SSA را نمایش می‌دهد.



شکل ۱: مراحل اجرای تحلیل مجموعه‌ی مقادیر تکین.

الف) نشاندن[۴]

نشاندن را می‌توان به عنوان یک نگاشت به منظور تبدیل یک سری زمانی حقیقی مقدار، غیر صفر و یک بعدی با طول N $(N > ۲)$ همچون $Y_N = (y_۱, ..., y_N)^T$، به سری‌های زمانی چند بعدی $X_k, ..., X_۱$ (هرکدام به طول L) با بردارهای $X_i = (y_i, ..., y_{i+L-۱})^T \in R^L$ که $i = ۱, ..., k$، در نظر گرفت که در آن عدد صحیح L $(۲ \le L \le N - ۱)$ طول پنجره[۵] نامیده می‌شود و $K = N - L + ۱$. نتیجه‌ی گام نشاندن، ماتریس $X = [X_۱, ..., X_k]$ است که بردارهای تاخیر[۶]، ستون‌های این ماتریس را تشکیل می‌دهند و آن را ماتریس مسیر [۷] می‌نامند:

$$X = (x_{ij})_{i,j=۱}^{L,K} \begin{bmatrix} y_۱ & y_۲ & y_۳ & \cdots & y_k \\ y_۲ & y_۳ & y_۴ & \cdots & y_{k+۱} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+۱} & y_{L+۲} & \cdots & y_N \end{bmatrix}. \quad (۱)$$

ماتریس X را می‌توان به عنوان یک مجموعه داده‌ی چند متغیره با L مشخصه و k مشاهده در نظرگرفت. با توجه به فرمول (۱)، واضح است که $x_{ij} = y_{i+j-۱}$. چنین ماتریسی که عناصر روی قطرهای فرعی آن با هم برابرند، ماتریس هنکل[۸]

[۴] Embeding
[۵] Window Length
[۶] Lagged Vector
[۷] Trajectory Matrix
[۸] Hankel Matrix

می‌نامند. با توجه به ساختار این ماتریس، به ازای هر L و N ثابت، همواره رابطه‌ی یک به یکی میان ماتریس مسیر و سری زمانی اصلی وجود دارد. در حقیقت، درایه‌های ستون اول و سطر آخر (یا سطر اول و ستون آخر) ماتریس مسیر همان سری زمانی تک متغیره اصلی است.

ب) تجزیه‌ی ویژه‌مقدار

دومین گام از مرحله‌ی نخست، تجزیه‌ی ماتریس مسیر بر اساس تجزیه‌ی ویژه‌مقدار است. در این گام، ماتریس مسیر به کمک تجزیه‌ی ویژه‌مقدار به مجموع ماتریس‌های مقدماتی متعامد (هر یک با رتبه‌ی یک) تجزیه می‌شوند. ماتریس $S = X'X$ با بعد $L \times L$ را در نظر بگیرید. فرض کنید، $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L$ بیانگر ویژه‌مقدارهای[۹] این ماتریس و $U_1, ..., U_L$ مجموعه‌ی یکامتعامد[۱۰] از ویژه‌بردارهای[۱۱] متناظر با این ویژه‌مقدارها باشند. قرار دهید:

$$d = max\{i; \lambda_i > 0\},$$

که d همان رتبه‌ی ماتریس X است. در این صورت، تجزیه‌ی ویژه‌مقدار ماتریس X به شکل زیر خواهد بود:

$$X = X_1 + ... + X_d, \qquad (۲)$$

که برای

$$V_i = \frac{X'U_i}{\sqrt{\lambda_i}}, \quad X_i = \sqrt{\lambda_i} U_i V_i^T, \quad i = 1, .., d,$$

است. اگر ویژه‌مقدارها غیرتکراری باشند، بسط (۲) به شکلی یکتا تعریف می‌شود.

ج) گروه‌بندی[۱۲]

در گروه‌بندی، d ماتریس موجود در رابطه‌ی (۲) به چند زیرگروه تقسیم می‌شوند. فرض کنید، $I = \{1, ..., d\}$ به m زیرگروه $I_m, ..., I_1$ افراز شود؛ در این صورت، رابطه‌ی (۲) را می‌توان به شکل زیر بازسازی کرد:

$$X = X_{I_1} + ... + X_{I_m}, \qquad (۳)$$

که در این رابطه

$$X_{I_j} = \sum_{l \in I_j} X_l, \qquad j = 1, ..., m.$$

در ساده‌ترین حالت ممکن $(m = 2)$، گروه نخست همان مولفه‌ی اصلی سری (سیگنال) و گروه باقی‌مانده، نوفه در نظر گرفته می‌شود. در این حالت، r مقدار از بزرگترین مقدارهای تکین (ریشه‌ی دوم ویژه‌مقدارها) و ویژه‌بردارهای متناظر آن‌ها به منظور تقریب سری اصلی انتخاب شده و سایر مقدارهای تکین به عنوان مولفه‌های نوفه در نظر گرفته می‌شوند (در ادبیات SSA، پارامتر r را نقطه‌ی برش[۱۳] می‌نامند).

