Análise de séries temporais em epidemiologia: uma introdução sobre os aspectos metodológicos

Time series analysis in epidemiology: an introduction to methodological aspects

Maria do Rosário Dias de Oliveira Latorre
Departamento de Epidemiologia
Faculdade de Saúde Pública
Universidade de São Paulo
Endereço para correspondência/Correspondence to:
Av.Dr. Arnaldo, 715
01246-904 São Paulo - SP
e-mail: mdrddola@usp.br

Maria Regina Alves Cardoso
Departamento de Epidemiologia
Faculdade de Saúde Pública
Universidade de São Paulo

Apoio (bolsa pesquisador)
Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq: 300328/97-9)

Resumo

Este é um artigo introdutório sobre análise de séries temporais, onde se pretende apresentar, de maneira sumária, alguns modelos estatísticos mais utilizados em análise de séries temporais .

Uma série temporal, também denominada série histórica, é uma seqüência de dados obtidos em intervalos regulares de tempo durante um período específico. Na análise de uma série temporal, primeiramente deseja-se modelar o fenômeno estudado para, a partir daí, descrever o comportamento da série, fazer estimativas e, por último, avaliar quais os fatores que influenciaram o comportamento da série, buscando definir relações de causa e efeito entre duas ou mais séries. Para tanto, há um conjunto de técnicas estatísticas disponíveis que dependem do modelo definido (ou estimado para a série), bem como do tipo de série analisada e do objetivo do trabalho.

Para analise de tendências, podem se ajustar modelos de regressão polinomial baseados na série inteira ou em vizinhança de um determinado ponto. Isso também pode ser realizado com funções matemáticas. Define-se como um fenômeno sazonal aquele que ocorre regularmente em períodos fixos de tempo e, se existir sazonalidade dita determinística na série, podem-se utilizar modelos de regressão que incorporem funções do tipo seno ou cosseno à variável tempo.

Os modelos auto-regressivos formam outra classe de modelos. Na análise do comportamento de uma série histórica livre de tendência e de sazonalidade podem ser utilizados modelos auto-regressivos (AR) ou que incorporem médias móveis (ARMA). Quando há tendência, utilizam-se os modelos auto-regressivos integrados de médias móveis (ARIMA) e, para incorporar o componente de sazonalidade, utilizam-se os modelos SARIMA.

Por último há os modelos lineares generalizados. Neste grupo de modelos estatísticos, a variável resposta é um processo de contagem e as variáveis independentes são variáveis candidatas a explicar o comportamento

da série ao longo do tempo. Estes modelos são indicados quando as variáveis em estudo não têm aderência à distribuição normal, principalmente pelo fato de serem processos de contagem . Estes modelos compõem um grupo de distribuições de probabilidades conhecido como família exponencial de distribuições que englobam diversas funções aditivas, como a regressão linear, de Poisson, logística, log-linear etc. Os modelos aditivos generalizados são uma extensão desta classe de modelos, nos quais cada variável independente analisada não entra no modelo com o seu valor, mas sim, adotando uma função não paramétrica de forma não especificada, estimada a partir de curvas de alisamento.

Palavras chave: Séries temporais. Séries históricas. Modelos estatísticos. Tendência. Sazonalidade

Abstract

A time series, also denominated historical series, is a sequence of data obtained in regular intervals of time during a specific period. In the analysis of a time series, one first wants to model the study phenomenon and, from this, to describe the behaviour of the series, to make estimates, and, in the end, to evaluate the factors that may have influenced the behaviour of the series, with the objective of defining cause-effect relationships between two or more series. For this, there is a set of available statistical techniques which depend upon the defined model (or that estimated for the series), the type of the study series, and of the objective of the work.

To analyse trends, it is possible to adjust polynomial regression models based on the whole series or on the neighbourhood of a specific point. This can also be done with mathematical functions. A seasonal phenomenon is defined as the one that occurs regularly in fixed periods of time and, if there is seasonality considered as deterministic in the series, one can use regression models which include functions like seno or cosseno to the variable time.

