مجله تخصصی اپیدمیولوژی ایران؛ ۱۳۹٤، دوره ۱۱، شماره ۱: صفحات ۱۰۲-۹٤.

مقاله مروری

چگونه یک مدل مناسب برای داده‌های سری زمانی انتخاب کنیم؟

جعفر حسن‌زاده[۱]، فرید نجفی[۲]، مهدی مرادی‌نظر[۳]

[۱] دانشیار گروه اپیدمیولوژی، دانشکده بهداشت و تغذیه، دانشگاه علوم پزشکی شیراز، ایران

[۲] دانشیار گروه اپیدمیولوژی، مرکز تحقیقات عوامل محیطی مؤثر بر سلامت، دانشکده بهداشت، دانشگاه علوم پزشکی کرمانشاه، ایران

[۳] دانشجوی دکترای اپیدمیولوژی، مرکز تحقیقات عوامل محیطی مؤثر بر سلامت، دانشکده بهداشت، دانشگاه علوم پزشکی کرمانشاه، ایران

نویسنده رابط: مهدی مرادی نظر، نشانی: مرکز تحقیقات عوامل محیطی مؤثر بر سلامت، دانشکده بهداشت، دانشگاه علوم پزشکی کرمانشاه، تلفن: ۰۹۱۸۸۵۸۶۵۲۰، آدرس پست الکترونیک: m.moradinazar@gmail.com

تاریخ دریافت: ۹۲/۰۹/۱۲؛ پذیرش: ۹۳/۰۷/۰۵

مقدمه و اهداف: سری‌های زمانی مجموعه‌ای از مشاهدات هستند که برحسب زمان مرتب‌شده است. هدف اصلی در برپا کردن یک سری زمانی معمولاً پیش‌بینی مقادیر آینده می‌باشد. نخستین گام در سری‌های زمانی، رسم نمودار داده‌ها است. با استفاده از رسم نمودار می‌توان اطلاعات کلی ازجمله روند صعودی یا نزولی، وجود الگوی فصلی، روند دوره‌ای و وجود داده‌های پرت درداده‌ها را تشخیص داد. پس از رسم نمودار برای این‌که پیش‌بینی مناسبی وجود داشته باشد، باید داده‌ها را ایستا کرد. می‌توان داده‌ها را با استفاده از تفاضل گیری یا تجزیه به مؤلفه‌های تشکیل‌دهنده‌ی آن، ایستا نمود. پس از ایستا کردن داده‌ها می‌توان با استفاده از نمودار نگاره مرتبه میانگین متحرک و مرتبه اتو رگرسیون مدل را شناسایی نمود. لازم است پارامترها به‌دست‌آمده را با استفاده از آزمون T ازنظر معنی‌داری موردبررسی قرارداد. در صورت معنی‌دار بودن و عدم وابستگی در باقی‌مانده می‌توان پیش‌بینی مناسبی با کمک داده‌های گذشته انجام داد، هم‌چنین مقادیر پیش‌بینی‌شده را می‌توان با استفاده از میانگین مطلق درصد خطا ارزیابی نمود.

واژگان کلیدی: سری‌های زمانی، شناسایی مدل، ایستا کردن، پیش‌بینی

مقدمه

در سال‌های اخیر با ثبت منظم داده‌ها و به وجود آمدن بانک‌های مختلف اطلاعاتی در سراسر دنیا تمایل به استفاده ازسری‌های زمانی افزایش‌یافته است. سری‌های زمانی که زیرشاخه‌ای از آمار و احتمالات می‌باشد، مجموعه‌ای از مشاهدات است که برحسب زمان مرتب‌شده است (۱). سری‌های زمانی از دیرباز در بیش‌تر رشته‌ها مانند زمین‌شناسی، اقتصاد، مهندسی ارتباطات و غیره کاربرد فراوانی داشته است. امروزه با آگاه‌تر شدن پژوهشگران از کاربردهای سری‌های زمانی به‌طور وسیع در خدمات بهداشتی و درمانی برای پیش‌بینی طغیان بیماری‌ها، تعداد بیماران مراجعه‌کننده، تعداد پرسنل موردنیاز در بخش‌های مختلف درمانی، مورداستفاده قرار می‌گیرد (۲). تفاوتی که بین سری‌های زمانی با سایر روش‌های مدل‌سازی ازجمله رگرسیون وجود دارد، این است که در سری‌های زمانی با استفاده از داده‌های قبلی مقادیر آینده را پیش‌بینی می‌کنند، درحالی‌که درروش‌های مدل‌سازی اغلب با استفاده از متغیرهای دیگر سعی می‌شود متغیر موردنظر پیش‌بینی شود. به همین دلیل معمولاً قدرت سری‌های زمانی در پیش‌بینی کم‌تر است، اما به دلیل این‌که به اطلاعات جانبی کم‌تری نیاز دارد، تمایل به استفاده از آن زیاد می‌باشد، در این مقاله سعی بر آن است که مراحل برپا کردن یک مدل سری زمانی را به‌صورت مختصر و به زبان ساده با استفاده از داده‌های مرگ‌ومیر ناشی از تصادفات در استان کرمانشاه توضیح داده شود.