[۹] Eigen values
[۱۰] Ortho normal
[۱۱] Eigen vectors
[۱۲] Grouping
[۱۳] Cut Point

د) میانگین‌گیری قطری[۱۴]

وجود رابطه‌ی یک به یک میان ماتریس هنکل حاصل از SSA و سری تحت مطالعه، یکی از ویژگی‌های مفید این روش است. از سوی دیگر، ماتریس‌هایی که در مرحله‌ی گروه‌بندی حاصل می‌شوند، دارای خاصیت هنکلی نیستند و از این رو، نمی‌توان به کمک آن‌ها تقریب یکتایی از سری اصلی به دست آورد. یک روش بهینه برای حل این مشکل استفاده از میانگین‌گیری قطری است. (گولیندینا و همکاران [۹]). هدف از میانگین‌گیری قطری، تبدیل یک ماتریس به ماتریس هنکلی است که در ادامه، می‌توان آن را به یک سری زمانی برگرداند. اگر z_{ij} درایه‌ی ijام ماتریس Z باشد، kامین جمله از سری حاصل با میانگین‌گیری z_{ij} بر روی تمامی i و j هایی که $i+j=k+1$، بدست می‌آید. این رویه، میانگین‌گیری قطری یا هنکل‌سازی ماتریس Z نامیده می‌شود. بنابراین، هنکل‌سازی را می‌توان به این شکل فرمول‌بندی کرد: فرض کنید، Y یک ماتریس $L\times K$ با عناصر y_{ij}، $1\leqslant j\leqslant K$ و $1\leqslant i\leqslant L$، باشد. قرار دهید، $s_1=\min\{L,i+j-1\}$ و $s_2=\max\{1,i+j-N-1+L\}$. از این رو، عنصر $\tilde{y}_{ij}$ ماتریس $\mathcal{H}Y$ ($\mathcal{H}$ عملگر هنکل‌سازی است) برابر است با [۳]:

$$\tilde{y}_{ij}=\frac{1}{s_2-s_1+1}\sum_{l=s_1}^{s_2}y_{l,i+j-l-1},$$

بنابراین، با هنکل‌سازی رابطه‌ی (۳)، بسط

$$X=\tilde{X}_{I_1}+...+\tilde{X}_{I_m},$$

به دست می‌آید که در آن داریم:

$$\tilde{X}_{I_j}=\mathcal{H}X_{I_j},\quad j=1,...,m.$$

با انتخاب مناسب r مقدار از m زیر سری، سری بازسازی شده از مرحله‌ی میانگین‌گیری قطری، یک سری با سطح نوفه‌ی اندک خواهد بود و از این رو، می‌توان از آن برای پیش‌بینی مقادیر آتی سری استفاده کرد.

۲.۲ پیش‌بینی با استفاده از روش تحلیل مجموعه‌ی مقادیر تکین

در روش SSA، با تفکیک سری اصلی به چند زیرسری، سعی در حذف نوفه از سری زمانی تحت مطالعه را داریم؛ حال، اگر سری به درستی پالایش شده و سپس برای پیش‌بینی به کارگرفته شود، انتظار می‌رود که دقت پیش‌بینی حاصل از این روش نیز بالاتر رود. در ادامه، به بیان دو روش اصلی در پیش‌بینی بر اساس SSA خواهیم پرداخت.

الف) پیش‌بینی تحلیل مجموعه‌ی مقادیر تکین به روش بازگشتی[۱۵]

از آنجایی که این الگوریتم پیش‌بینی مبتنی بر یک رابطه‌ی بازگشتی است، عنوان $RSSA$ برای آن در نظرگرفته شده است که در آن، حرف R از ابتدای کلمه‌ی $Recurrent$ گرفته شده است. این روش را دانیلو [۷] و دانیلو و ژیگلافسکی [۸] معرفی کردند.

اساس این روش مبتنی بر این فرض است که اگر زیرفضای r-بعدی $\mathfrak{L}_r=\{U_1,...,U_r\}$ از $\mathbb{R}^L$ یک فضای عمودی نباشد (U_i، ویژه‌بردارهای چپ

[۱۴] Diagonal Averaging

[۱۵] Recurrent Singular Spectrum Analysis (RSSA)

ماتریس مسیر X هستند)، آنگاه هر بردار در این زیرفضا در یک رابطه‌ی بازگشتی صدق می‌کند، به طوری که آخرین مولفه‌ی آن را می‌توان به شکل یک ترکیب خطی منحصربه فرد از سایر مولفه‌ها نوشت؛ برای اثبات، گولیندینا و همکاران [۹] را ببینید. بنابراین، هر یک از گروه‌های ایجادشده در مرحله‌ی گروه‌بندی را، با فرض وجود شرط فوق، می‌توان به کمک این رابطه‌ی بازگشتی ادامه داد. از این رو، اگر سری زمانی به گونه‌ای باشد که بتوان چنین رابطه‌ی بازگشتی را برای آن توصیف کرد، آنگاه M جمله‌ی نهایی سری به کمک یک رابطه‌ی بازگشتی منحصربه فرد به شکل زیر برآورد می‌شوند: (۵)

$$g_i = \begin{cases} \tilde{y}_i, & i = ۱, ..., N, \\ \sum_{j=۱}^{L-۱} a_j g_{i-j}, & i = N+۱, ..., N+M. \end{cases} \quad (۴)$$

در رابطه‌ی فوق، بردار ضرایب، $\mathfrak{R} = (a_۱, ..., a_{L-۱})$، به کمک رابطه زیر محاسبه می‌شود:

$$\mathfrak{R} = \frac{۱}{۱ - (\pi_۱^۲ + ... + \pi_r^۲)} \sum_{i=۱}^{r} \pi_i P_i^{\nabla},$$

که در آن $\{P_۱, ..., P_r\}$ یک پایه‌ی متعامد یکه برای زیرفضای $\mathfrak{L}_r$، P_i^{∇} برداری شامل $L-۱$ مولفه‌ی نخست P_i، π_i مولفه آخر بردار P_i و r نقطه‌ی برش تعیین شده در مرحله‌ی گروه‌بندی است. فرمول بازگشتی (۴) در حقیقت بیانگر پیش‌بینی سری زمانی تا M گام جلوتر است.