In the analysis of the behaviour of a time series without trend and seasonality, the auto-regressive models (AR) or models which incorporate moving averages (ARMA) can be used. When trend is present, one can use auto-regressive models integrated with moving averages (ARIMA) and to incorporate the seasonality component the SARIMA models are used.

The generalized linear models constitute another class of models. In this group of statistical models, the response variable is a counting process and the independent variables are those which are candidates to explain the behaviour of the series throughout the time. This class of models is indicated when the study variables do not follow the Normal distribution, mainly because they are counting processes. These models represent a group of probability distributions known as exponential family of

distributions that incorporates many additive functions like the linear regression, Poisson, logistic, log-linear, etc. The generalized additive models are an extension of this class of models, in which each independent variable analysed does not enter in the model with its own value, but adopting a non parametric function in a non specific manner, which is estimated from smoothing curves.

Keywords: Time series. Models, statistical. Trends. Seasonality.

Este número da Revista Brasileira de Epidemiologia tem como ênfase artigos científicos que analisaram séries históricas. Este tipo de análise é muito comum em Epidemiologia, quando se pretende analisar o comportamento de algum fenômeno ao longo do tempo. Neste número há diversos exemplos de séries históricas e espera-se, com isso, mostrar as inúmeras aplicações da análise de séries temporais, estimulando os leitores ao aprofundamento das mesmas. Este é um artigo introdutório sobre análise de séries temporais, onde se pretende apresentar, de maneira sumária, alguns modelos estatísticos mais utilizados Não há a pretensão de ser um livro texto, e detalhes de modelagem podem ser vistos nos livros textos indicados nas referências bibliográficas.

Definição

Uma série temporal, também denominada série histórica, é uma seqüência de dados obtidos em intervalos regulares de tempo durante um período específico[1,2]. Este conjunto pode ser obtido através de observações periódicas do evento de interesse como, por exemplo, o valor máximo diário da concentração de ozônio no ar no Município de São Paulo, ou através de processos de contagem como o total mensal de óbitos por câncer no Rio Grande do Sul. Se a série histórica for denominada como Z, o valor da série no momento t pode ser escrito como Z_t (t=1,2,...,n). Denomina-se trajetória de um processo, a curva obtida no gráfico da série histórica e o conjunto de todas as possíveis trajetórias é denominado como um processo estocástico. Considera-se que uma série temporal é uma amostra deste processo.

O conjunto de observações ordenadas no tempo pode ser *discreto* como o número de atendimentos diários em um Pronto Socorro ou o número mensal de casos notificados de uma doença específica; ou *contínuo,* como o registro de um eletrocardiograma de uma pessoa ou o registro dos valores de temperatura e umidade ao longo do dia. Pode-se obter uma série temporal discreta a partir de uma amostra de pontos

de uma série contínua ou por meio de um parâmetro como, por exemplo, a média de períodos fixos de tempo.

Na análise de uma série temporal, primeiramente deseja-se modelar o fenômeno estudado para, a partir daí, descrever o comportamento da série, fazer estimativas e, por último, avaliar quais os fatores que influenciaram o comportamento da série, buscando definir relações de causa e efeito entre duas ou mais séries. Para tanto, há um conjunto de técnicas estatísticas disponíveis que dependem do modelo definido (ou estimado para a série), bem como do tipo de série analisada e do objetivo do trabalho.

Componentes

Uma série histórica pode ser composta por três componentes não observáveis[2,3]: tendência (T_t), sazonalidade (S_t) e a variação aleatória denominada de ruído branco (a_t). A primeira escolha para a elaboração de um modelo seria um relacionamento aditivo destes componentes: $Z_t=T_t+S_t+a_t$. Pode-se construir, também, um modelo multiplicativo ($Z_t=T_t . S_t . a_t$) ou realizar-se a transformação log, no modelo multiplicativo, quando ele se transforma no modelo log-linear. Ao analisar uma série histórica, deve-se estudar cada um destes componentes separadamente, retirando-se o efeito dos outros.