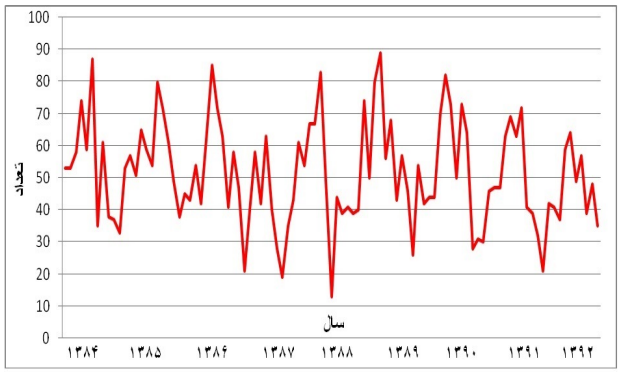
نوع داده‌ها و هدف از تجزیه‌وتحلیل سری‌های زمانی

داده‌ها در سری زمانی مجموعه‌ای از مشاهده‌های متوالی می‌باشد که به‌صورت منظم با فواصل زمانی یکسان مرتب‌شده است (۳). زمان در سری زمانی بسته به نوع متغیر موردبررسی و هدف می‌تواند روز، هفته، ماه یا سال باشد. در نمودار شماره ۱ روند مرگ‌ومیر ناشی از تصادفات در استان کرمانشاه با فواصل زمانی یک ماه طی سال‌های ۹۲-۱۳۸۴ نشان داده‌شده است. اگر داده‌ها از یکدیگر مستقل و تصادفی باشند (مقادیر گذشته تأثیری روی مقادیر حال و آینده نداشته باشد) پیش‌بینی با استفاده از این داده‌ها امکان‌پذیر نمی‌باشد. به این داده‌ها که دارای دنباله تصادفی، مستقل و هم توزیع با میانگین صفر هستند، اغتشاش خالص[۱] می‌گویند (۴)، اما اگر داده‌ها به یکدیگر وابسته و غیر

[۱] White noise

تصادفی مانند تعداد تصادفات منجر به فوت در یک خیابان یا شهر، تعداد بیماران مراجعه‌کننده به بیمارستان‌های تحت پوشش یک دانشگاه یا متوسط دمای روزانه هوای استان- باشند؛ در این داده‌ها اغلب بین بازه‌های زمانی (lag) هم‌بستگی وجود دارد، در این داده‌ها که مقادیر حال وابسته به مقادیر گذشته می‌باشد، می‌توان با استفاده از مقادیر گذشته مقادیر آینده را پیش‌بینی نمود. اگر یک سری زمانی را کاملاً بتوان پیش‌بینی کرد، آن را سری زمانی غیر تصادفی می‌گویند، اما معمولاً نمی‌توان یک داده را به‌طور کامل پیش‌بینی نمود (۵).

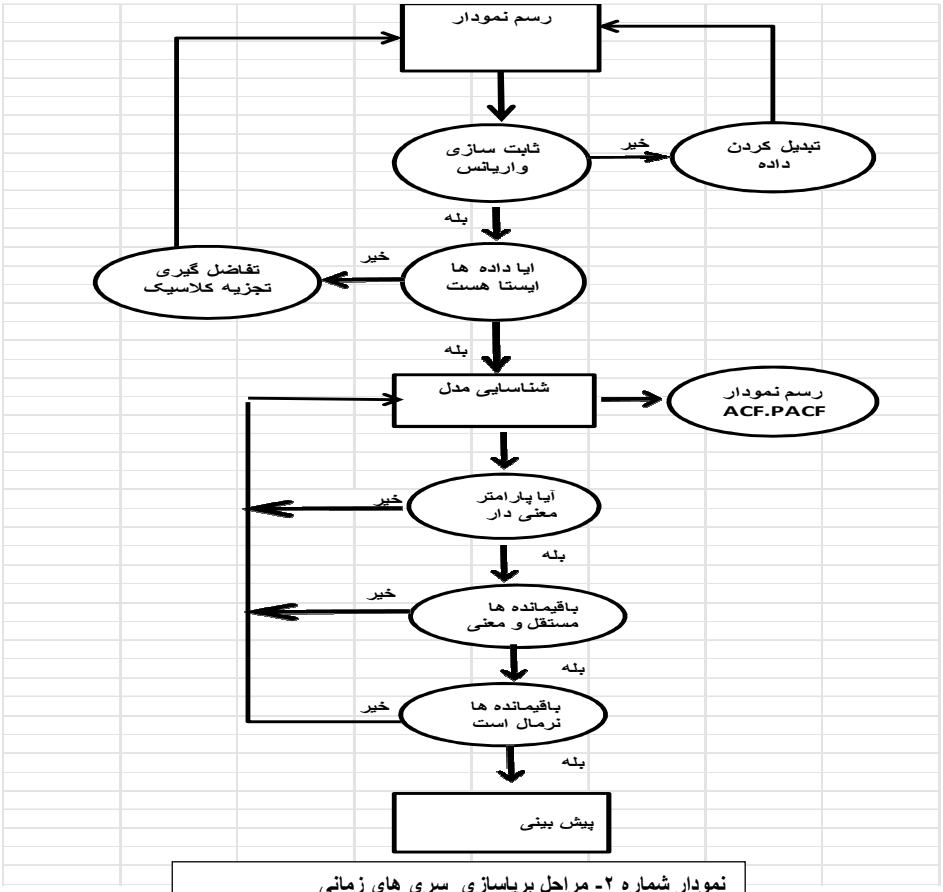
هدف اصلی از تجزیه‌وتحلیل سری زمانی معمولاً پیش‌بینی مقادیر آینده می‌باشد، سه هدف دیگر از بررسی سری‌های زمانی که معمولاً نسبت به پیش‌بینی کم‌تر موردتوجه قرار می‌گیرد، توصیف (Explanation) تشریح (Description) و کنترل (Control) می‌باشد. توصیف داده‌ها معمولاً با استفاده از رسم نمودار صورت می‌گیرد، با استفاده از رسم نمودار می‌توان توصیف ساده‌ای از خواص داده‌ها مانند صعودی یا نزولی بودن روند داده‌ها، وجود الگوی فصلی و وجود تغییر ناگهانی یا آرام در روند سری‌های زمانی انجام داد. توصیف یا رسم نمودار مرحله‌ی نخست در تجزیه و تحلیل سری‌های زمانی است. هر چند رسم نمودار بسیار مقدماتی و شهودی می‌باشد، اما در شناسایی مدل نهایی می‌تواند بسیار کمک کننده باشد، برای مثال از روی رسم نمودار شماره ۱ می‌توان پی برد که مرگ‌ومیر ناشی از تصادفات در استان کرمانشاه دارای الگوی فصلی است که در فصل تابستان بیش‌ترین و در زمستان کم‌ترین تعداد موارد مرگ در اثر تصادفات رخ می‌دهد و به‌طور ساده می‌توان گفت روند مرگ‌ومیر ناشی از تصادفات هر چند در سه سال اخیر کاهشی بود، اما تقریباً دارای روند ثابتی است. دومین هدف از تجزیه و تحلیل سری زمانی تشریح است که از تغییرات در یک سری برای بیان تغییرات در سری دیگر استفاده می‌شود. سومین هدف کنترل می‌باشد که رابطه‌ی بسیار نزدیک با پیش‌بینی دارد، یکی از کاربردهای کنترل استفاده در مرکز مبارزه با بیماری‌ها برای دادن هشدار قبل از طغیان بیماری می‌باشد (۱،۳).