نکته‌ای که در این روش حائز اهمیت است، عدم وابستگی پیش‌بینی به یک الگوی از پیش تعیین شده است؛ در حقیقت، در این الگوریتم، پیش‌بینی‌ها تنها با استفاده از رابطه‌های موجود میان داده‌های تحت مطالعه به دست می‌آیند.

ب) پیش‌بینی تحلیل مجموعه‌ی مقادیر تکین به روش برداری[۱۶]

در این روش، به منظور پیش‌بینی M گام جلوتر یک سری به کمک رابطه‌ی بازگشتی (۵)، ماتریس مسیر بازسازی شده را با اضافه کردن تعداد $M+L-۱$ ستون جدید برآورد می‌کنیم.

$$Z_i = \begin{cases} \tilde{X}_i, & i = ۱, ..., k, \\ \mathcal{P}^{(\vartheta)} Z_{i-۱}, & i = k+۱, ..., k+M+L-۱. \end{cases}$$

در رابطه‌ی فوق،

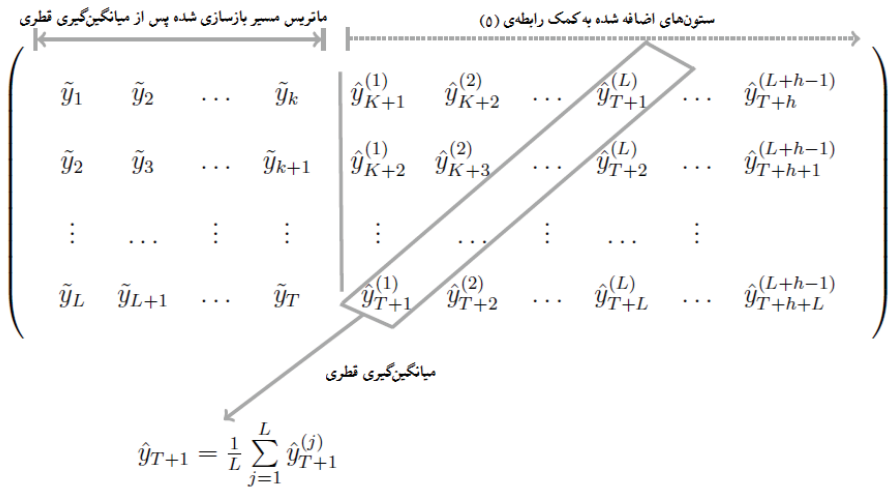
$$\mathcal{P}^{(\vartheta)} = \begin{bmatrix} \Pi \\ \mathfrak{R} \end{bmatrix},$$

$$\Pi = \sum_{i=۱}^{r} P_i^{\nabla} P_i^{\nabla T} + \frac{۱}{۱ - (\pi_۱^۲ + ... + \pi_r^۲)} \mathfrak{R}\mathfrak{R}^T.$$

اکنون، با اعمال میانگین‌گیری قطری بر ماتریس حاصل (Z)، مولفه‌های $N+۱$ تا $M+N$ سری بیانگر پیش‌بینی سری زمانی تا M گام بعد خواهند بود. شکل ۲، پیش‌بینی برداری SSA را در یک گام جلوتر نمایش می‌دهد.

همان‌طور که مشاهده می‌شود، پیش‌بینی به روش برداری برای یک افق یک ساله بر اساس L مقدار است؛ این درحالی است که با توجه به آنچه که در قسمت قبل ذکر شد، این پیش‌بینی در روش

[۱۶] Vertical Singular Spectrum Analysis



شکل ۲: پیش‌بینی برداری تحلیل مجموعه‌ی مقادیر تکین.

بازگشتی بر اساس یک مقدار (اولین L در روش برداری) انجام می‌گیرد.

گولیندینا و همکاران [۹] و پیلیشُو [۱۳]، در مطالعه‌های خود به مقایسه‌ی توانایی این دو روش پیش‌بینی پرداخته‌اند. از نتایج این بررسی‌ها چنین می‌توان استنباط کرد که روش $VSSA$ محافظه‌کارتر از روش $RSSA$ است [۳].

۳.۲ تعیین پارامترهای روش تحلیل مجموعه‌ی مقادیر تکین

همانطور که ذکر شد، در اولین گام از روش SSA، سری اصلی به چند زیرسری هر یک با طول L تقسیم می‌شوند. بنابراین، یکی از پارامترهای مهم SSA، مقدار طول پنجره‌ی L است که بایستی قبل از مرحله‌ی گروه‌بندی انتخاب شود. نتایج کاربردی نشان می‌دهد که دقت روش SSA بستگی زیادی به این پارامتر دارد و مقادیر نامناسب آن منجر به نتایج نادقیق و نامطمئن خواهد شد (برای مثال، گولیندینا و همکاران [۹] را ببینید).

معیارهای مختلفی به منظور انتخاب پارامتر طول پنجره معرفی شده است که می‌توان آن‌ها را به دو گروه کلی تقسیم کرد:

۱- معیارهای عمومی؛

۲- معیارهای مسئله محور.

در گروه اول، طول پنجره بدون در نظر گرفتن ماهیت داده‌ها و بر اساس ویژگی‌های مختلف روش SSA انتخاب می‌شوند. معیارهایی همچون اثر ماتریس کوواریانس تاخیر، رتبه‌ی ماتریس مسیر، تعداد عناصر ماتریس مسیر (برای مطالعه‌ی بیشتر هر سه معیار محمودوند [۳] را ببینید) و تفکیک‌پذیری (حسنی [۱۰، ۱۱] و محمودوند و همکاران [۱۲]) در این گروه جای دارند.