نمودار شماره ۱- روندمرگ‌ومیر ناشی از تصادفات در استان کرمانشاه طی سال‌های ۹۲-۱۳۸٤

مراحل برپا کردن مدل مناسب

برای انتخاب مدل مناسب برای داده‌ها از الگویی که در سال ۱۹۷۰ توسط باکس وجنگیز (Box&Jenkins) ابداع شد و به‌سرعت نیز تکامل یافت؛ استفاده می‌شود که اساس آن در ۳ مرحله تعیین مدل، برآورد و آزمون پارامترها و کاربرد مدل برای پیش‌بینی می‌باشد (۵). همان‌طور که نمودار شماره ۲ نشان می‌دهد، این مدل یک مدل با مراحل تکراری است که تا زمانی که مدل مناسب پیدا نشود؛ مراحل تکرار می‌گردد (نمودار شماره ۲).



نمودار شماره ۲- مراحل برپاسازی سری های زمانی

ایستا کردن داده‌ها

پس از رسم نمودار داده‌ها و به دست آوردن یک ویژگی ابتدایی از داده‌ها، اگر داده‌ها دارای روند صعودی یا نزولی یا الگوی فصلی مانند داده‌های مرگ‌ومیر ناشی از تصادفات در استان کرمانشاه باشد، داده‌ها باید ایستا شود، منظور از ایستا کردن داده ثابت‌سازی واریانس و میانگین در طول زمان می‌باشد، برای ایستا کردن داده‌ها می‌توان از دو روش تفاضل گیری و تجزیه کلاسیکی به اجزای تشکیل‌دهنده (مؤلفه روند، مؤلفه فصلی و مؤلفه باقیمانده‌های) استفاده کرد. درروش تفاضل گیری هر مشاهده را از مشاهده قبلی کم می‌کنند، برای مثال جدول شماره ۱ تعداد تصادفات منجر به فوت در استان کرمانشاه در هرماه آورده شده است. در تفاضل گیری مرتبه اول برای ماه سوم عدد ۷۵ مربوط به ماه سوم را از عدد ۷۰ مربوط به ماه قبل کسر نمود. لازم به ذکر است در هر بار تفاضل گیری یکی از تعداد بازه‌های زمانی (lag) کاسته می‌شود؛ بنابراین اگر از یک داده سه بار تفاضل گیری انجام شود، سه بازه زمانی کاسته می‌شود (۶).

جدول شماره ۱- تعداد تصادفات منجر به مرگ در استان کرمانشاه و تفاضل گیری مرتبه اول و دوم

ماه	تعداد فوت	تفاضل گیری مرتبه اول	تفاضل گیری مرتبه دوم
۱	۶۰		
۲	۷۰	۷۰-۶۰=۱۰	
۳	۷۵	۷۵-۷۰=۵	۵-۱۰=-۵
۴	۷۲	۷۲-۷۵=-۳	-۳-۵= -۸
۵	۸۰	۸۰-۷۲=۴	۴-(-۳)=۷
۶	۹۶	۹۶-۸۰=۱۶	۱۶-۴=۱۲
۷	۱۰۰	۱۰۰-۹۶=۴	۴-۱۶=-۱۲
۸	۸۷	۸۷-۱۰۰=-۱۳	۱۳-۴=۹

اگر داده‌ها فقط دارای روند باشند، تفاضل گیری مرتبه‌ای اول کافی است، اما اگر داده‌ها دارای الگوی فصلی باشند مانند داده‌های مرگ‌ومیر ناشی از تصادفات در استان کرمانشاه یا واریانس داده‌ها در طول زمان تغییر کند، تفاضل گیری مرتبه دوم معمولاً نیاز می‌باشد. اگر از یک تابع ایستا تفاضل گیری انجام شود، داده حاصل نیز ایستا است، اما تفاضل گیری غیرضروری علاوه بر این‌که تعداد بازه‌های موردبررسی را کم می‌کند، باعث پیچیده شدن داده‌ها و هم‌بستگی کاذب بین داده‌ها می‌شود، اما درمجموع تفاضل گیری غیرضروری بهتر از تفاضل نگرفتن ضروری است (۱).

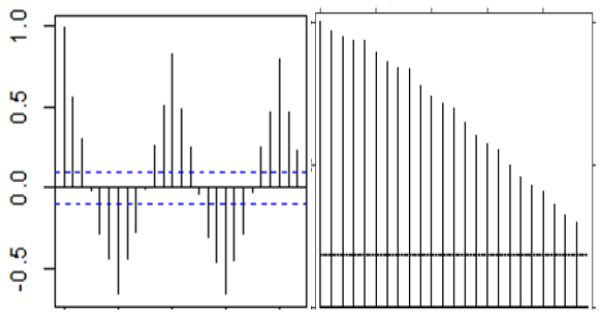
ثابت‌سازی واریانس نسبت به میانگین بسیار مشکل‌تر و پیچیده‌تر است، برخی اوقات ثابت‌سازی واریانس فقط با انجام تفاضل گیری امکان‌پذیر نیست؛ بنابراین زمانی که واریانس داده‌ها در طول زمان بسیارمتغیر باشد؛ ابتدا داده‌ها تبدیل می‌شود، برای مثلاً اگر سطح سری به‌صورت نمایی تغییر کند؛ باید از داده‌ها لگاریتم گرفته شود. اگر داده‌ها شبیه توزیع پوآسن باشد، باید ریشه دوم آن‌ها برآورد شود (۱). ازآنجایی‌که احتمال دارد پس از تفاضل گیری مرتبه اول یا دوم بعضی از داده‌ها منفی شود؛ بنابراین باید پیش از تفاضل گیری تبدیل انجام شود. تفاضل گیری به‌منظور ایستا کردن داده‌ها که به‌عنوان یک روش شهودی و غیراستدلالی شناخته می‌شود، پایه و اساس سری‌های زمانی می‌باشد (۷). در داده‌های مرگ‌ومیر ناشی از تصادفات در استان کرمانشاه هرچند دارای الگوی فصلی و ناایستایی در واریانس است، اما با توجه به این‌که ناایستایی در واریانس کم است، احتمالاً نیاز به تبدیل وجود ندارد.

روش دیگری برای ایستا کردن داده‌ها، تجزیه یک سری زمانی به مؤلفه‌های تشکیل‌دهنده (Seasonal decomposition) آن می‌باشد، این روش مشابه تفاضل گیری است؛ با این تفاوت که نیاز است مؤلفه‌های تشکیل‌دهنده یک سری زمانی-الگوهای فصلی، دوره‌ای و روند سری- به‌درستی شناسایی شود. برای شناسایی مؤلفه‌های تشکیل‌دهنده یک سری زمانی می‌توان از تابع خود هم‌بستگی و خود هم‌بستگی جزئی کمک گرفت، برای مثال اگر تابع خود هم‌بستگی (ACF)[۱] داده‌های مرگ‌ومیر ناشی از تصادفات در استان کرمانشاه که دارای الگوی فصلی می‌باشد؛ رسم شود، در وقفه‌های ۴، ۸، ۱۲، ... روند کاهشی نشان می‌دهد که احتمال معنی‌دار شدن در وقفه‌های یادشده هست و سایر وقفه‌ها احتمالاً صفر خواهد بود، اما اگر داده‌ها دارای الگوی ماهیانه باشد، به‌جای وقفه‌های یادشده، وقفه‌های ۱۲، ۲۴، ۳۶، ... دارای این ویژگی می‌باشند (۹-۷).

شناسایی الگو

پس از تفاضل گیری و یا تجربه داده‌ها به مؤلفه‌های تشکیل‌دهنده آن، برای اطمینان از ایستا شدن داده‌ها یا نیاز به تفاضل گیری مرتبه بالاتر یا استفاده از تبادل دیگر برای ثابت‌سازی واریانس وجود دارد و از طرف دیگر برای بررسی این که آیا پس از ایستا کردن داده‌ها، می‌توان الگویی برای افزایش توان پیش‌بینی از داخل باقی‌مانده‌ها[۲] که روند صعودی یا نزولی یا الگوی فصلی را از خود نشان نمی‌دهد، پیدا نمود. برای این کار از نمودار نگاره[۳] استفاده می‌شود. نمودار نگاره از دو نمودار تابع خود هم‌بستگی (ACF) و خود هم‌بستگی جزئی (PACF)[۴] تشکیل‌شده است. از تابع خود هم‌بستگی برای نشان دادن مرتبه اتورگرسیو (یعنی داده باقی‌مانده حال تحت تأثیر چند داده قبل قرار می‌گیرد) و از تابع خود هم‌بستگی جزئی برای نشان دادن مرتبه میانگین متحرک (یعنی مجموع چند داده باقی‌مانده را می‌توان صفر در نظر گرفت) استفاده می‌شود (۹).

تابع خود هم‌بستگی که روی باقی‌مانده‌ها انجام می‌شود، اگر دارای الگوی تدریجی نزولی و یا نمایی بود (نمودار شماره ۳) بیان‌گر از نا ایستا بودن مدل و نیاز به تفاضل گیری مجدد و یا تغییر در تفاضل گیری‌های صورت گرفته و یا شاید هم تبدیل دارد.