در گروه دوم، طول پنجره با توجه به هدف تحلیل و ماهیت داده‌های تحت بررسی انتخاب می‌شود. معیارهایی همچون دقت بازسازی و پیش‌بینی یک

سری نیز در این گروه قرار می‌گیرند. محمودوند و همکاران [۳] به کمک شبیه‌سازی نشان دادند که میان مقادیر بهینه‌ی پارامتر L در بازسازی و پیش‌بینی یک سری زمانی تفاوت‌های معنی‌داری وجود دارد. بر اساس این بررسی، مقدار بهینه برای بازسازی یک سری مقداری برابر میانه‌ی سری است؛ در حالی که مقدار بهینه‌ی این پارامتر در پیش‌بینی یک سری زمانی مفروض به شرایط مختلفی بستگی دارد.

از سوی دیگر، در قسمت گروه‌بندی نیز تعداد m زیرگروه برای بازسازی مولفه‌های سری مورد استفاده قرار می‌گیرند. واضح است که انتخاب نادرست زیرگروه‌ها منجر به بازسازی نادقیق مولفه‌ها خواهد شد. بنابراین، چگونگی گروه‌بندی نیز تاثیر بسیار زیادی در نتایج خواهد داشت.

دو روش متداول در گروه‌بندی شامل رسم نمودار مقادیر تکین ماتریس مسیر و استفاده از مفهوم تفکیک‌پذیری است [۹]. در مورد اول، معمولاً در صورتی که دو زیرسری در یک گروه قرار داشته باشند، مقادیر تکین آنها به هم نزدیک خواهد بود. بنابراین، رسم نمودار مقادیر تکین به عنوان یک راه حل بصری برای گروه‌بندی مورد استفاده قرار می‌گیرد. در مورد دوم، همانطور که در بالا اشاره شد، سری زمانی اصلی در مرحله‌ی گروه‌بندی به چند زیر سری تفکیک می‌شود؛ نکته‌ای که در این میان از اهمیت بالایی برخوردار است، دقت در تعیین این زیرگروه‌ها است. عدم توجه کافی به انتخاب گروه‌های مناسب منجر به تداخل نوفه در سیگنال (سیگنال در نوفه) شده و در این صورت از کارایی SSA کاسته می‌شود. برای بررسی این موضوع، در ادبیات SSA مفهوم تفکیک‌پذیری مطرح شده است. در ادامه، این مفهوم را با جزئیات بیش‌تری بررسی می‌کنیم.

فرض کنید، سری زمانی Y_N مجموع دو سری زمانی $Y_N^{(۱)}$ و $Y_N^{(۲)}$ است. طول پنجره L را ثابت در نظر گرفته و ماتریس‌های مسیر مربوط به سری‌های Y_N، $Y_N^{(۱)}$ و $Y_N^{(۲)}$ را به ترتیب با X، $X^{(۱)}$ و $X^{(۲)}$ نشان می‌دهیم.

تعریف ۱. *دو سری زمانی $Y_N^{(۱)}$ و $Y_N^{(۲)}$ تفکیک‌پذیری ضعیف دارند، هرگاه، هر سطر (ستون) از ماتریس مسیر $X^{(۱)}$ بر هر سطر (ستون) ماتریس مسیر $X^{(۲)}$ عمود باشد [۹].*

به منظور اندازه‌گیری میزان تفکیک‌پذیری ضعیف، نوع خاصی از ضریب همبستگی، موسوم به ضریب همبستگی موزون یا w-همبستگی، مورد استفاده قرار می‌گیرد. برای هر دو بردار $Y_N^{(۱)}$ و $Y_N^{(۲)}$، مقدار w-همبستگی به کمک رابطه‌ی زیر محاسبه می‌شود:

$$\rho^{(\omega)} = \frac{\langle Y_N^{(۱)}, Y_N^{(۲)} \rangle}{\sqrt{\langle Y_N^{(۱)}, Y_N^{(۱)} \rangle . \langle Y_N^{(۲)}, Y_N^{(۲)} \rangle}},$$

که در این رابطه،

$$\langle Y_N^{(s)}, Y_N^{(t)} \rangle = \sum_{j=۱}^{N} \omega_i^{L,N} y_j^{(s)} y_j^{(t)}, \quad s,t = ۱,۲,$$

و در آن،

$$\omega_j^{L,N} = \min\{j, L^*, N-j+۱\},$$

و

$$L^* = \min\{L, N-L+۱\}.$$

چنانچه مقدار w-همبستگی صفر یا کوچک باشد، دو بردار تفکیک‌پذیری ضعیف دارند.

همچنین برای مطالعه‌ی بیشتر در خصوص انتخاب پارامتر بهینه‌ی نقطه‌ی برش می‌توانید به محمودوند [۳] یا محمودوند و همکاران [۱۲] مراجعه کنید.

۳ بازسازی و پیش‌بینی سری ورودی ماهانه‌ی اتهام

در این بخش، پس از معرفی داده‌های تحت بررسی، به مدل‌بندی و پیش‌بینی سری مفروض به دو روش $RSSA$ و $VSSA$ خواهیم پرداخت.

۱.۳ داده‌ها

پس از ثبت اطلاعات پرونده‌های قضایی هر ماه در سامانه توسط کاربران هر واحد قضایی، واحد قضایی مرجع در اوایل ماه بعد، داده‌های ثبت شده در ماه قبل را تثبیت کرده و در سامانه‌ی آماری، اطلاعات جامعی همچون تعداد موجودی اول ماه، ورودی در طی ماه، خروجی در طی ماه و مانده پرونده‌های آن ماه قابل رویت خواهد بود. این داده‌ها ملاک گزارش‌گیری و تحلیل‌های آماری آتی خواهند بود.

در حال حاضر، دو سامانه‌ی شهر تهران و شهرستانهای استان تهران به منظور ثبت اطلاعات پرونده‌های قضایی استان تهران وجود دارد. به دلیل فراگیری بیشتر استفاده از این سامانه در شهر تهران و در نتیجه، جامعیت بالاتر داده‌ها، در این مقاله گزارش‌گیری از اطلاعات شهر تهران انجام شده است.

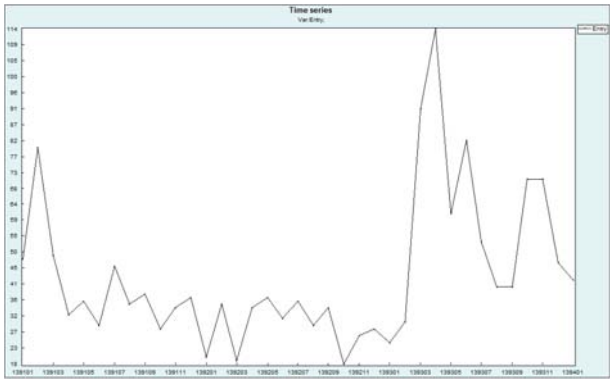
همچنین، نظر به کثرت تعداد عناوین اتهام در مفاهیم قضایی (چیزی حدود ۲۰۷۷ عنوان اتهام)، مدل‌بندی سری تحت بررسی به دلخواه و بر روی یکی از اتهام‌های متداول در شهر تهران (کیف‌زنی) انجام شده است؛ هرچند این تحلیل قابل تعمیم به دیگر عناوین اتهام یا خواسته نیز خواهد بود. همچنین، گزارش‌گیری این مقاله محدود به فروردین ماه ۱۳۹۱ تا فرودین ماه ۱۳۹۴ (حداکثر بازه‌ی زمانی با داده‌های ورودی ماهانه‌ی قابل دسترسی در سامانه‌ی جامع آماری این ارگان تا زمان انجام این مطالعه) بوده است (ثبت اطلاعات در سامانه‌ی آماری این ارگان از سال ۱۳۸۹ آغاز شده است، با این حال، داده‌های ماهانه به شکل کامل تنها برای سال‌های ۱۳۹۱ و بعد از آن در دسترس است).

شایان ذکر است که عواملی همچون خطای کاربر، عدم ثبت درست و کامل پرونده‌ها توسط واحدهای قضایی و نیز تاثیرپذیری شدید داده‌ها (در اینجا، ورودی اتهام) از وضعیت اقتصادی، اجتماعی یا سیاسی جامعه تنزل اعتبار داده‌ها را ممکن می‌سازد. با این حال، از آنجا که این داده‌ها، با وجود تمام عوامل خطاساز مذکور، مرجع اصلی تحلیل‌های آماری به حساب می‌آیند، سعی در مدل‌بندی هرچه دقیق‌تر آنها در این مقاله شده است.

۲.۳ مدل‌بندی و پیش‌بینی به روش SSA

سری ورودی ماهانه‌ی اتهام کیف‌زنی از فروردین ۱۳۹۱ تا فرودین ماه ۱۳۹۴ در شکل (۳) نمایش داده شده است. همانطورکه مشاهده می‌کنید، این سری دارای روند ناخطی بوده و الگوی دوره‌ی آن در دو سال ۱۳۹۱ و ۱۳۹۲ تقریباً مشابه و کاملا متفاوت از الگوی سال ۱۳۹۳ است. بررسی تغییرات وضعیت اقتصادی جامعه (در اینجا، شهر تهران) و سطح معیشتی افراد این جامعه در سال ۱۳۹۳ نسبت به دو سال قبل از آن ممکن است ما را در فهم علت این تفاوت چشمگیر در سطح ورودی پرونده‌ی اتهام

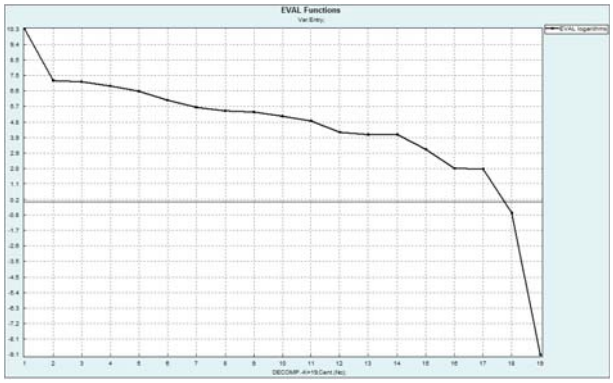
کیف‌زنی یاری رساند.



شکل ۳: سری زمانی ورودی اتهام کیف‌زنی از فروردین ۱۳۹۱ تا فروردین ۱۳۹۴.

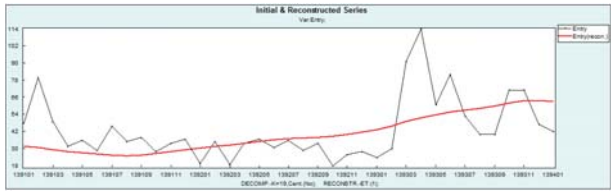
با انتخاب L برابر میانه‌ی طول سری در شکل (۴)، لگاریتم ۱۹ ویژه‌مقدار حاصل از ماتریس مسیر نشان داده شده است.