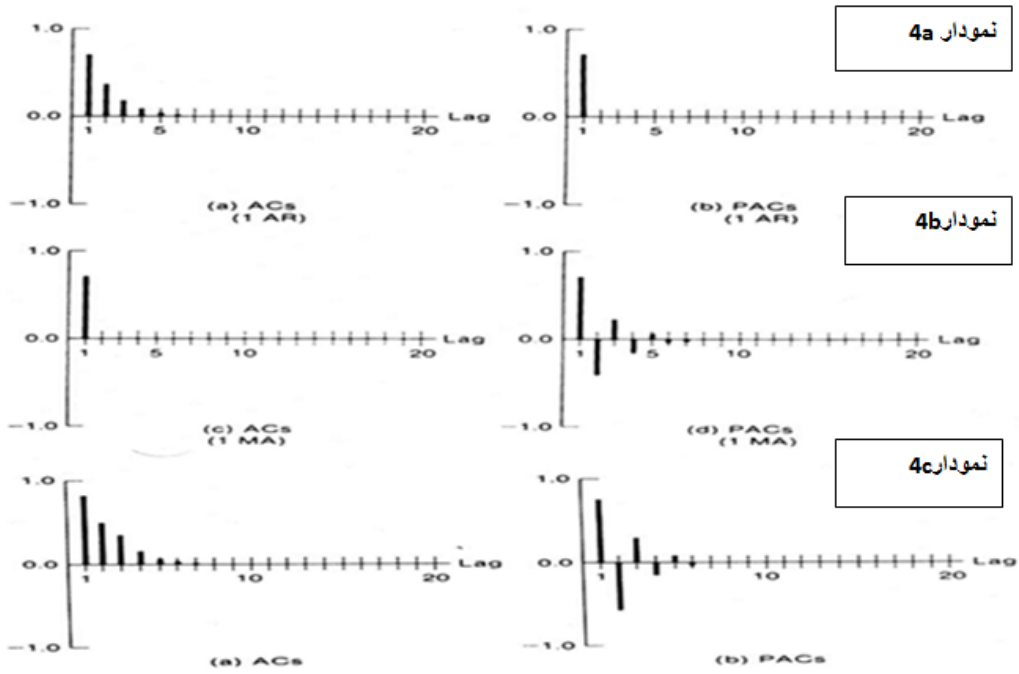
نمودار شماره۳- روند نزولی تابع ACF

[۱] Autocorrelation Function; ACF

[۲] Residual
[۳] Correlogram
[۴] Partial Autocorrelation Function; PACF

درصورتی‌که تابع خود هم‌بستگی دارای روند نزولی به سمت صفر باشد و تابع خود هم‌بستگی جزئی دارای وقفه‌های معنی‌دار در ابتدای تابع باشد (نمودار 4a) بیان‌گر از یک الگوی اتورگرسیو دارد با مرتبه AR(P) دارد که مرتبه (P) آن بستگی به تعداد وقفه‌های معنی‌دار دارد، اما اگر تابع خود هم‌بستگی جزئی دارای روند نزولی به سمت صفر و تابع خود هم‌بستگی دارای وقفه‌های معنی‌دار در ابتدای تابع باشد (شکل 4b) نشان می‌دهد مدل دارای یک الگوی میانگین متحرک با مرتبه (q) MA است (برعکس تابع AR) اما اگر هردو تابع PACF و ACF دارای روند نزولی به سمت صفر باشد (نمودار شماره 4c)

بیان‌گر از یک روند مرکب ARMA(p,q) دارد (۸). در صورت نا ایستا بودن داده‌ها و نیاز به تفاضل گیری، مدل از ARMA(p,q) به ARIMA(p,d,q) تبدیل می‌شود، که به ازای هر تفاضل گیری یک واحد به مرتبه d اضافه می‌گردد. شایان ذکر است در صورت وجود الگوی فصلی تابع خود هم‌بستگی و تابع خود هم‌بستگی جزئی یک روند متناوب را نشان می‌دهد (۱۰).



نمودار شماره 4a تا4c: الگوهای مختلف ACF و PACF

شناسایی الگوی مناسب برای داده‌ها کاری بسیار پیچیده و نیاز به تجربه دارد. چندین روش برای شناسایی الگوی مناسب برای داده‌ها وجود دارد، که بهترین آن‌ها شناسایی الگوی مناسب با استفاده از نمودار سری زمانی انجام می‌گیرد که این روش هرچند شهودی می‌باشد، اما بسیار کمک کننده می‌باشد (۱۱).

انتخاب مدل

پیش از انتخاب مدل مناسب، لازم است چند نکته بیان گردد، نخست این‌که هر اندازه تعداد پارامترهای مورداستفاده کم‌تر باشد، مدل مناسب‌تر است؛ برای مثال در شرایط تقریباً برابر مدل اتورگرسیو ۱ بهتر از مدل اتورگرسیو ۲ می‌باشد، به این ویژگی اصل صرفه‌جویی یا امساک می‌گویند و دوم ممکن است مدل و تعداد پارامتری که از روی تابع نگاره به دست آورده می‌شود، بهترین انتخاب نباشد. برای توضیح بیش‌تر فرض کنید مدلی که از روی نمودار نگاره پیداشده است، MA(2) باشد (تابع خود هم‌بستگی دارای روند نزولی به سمت صفر، تابع خود هم‌بستگی جزئی تا وقف دوم معنی‌دار باشد)، اما بهترین مدل یک مدل ARMA(۱,۱) باشد و سوم این‌که در یک سری زمانی ممکن است چندین الگوی مناسب وجود داشته باشد. بنابراین نیاز است الگوهای مختلف به‌دست‌آمده با یکدیگر مقایسه شود و با در نظر گرفتن چندین معیار نسبت به انتخاب مدل اصلاح اقدام شود (۴،۱۰).

پس از شناسایی الگوی و انتخاب پارامتری مناسب، پارامتر موردنظر را ازنظر معنی‌داری با آزمون t مورد ارزشیابی قرار داده

شد، که در صورت معنی‌داری، بیان‌گر از مؤثر بودن پارامترهای انتخاب‌شده دارد، اگر معنی‌دار نبود، باید پارامتر را از مدل حذف نمود؛ و پارامترهای دیگر را مورد ارزیابی قرارداد. پس از انتخاب پارامتر معنی‌دار، نیاز است باقی‌مانده‌های مدل را ازنظر استقلال و معنی‌داری موردبررسی قرارداد که برای استقلال باقی‌مانده‌ها از آزمون پورت مانتو که مقدار به‌دست‌آمده از فرمول

$$Q^* = n'(n'+2)\sum_{l=1}^{k}(n'-l)^{-1}r_e^2(\bar{a})$$

که در آن a درجه آزادی re اتوکواریانس بین داده‌ها با مقدار مربع کای با درجه آزادی L-AR(p) -MA(q) که L تعداد وقفه‌ها می‌باشد، موردبررسی قرارداده شود، که اگر Q تست پورت مانتو کوچک‌تر باشد، استقلال باقی‌مانده‌ها تأیید می‌گردد. پس از تأیید استقلال داده‌های باقی‌مانده، باقی‌مانده را ازنظر نرمال بودن مورد ارزیابی قرارداده شود. برای نرمال بودن می‌توان از آزمون‌های مختلف مانند کولموگروف اسمیرنوف، مربع کای و هیستوگرام باقی‌مانده که باید دارای توزیع با میانگین صفر و واریانس یک باشد، استفاده کرد (۱۴-۱۲) اگر استقلال و نرمال بودن داده‌ها تأیید شد، مدل مناسب می‌باشد، اما اگر باقی‌مانده‌ها وابسته یا توزیع باقی‌مانده‌ها نرمال نبود، نیاز است به دنبال مدل دیگری بود (۴،۱۵).