اولین لگاریتم ویژه‌مقدار، با بیشترین اختلاف مقدار از



شکل ۴: لگاریتم ویژه‌مقدارهای سری زمانی اتهام کیف‌زنی از فروردین ۱۳۹۱ تا فروردین ۱۳۹۴.

دیگر مقادیر، بیانگر روند سری تحت مطالعه است. سری بازسازی شده براساس ویژه سه‌گانه‌ی[۱۷]نخست (گردایه‌ی $(\sqrt{\lambda_i}, U_i, V_i)$ را ویژه سه‌گانه‌ی iام SVD معادله‌ی (۱) می‌نامند) در شکل (۵) نمایش داده شده است. همانطور که مشاهده می‌کنید، این مولفه تقریباً روند سری را بازسازی کرده است.

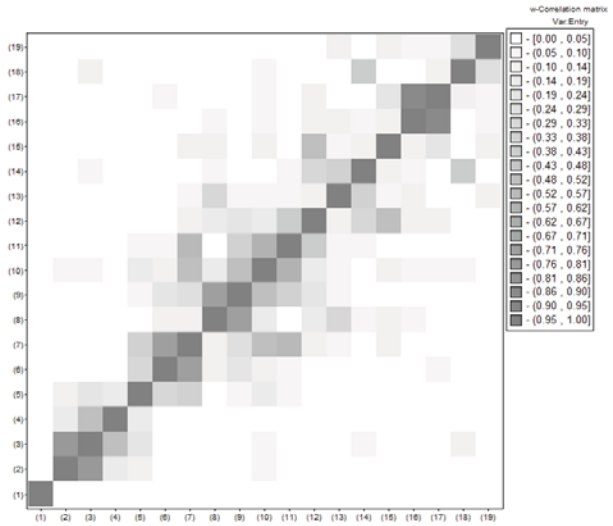


شکل ۵: استخراج مولفه‌ی روند از سری زمانی اتهام کیف زنی به وسیله‌ی ویژه‌سه‌گانه‌ی نخست.

استفاده از ماتریس قدر مطلق مقادیر w-همبستگی‌ها راهی دیگر در تشخیص یک گروه‌بندی مناسب است. شکل (۶)، w-همبستگی‌ها را برای ۱۹ مولفه‌ی بازسازی شده در مقیاس ۲۰ درجه‌ی خاکستری، از سفید تا سیاه متناظر با مقادیر قدرمطلق همبستگی‌ها از صفر تا یک نشان می‌دهد. با توجه به این شکل، تقریباً از حول مولفه‌ی سوم به بعد، همبستگی‌ها با خاکستری‌تر شدن شکل آشکار می‌شود. بنابراین، با توجه به اینکه w - همبستگی میان مولفه‌های ۱-۳ و سایر مولفه‌ها (که این مولفه‌ها در بین خود نیز همبستگی دارند) صفر است، می‌توان مولفه‌های ۱-۳ را در یک گروه (سیگنال) و سایر مولفه‌ها را در گروه دیگر (نوفه) قرار داد.

از سوی دیگر، زوج ۲-۳ با مقادیر تکین تقریباً مساوی متناظر با مولفه‌ی دوره‌ی این سری است. دقت کنید که زوج ۳-۴، به وضوح، مقادیر تکین برابری نداشته و حول این مولفه، افت معنی‌داری در لگاریتم ویژه‌مقدارها رخ می‌دهد که می‌توان آن را نقطه‌ی شروع سطح نوفه دانست. از آنجا که ویژه‌سه‌گانه‌ها به شکل تصویری از ماتریس

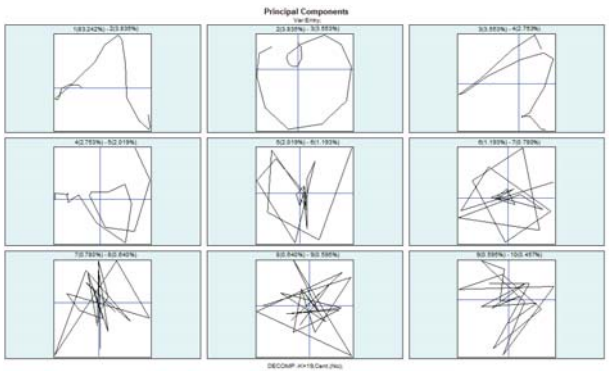
[۱۷] Eigentriple



شکل ۶: ماتریس w-همبستگی برای ۱۹ مولفه‌ی بازسازی شده‌ی سری زمانی اتهام کیف‌زنی.

مسیر X بر روی خود آنها بدست می‌آیند، انتظار داریم که رفتاری شبیه به مولفه‌های اصلی از خود نشان دهند. از این رو، مولفه‌های اصلی (به شکل سری زمانی) نُه ویژه سه‌گانه‌ی نخست در شکل (۷) رسم شده است (شکل دیگر مولفه‌های اصلی مشابه با مولفه‌ی اصلی ۸ به بعد است). مطابق شکل، مولفه‌ی نخست، سهمی حدود ۸۳٫۲۴درصد از کل مولفه‌های اصلی را داشته و سهم مجموع مولفه‌ی دوم و سوم (با الگوی تقریباً برابر) برابر ۷٫۳۸ درصد است. این درحالی است که الگوی مولفه‌های چهارم به بعد کاملا متفاوت از یکدیگر بوده و سهم کمتری نیز از کل تغییرات موجود را در برداشته و در نتیجه، می‌توان آنها را جزو نوفه‌ی سری در نظر گرفت. از این رو، می‌توان نتیجه گرفت که استفاده از سه مولفه‌ی نخست این سری، می‌تواند در تعیین الگوی کلی آن و بدون وارد کردن نوفه به سری بازسازی شده موثر واقع شوند.

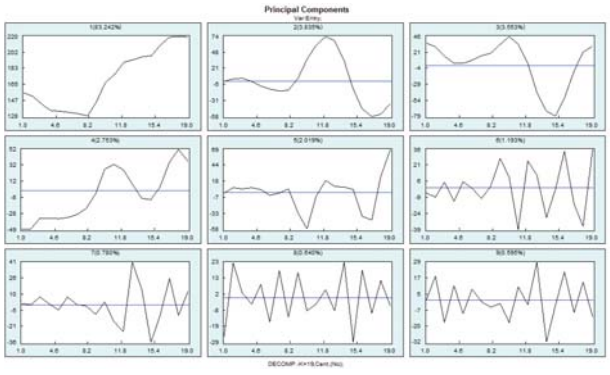
شکل (۸) پراکندگی مولفه‌های اصلی زوج شده را در



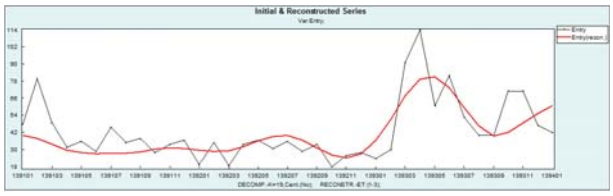
شکل ۷: مولفه‌های اصلی نُه ویژه سه‌گانه‌ی نخست سری زمانی اتهام کیف‌زنی.

سری زمانی نشان می‌دهد؛ این زوج‌ها بر اساس سهم خود در گام SVD (از چپ به راست) مرتب شده‌اند. به کمک این شکل نیز می‌توان به دوره‌ی ویژه سه‌گانه‌ها پی‌برد. اگر ویژه سه‌گانه‌ای دارای دوره باشد، شکل آن یک چند ضلعی منتظم است. با توجه به شکل و در تایید نتیجه‌ی حاصل از شکل (۷) مشاهده می‌کنید که تنها (و البته، تقریباً) می‌توان ویژه سه‌گانه‌ی ۲-۳ را متناظر با دوره‌ی سری زمانی در نظر گرفت. دقت کنید که چنین پراکندگی نامنظمی، ناشی از ماهیت داده‌های تحت مطالعه و تفاوت قابل توجه‌ی الگوی سری در سال‌های تحت بررسی است. (دلایل پراکندگی بالا و چنین الگویی در داده‌ها در بخش پایانی مورد بحث قرار گرفته است.)

بنابراین، با انتخاب ویژه سه‌گانه‌ی نخست (ویژه‌برداری که به کندی تغییر می‌کند) به عنوان مولفه‌ی روند و ۲-۳ به عنوان مولفه‌ی دوره، سری را بازسازی می‌کنیم. شکل (۹) سری بازسازی شده را نمایش می‌دهد. همانطور که مشاهده می‌کنید، SSA در مجموع در شناسایی جهت تغییرات سری موفق عمل کرده و مقادیر سری بازسازی شده به مقادیر واقعی نزدیک هستند.



شکل ۸: پراکندگی متناظر با نُه مولفه‌های اصلی زوج شده‌ی سری زمانی ورودی اتهام کیف‌زنی.



شکل ۹: سری ورودی ماهانه‌ی اتهام کیف‌زنی و سری بازسازی شده به روش SSA.

در ادامه، مقادیر ورودی اتهام را برای یک افق ۲ ماهه (پایان فصل بهار) و به دو روش $RSSA$ و $VSSA$ پیش‌بینی می‌کنیم. (دقت کنید که با توجه به ماهیت داده‌ها و فراوانی و تغییرپذیری مداوم عامل‌های اثرگذار بر آنها، پیش‌بینی افق‌های بلند مدت از اعتبار کمتری برخوردار خواهد بود.) نتایج در جدول (۱) ارایه شده است.

باتوجه به شکل فوق و در پیروی از الگوی سال ۱۳۹۳ در ماه‌های متناظر، انتظار می‌رود که ورودی این اتهام در دو ماه آخر فصل بهار افزایشی باشد.

بر اساس جدول، الگوی افزایشی در مقادیر پیش‌بینی شده به هر دو روش حفظ شده است. همچنین، اختلاف مقادیر پیش‌بینی شده از واقعی در روش برداری نسبت به بازگشتی کمتر است (همانطور که قبلا ذکر شد، روش $VSSA$ نسبت به $RSSA$ محافظه‌کارتر است). لازم به ذکر است که از آنجا که طول دوره‌ی پیش‌بینی کوتاه است، دقت پیش‌بینی را می‌توان با مشاهده‌ی اختلاف پیش‌بینی‌ها از مقادیر واقعی و بدون نیاز به ارائه‌ی معیارهایی همچون میانگین توان دوم خطا بررسی کرد.

۴ بحث و نتیجه‌گیری

آنچه در این مقاله به آن اشاره شد، بررسی امکان و کیفیت بازسازی و در ادامه، پیش‌بینی سری‌های قضایی (در این مقاله، ورودی اتهام) به روش SSA بود. در این میان، ارایه‌ی نکاتی در خصوص ماهیت سری‌های قضایی و نقش آن در نتایج بدست آمده، خالی از لطف نیست.