پیش‌بینی

پس از شناسایی الگو و تعیین پارامترهای مناسب می‌توان مقادیر آینده را پیش‌بینی نمود. در هنگام پیش‌بینی باید در نظر داشت که هراندازه تعداد وقفه‌های پیش‌بینی‌شده بیش‌تر شود، احتمال خطا افزایش می‌یابد. درمجموع بهتر است تعداد وقفه‌هایی که پیش‌بینی می‌شود، کم‌تر از یک‌سوم وقفه‌های موردبررسی باشد (۶). برای اندازه‌گیری خطای پیش‌بینی روش‌های متفاوتی وجود دارد، که مهم‌ترین آن‌ها عبارت است از

a) میانگین مطلق انحراف (Mean Absolute Deviation (MAD))

$$MAD = \frac{\sum_{t=1}^{n}|y_t - \hat{y}_t|}{n}$$

در این روش yt مقدار واقعی، Ŷt مقدار برآورد شده و n تعداد مشاهده‌ها می‌باشد. میانگین مطلق انحراف تصور دقیقی از مقدار خطای پیش‌بینی به پژوهشگر می‌دهد.

b) مجذور میانگین خطاها (MSD)[۱]

$$MSD = \frac{\sum_{t=1}^{n}|y_t - \hat{y}_t|^2}{n}$$

این روش نسبت به MAD حساسیت بیش‌تری به خطای پیش‌بینی دارد. در این روش yt مقدار واقعی، Ŷt مقدار برآورد شده و n تعداد مشاهدات می‌باشد، استفاده نمود (۸، ۱۶و۱۷)

C) میانگین مطلق درصد خطا (MAPE)[۲]

$$MAPE = \frac{\sum|(yt - \hat{y}t)|/yt}{n} \times 100 \qquad (yt \neq 0)$$

که در آن yt مقدار واقعی، Ŷt مقدار برآورد شده و n تعداد مشاهده‌ها می‌باشد، که خطای پیش‌بینی را به‌صورت درصد نشان می‌دهد. این شاخص نسبت به دو شاخص بالا معیار مناسب‌تری می‌باشد، یکی به این دلیل که تفاوت بین مقدار واقعی و مقدار برآور شده را بر مقادیر واقعی تقسیم می‌کند؛ درنتیجه تحت تأثیر اندازه عددی مقدار موردبررسی قرار نمی‌گیرد و دوم این‌که می‌توان میزان دقت دو مدل که از دو داده مختلف مثلاً تصادفات و خودکشی به‌دست‌آمده را باهم مقایسه نمود (۱۸).

d) شاخص AIC

$$AIC = 2k - 2\ln(L)$$

یکی از بهترین روش‌ها برای اندازه‌گیری مقدار خطا، آماره AIC می‌باشد که پارامتر K بیانگر تعداد بازه‌های زمانی موردنظر برای پیش‌بینی و L مقدار حداکثر درستنمایی[۳] است.

E) شاخص شوارتز[۴]

$$SBC = -2 \cdot \ln L + k\ln(n)$$

شاخص شوارتز یا شاخص بیزین[۵] از دیگر روش‌هایی است که می‌توان برای سنجش مقدار خطا استفاده کرد. در این روش L نشان‌دهنده‌ی حداکثر درست نمایی، K تعداد بازه‌های زمانی پیش‌بینی‌شده و n تعداد بازه‌های زمانی مشاهده‌شده است.

---

[۱] Mean Squared Deviation; MSD
[۲] Mean Absolute Percentage Error; MAPE
[۳] Maximum Likelihood
[۴] Schwartz-Bayesian criteria
[۵] Bayesian information criteria

هرچقدر مقدار شاخص‌های ارائه‌شده در بالا کوچک‌تر باشد، نشان‌دهنده مدل بهتر است. این شاخص‌ها معمولاً رابطه‌ی بسیار نزدیکی با یکدیگر دارند. مثلاً اگر یک مدل نسبت به مدل دیگری شاخص AIC کوچک‌تری داشته باشد، احتمالاً مجذور میانگین خطاها و شاخص شوارتز کوچک‌تری نیز خواهد داشت (۱۵). در جدول شماره ۲ شاخص‌های مختلف برازش و پیش‌بینی مدل با استفاده از داده‌های مرگ‌ومیر ناشی از تصادفات در استان کرمانشاه با یکدیگرمقایسه شدند. همان‌طور که قبلاً گفته شد رسم نمودار در شناسایی مدل مناسب می‌تواند بسیار کمک کننده باشد.از روی نمودار مشخص شد که داده‌ها دارای الگوی فصلی است. شاخص‌های برازش نیز تأیید کردند که مدل‌های SARIMA بهتر از مدل‌های ARIMA هستند. از سوی دیگر هم‌بستگی بسیار زیادی بین شاخص‌های مختلف وجود دارد، به‌طوری که مدل SARIMA (۱,۱,۱)*(۰,۱,۱)۱۲ که شاخص AIC کوچک‌تری نسبت به سایر مدل‌ها دارد، شاخص شوارتز و میانگین مطلق درصد خطا و میانگین مطلق انحراف کم‌تری نیز دارد. علاوه بر این، شاخص‌ها، شاخص‌های دیگری مانند Stationary R2، maximum likelihood و Hannan–Quinn information criterion (HQIC) وجود دارد، که در شناسایی مدل می‌تواند بسیار کمک کننده باشد، اما شاخص‌های گفته شده پرکاربردترین شاخص‌ها برای ارزشیابی مدل می‌باشد.