آنچه مسلم است، تغییرپذیری شدید و عدم پیروی از یک الگوی مشخص در دوره‌های تکراری در شهر یا استان تهران است که بخشی از آن را می‌توان به دلیل ماهیت ذاتی این داده‌ها و تاثیرپذیری شدید آنها از وضعیت اجتماعی، سیاسی، اقتصادی و ... در جامعه و بخش دیگر را در وجود خطا و مشکلات موجود در ثبت این داده‌ها یافت. یکی از مهمترین و فراوان‌ترین اشتباه‌های کاربری، عدم ثبت یک پرونده در ماه ورودی و اعمال آن در ماه بعد بوده که این کار منجر به ایجاد مغایرت در تعداد داده‌های

جدول ۱: تعداد ورودی اتهام کیف‌زنی و مقادیر پیش‌بینی شده‌ی آن به تفکیک روش بازگشتی ($RSSA$) و برداری ($VSSA$) در اردیبهشت و خرداد ماه سال ۱۳۹۴ در شهر تهران.

ماه	مقدار واقعی	$RSSA$	$VSSA$
اردیبهشت	۵۲	۶۹	۶۳
خرداد	۶۱	۷۷	۶۹

تثبیت شده در دو ماه مذکور می‌شود. از این رو، بدیهی است که ثبت و سعی در رفع مشکلات آن، نقش اساسی در بهبود و افزایش دقت بازسازی و پیش‌بینی سری‌های تحت مطالعه (از جمله به روش SSA) در این حوزه خواهد داشت.

در خصوص توانمندی روش SSA نیز در این مقاله به استناد مطالعه‌های صورت گرفته، طول پنجره‌ای برابر میانه‌ی داده‌ها مورد استفاده قرار گرفت؛ با این حال، با توجه به تفاوت الگویی موجود در این داده‌ها در مقاطع زمانی مختلف، می‌توان امکان استفاده و عملکرد دیگر مقادیر انتخابی برای این پارامتر را با تعریف و اعمال قیدهایی برای آن مورد مطالعه قرار داد.

از سوی دیگر، از دیگر ویرایش‌های قدرتمند SSA (همچون SSA بر مبنای برآوردگر کمترین واریانس و نظریه‌ی پرشیدگی) به منظور افزایش دقت و بهبود نتایج بازسازی و پیش‌بینی نیز می‌توان بهره جست.

مطالعه بر روی بحث دقت پیش‌بینی روش‌های مختلف برپایه‌ی SSA (که منجر به استفاده از داده‌های بیشتر و اختصاص وزن‌های بیشتر به داده‌های به روزتر می‌شود) و همچنین، بررسی و مقایسه‌ی دقت آنها از دیگر جنبه‌های ممکن در جهت بسط و توسیع روش SSA در حوزه‌ی داده‌های قضایی خواهد بود.

مراجع

[۱] آل‌حسینی، فاطمه سادات. (۱۳۹۱). مقایسه روش تحلیل مجموعه مقادیر تکین در پیش بینی نرخ مرگ و میر با روش‌هایی از خانواده لی-کارتر. پایان نامه کارشناسی ارشد، دانشگاه شهیدبهشتی، دانشکده علوم ریاضی.

[۲] آل‌حسینی، فاطمه سادات. (۱۳۹۴). پیش‌بینی سری‌های زمانی قضایی با استفاده از روش تحلیل مجموعه‌ی مقادیر تکین. نخستین همایش دانشجویی آمار، دانشگاه بوعلی‌سینا، همدان.

[۳] محمودوند، رحیم. (۱۳۹۱). گسترش برخی از مبانی نظری روش تحلیل مجموعه‌ی مقادیر تکین. دانشگاه شهید بهشتی، دانشکده علوم ریاضی، رساله‌ی دکتری.

[۴] نجاری، نادر. (۱۳۹۰). مقدمه‌ای بر تحلیل تکینی طیفی (SSA) و کاربردهای آن. دانشگاه شهید بهشتی، دانشکده علوم ریاضی، پایان نامه کارشناسی ارشد.

[5] Broomhead, D.S. and King, G.P. (1986). Extracting qualitative dynamics from experimental data. Physica D, 20, 217-236.

[6] Broomhead, D.S. and King, G.P. (1986). On the qualitative analysis of experimental dynamical systems. Nonlinear Phenomena and Chaos, Adam Hilger, Bristol, 12, 113-144.

[7] Danilov, D. (1997). Principal components in time series forecast. Journal of Computational and Graphical Statistics, 6 (1), 112-121.

[8] Danilov, D. and Zhigljavsky, A. (1997). Principal components of time series: the 'Caterpillar' method.

University of St. Petersburg Press (In Russian).

[9] Golyandina, N., Nekrutkin, V. and Zhigljavsky, A. (2001). Analysis of time series structure: SSA and related techniques. Chapman Hall/CRC, New York - London.

[10] Hassani, H., Mahmoudvand, R. and Zokaei, M. (2011). Separability and window length in singular spectrum analysis. Comptes Rendus Mathematique, 349 (17), 987–990.

[11] Hassani, H., Mahmoudvand, R., Zokaei, M. and Ghodsi, M. (2012). On the separability between signal and noise in singular spectrum analysis. Fluctuation and Noise Letters. Forthcomming.

[12] Mahmoudvand, R., Najari, N. and Zokaei, M. (2013). On the parameters for reconstruction and forecasting in the singular spectrum analysis. Communication in Statistics: Simulations and Computations, 42, 860-870.

[13] Pepelyshev, A. (2010). Comparison of recurrent and vector forecasting. In proceeding of UK-China SSA workshop, Cardiff, UK.