جدول شماره۲- شاخص‌ها و مدل‌های مختلف روی داده‌های مرگ‌ومیر ناشی از تصادفات در استان کرمانشاه طی سال‌های ۹۲-۱۳۸٤

مدل	AIC	SBC	MAD	MSD	MAPE
ARIMA (1,1,1)	۸۷۳	۸۸۳	۱۳	۲۵۳	۳۰
ARIMA (2,1,1)	۸۸۵	۸۹۵	۱۳	۲۸۸	۳۱
ARIMA (1,2,1)	۸۸۱	۸۸۹	۱۴	۳۳۴	۳۳
ARIMA (2,2,1)	۸۹۳	۹۰۱	۱۵	۳۷۲	۳۳
SARIMA (1,1,1)*(1,0,1)12	۸۰۱	۸۱۶	۱۲	۲۴۹	۲۸
SARIMA (1,1,1)*(0,1,1)12	۷۹۷	۸۰۷	۱۱	۲۴۹	۲۸
SARIMA (1,2,1)*(0,1,1)12	۸۰۷	۸۱۷	۱۲	۳۱۲	۲۹
SARIMA (1,2,1)*(1,1,1)12	۸۱۱	۸۲۶	۱۲	۳۰۸	۲۹

بحث

در سری‌های زمانی معمولاً از خود متغیر برای پیش‌بینی مقادیر آینده آن استفاده می‌کنند، اما می‌توان از چندین متغیر کمکی نیز بهره گرفت. برای مثال می‌توان برای پیش‌بینی تعداد موارد مرگ ناشی از تصادفات در استان کرمانشاه در هرماه علاوه بر داده‌های مرگ‌ومیر ناشی از تصادفات استان می‌توان از آمار و اطلاعات میزان مصرف بنزین در هرماه، تعداد روزهای بارانی در هرماه و اطلاعات دیگر استفاده نمود.که به این حالت سری‌های زمانی چندگانه[2] گفته می‌شود، این روش نسبت به روش سری‌های زمانی تک متغیره از کارایی و دقت بیشتری برخوردار است ولی متأسفانه به اطلاعات بیشتری نیاز دارد.

اگر واریانس داده‌ها تابعی از مقادیر گذشته باشد؛ نمی‌توان به‌راحتی واریانس داده‌ها را در طول زمان ثابت نگه داشت، بنابراین نمی‌توان از مدل‌های خانواده باکس وجنگینز یا مدل‌های آریما استفاده کرد، بنابراین نیاز است از مدل‌هایی که به فرض ثابت بودن واریانس نیاز ندارند، استفاده کرد، این مدل‌ها به خانواده ARCH & GRACH معروف هستند. پژوهشگران به چندین دلیل ترجیح می دهند از مدل‌های خانواده باکس وجنگینز استفاده کنند. اولاً برپا کردن این مدل‌ها راحت است، ثانیاً تفسیر این مدل‌ها راحت می‌باشد و با نرم‌افزارهای بیش‌تری می‌توان این کار را انجام داد.

تجزیه سری‌های زمانی به مؤلفه‌های تشکیل‌دهنده[1] آن مانند روند، تغییرات فصلی و سایر نوسانات نامنظم هرچند بهترین روش برای شناسایی الگو نمی‌باشد، اما برای شناسایی الگو مفید است. داده‌هایی که دارای روند فصلی هستند تفسیر ACF و PACF آن‌ها مشکل‌تر از داده‌های غیر فصلی می‌باشد، که دلیل آن ترکیب الگوی فصلی با الگوی غیر فصلی است، درنتیجه ACF و PACF ویژگی هر دو روند را دارد. که در وقفه‌های فصلی S,2S,3S به‌سرعت کاهش پیدا نمی‌کند و شاید هم معنی‌دار باشد، که در این حالت باید از تفاضل گیری فصلی استفاده نمود. لازم به ذکر

[1] Decomposition
[2] Multivariate time series

است در یک سری زمانی هر چه تعداد پارامترهای داخل مدل افزایش یابد، احتمالاً R2 (ضریب تعیین) بیش‌تر خواهد شد، اما تعداد پارامتر زیاد سبب می‌گردد استفاده از مدل، بسیار سخت و کاربرد آن کم‌تر شود؛ بنابراین هدف اصلی در پیش‌بینی یافتن می‌باشد که دارای پارامتر کم‌تری باشد. معیارهای AIC و SBC برای برازش کلی مدل با در نظر گرفتن اصل صرفه‌جویی مورداستفاده قرار می‌گیرد. اگر تعداد بازه‌های زمانی مورداستفاده کم باشد؛ بهتر است از معیار AIC و در بازه‌های زمانی با تعداد زیاد بهتر است معیار SBC استفاده شود، اما زمانی که برآوردها به‌سرعت همگرا نمی‌شود؛ باید بااحتیاط این دو شاخص مورد تفسیر قرار گیرد. درمجموع مدل مناسب برای برازش داده‌ها مدلی است که در آن مقدار ضرایب نشان‌دهنده ایستایی و معکوس‌پذیری داده‌ها و در طول دوره ثابت باشد. هم‌چنین در انتخاب پارامترهای آن اصل صرفه‌جویی رعایت شده باشد و تفاوت مقدار مشاهده‌شده و مقدار پیش‌بینی‌شده به‌وسیله مدل تنها به دلیل اغتشاش خالص باشد و نه به دلیل عدم شناسایی مدل مناسب باشد (۱۹،۲۰). برای برپا کردن یک سری زمانی از نرم‌افزارهای مختلف می‌توان استفاده نمود، بهترین نرم‌افزارها برای سری‌های زمانی نرم‌افزار STATA وR می‌باشد، نرم‌افزار SPSS و Minitab علاوه بر این‌که نمی‌توانند سری زمانی چند متغیره پیش‌بینی کنند؛ قدرت آن‌ها در شناسایی مدل کم است. هرچند برخی از نرم‌افزارهای آماری امکان انتخاب مدل مناسب را به‌طور خودکار فراهم کرده‌اند، که می‌تواند در شناسایی مدل مناسب بسیار کمک‌کننده باشد، اما این مدل در مواقعی که داده‌ها دارای بازه صفر، نیاز به تبدیل و یا داده پرت (به دلایل مختلف ازجمله اپیدمی یا اشتباه در ورود داده) درداده‌ها وجود داشته باشد؛ می‌تواند گمراه‌کننده باشد، علاوه براین توصیه می‌شود که استفاده‌کنندگان ازسری‌های زمانی مراحل انتخاب مدل را طی کنند؛ چون همان‌طور که قبلاً گفته شد امکان دارد بیش از یک مدل برای داده‌ها مناسب باشد.

نتیجه‌گیری

استفاده درست ازسری‌های زمانی بیشتر یک هنر است تا یک علم، بهترین راه برای فراگیری این هنر، تجربه و تکرار است. با توجه به این‌که پژوهشگران بیشتر اوقات در سری‌های زمانی مجبور می‌شوند برای شناسایی الگوی مناسب از روش‌های شهودی استفاده کنند؛ نباید تنها به یک روش برای شناسایی و انتخاب مدل اکتفا کرد، انتخاب مدل بر اساس توالی مراحل اجرای یک مدل سری زمانی و تکرار چندباره این توالی سبب شناسایی مدل یا مدل‌های مناسب خواهد شد.

منابع

1. Koopman SJ. Time series analysis by state space methods: OUP Oxford; 2012.. 93-113.
2. Hasanzadeh J, Amiresmaili M, Moosazadeh M, Najafi F, Moradinazar M. Implementing a Weather-Based Early Warning System to Prevent Traffic Accidents Fatalities. World Applied Sciences Journal. 2013; 24: 113-7.
3. Brockwell PJ, Davis RA. Introduction to time series and forecasting: Springer; 2002. 23-42.
4. Box GE, Jenkins GM, Reinsel GC. Time series analysis: forecasting and control: John Wiley & Sons; 2011.
5. Box GEP, Jenkins GM, Reinsel GC. Time series analysis: forecasting and control: Wiley; 2011. 31-53. 143-168
6. Chan NH. Time Series: Applications to Finance with R and S-Plus: John Wiley & Sons; 2011. 70-98.
7. Brockwell PJ, Davis RA. Time series: theory and methods: Springer; 2009. 44-79.
8. Brockwell PJ. Time Series Analysis: Wiley Online Library; 2005. 45-67.
9. Anderson TW. The statistical analysis of time series: John Wiley & Sons; 2011. 101-120.
10. Becketti S. Introduction to Time Series Using Stata. Stata Press books. 2013. 46-69.
11. Lütkepohl H. New introduction to multiple time series analysis. 2005. 112-145.
12. Salas J, Obeysekera J. ARMA model identification of hydrologic time series. Water Resources Research. 1982; 18: 1011-21.
13. Salas JD. Applied modeling of hydrologic time series: Water Resources Publication; 1980. 65-120.
14. Hollander M, Wolfe DA. Nonparametric statistical methods. 1999. 87-134.
15. Golyandina N, Nekrutkin V, Zhigljavsky AA. Analysis of time series structure: SSA and related techniques: CRC Press; 2010. 154-190.
16. Atkinson PM, Jeganathan C, Dash J, Atzberger C. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. Remote Sensing of Environment. 2012; 123: 400-17.
17. Meyer R, Krueger D. MINITAB Guide to Statistics: Prentice Hall PTR; 2001. 33-94.
18. Prado R, West M. Time Series: Modeling, Computation, and Interface: CRC Press; 2010. 110-143.
19. Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika. 1989; 76: 297-307.
20. Schmidt MB, Berri DJ. The impact of the 1981 and 1994-1995 strikes on Major League Baseball attendance: a time-series analysis. Applied Economics. 2002; 34: 471-8.

Iranian Journal of Epidemiology 2015; 11(1): 94-102.

Review Article

How to Choose an Appropriate Model for Time Series Data

Hasanzadeh J[1], F Najafi F[2], Moradinazar M[3]

1-Associate Professor, Department of Epidemiology, Shiraz university of Medical Sciences, Iran
2-Associate Professor, Department of Epidemiology, Research Center for Environmental Determinants of Health (RCEDH), Kermanshah University of Medical Sciences, Kermanshah, Iran
3- PhD candidate of epidemiology, Research Center for Environmental Determinants of Health (RCEDH), Kermanshah University of Medical Sciences, Kermanshah, Iran
Corresponding author: Moradinazar M., m.moradinazar@gmail.com

The time series is a collection of observation data that are arranged according to time. The main purpose of setting up a time series is to predict future values. The first step in time series data is graphed. Using graphs can provide general information such as uptrend or downtrend, seasonal patterns, periodic presence, and outliers in time series graphs. After graphing the data, if a good forecast is required, stationary data can be used. Differencing or decomposition methods can be used to make the data stationary. Then, a correlogram can be used to identify the order moving average and autoregressive model. The parameters of the model are examined using T-test. If the parameters are significant and the residue is independence, the predicted values can be evaluated using the mean absolute percentage error.

Keywords: Time series, Identify the model, Stationary, Prediction