Para analisar a tendência os dois métodos mais utilizados são: 1) ajuste de uma função polinomial do tempo e 2) análise do comportamento da série ao redor de um ponto, estimando a tendência naquele ponto[2,3]. Na primeira opção, utilizam-se os modelos de regressão polinomial e, na última, modelos auto-regressivos.

Modelos de regressão polinomial

Nos modelos de regressão polinomial, os valores da série são considerados como variável dependente (Y) e os períodos do estudo como variável independente (X). Primeiramente deve-se fazer o diagrama de dispersão de Z_t (Y) em relação ao tempo para visualizar qual a função que mais se ajusta à trajetória do processo: linear, parábola, exponencial etc. Para se evitar a correlação serial entre os termos da equação de regressão, recomenda-se[4,5] fazer a transformação da variável período na variável período-centralizada (período menos o ponto médio da série histórica), estimando-se, então, o modelo de regressão correspondente. Detalhes do processo de centralização da variável tempo podem ser vistos em Latorre[5]. Os exemplos da aplicação desta técnica, apresentados neste número da Revista Brasileira de Epidemiologia, são os trabalhos de Tomé e Latorre[6], Hallal e colaboradores[7] e Bastos e colaboradores[8].

Tomé e Latorre[6] analisaram as tendências da mortalidade infantil e seus componentes para o Município de Guarulhos, no período de 1971 a 1998, utilizando modelos de regressão polinomial. Verificaram que apenas no período de 1971 a 1980 todos os coeficientes apresentaram tendências decrescentes estatisticamente significativas. Na série histórica de 1981 a 1990, somente os coeficientes de mortalidade infantil (p=0,0058), o de mortalidade neonatal tardia (p=0,0105) e da pós-neonatal (p=0,0045) apresentaram tendências estatisticamente decrescentes, o mesmo acontecendo no período de 1991 a 1998 (respectivamente p<0,0001, p=0,0173 e p=0,0044).

Hallal e colaboradores[7] analisaram a mortalidade por câncer no Rio Grande do Sul entre 1979 e 1995, utilizando modelos de regressão linear simples. A tendência temporal das taxas padronizadas de mortalidade para o total dos cânceres, segundo sexo, foi de estabilidade, do ponto de vista estatístico, o mesmo acontecendo com as seguintes localizações: cólon/reto feminino e colo do útero e útero não especificado. Obtiveram tendências estatisticamente crescentes nos coeficientes de mortalidade por câncer de pulmão (ambos os sexos), mama feminina, próstata e cólon/reto masculino. Já a mortalidade por câncer de estômago (ambos os sexos) e esôfago masculino teve tendência de decréscimo estatisticamente significativo.

Bastos e colaboradores[8] analisaram a

tendência da epidemia de AIDS em adultos, no período de 1985 a 1997, no Município de São Paulo, tendo como enfoque principal os usuários de drogas injetáveis (UDI). Utilizaram modelos de regressão polinomial e observaram que no período de 1985 a 1992 houve tendência de ascensão dos casos de AIDS em UDI e não UDI; a partir deste ponto ocorreu um declínio para UDI e manutenção em platô elevado para os não UDI, pela tendência de crescimento constante entre mulheres e homens heterossexuais. Os modelos que mais se ajustaram foram os de segunda ordem (parábola), exceto para os heterossexuais não UDI, cuja tendência foi de aumento linear.

Às vezes, devido à grande oscilação dos pontos, é necessário suavizar a série reduzindo o ruído branco. Há várias técnicas de alisamento, sendo que a mais utilizada é a média móvel. Detalhes do processo de alisamento da série podem ser vistos em Moretin e Toloi[2] ou em Latorre[5].

As vantagens[2,3] da estimação da tendência utilizando modelos de regressão polinomial são o grande poder estatístico desta classe de modelos, fácil elaboração e interpretação. Entretanto, algumas vezes não há uma função definida, como a linear ou exponencial, tornando-se necessário que o pesquisador ajuste uma função matemática como a Kernel e outras ou utilize outra classe de modelos.

A segunda opção para a análise de séries históricas seria a estimação da tendência analisando o comportamento da série ao redor de um ponto, estimando a tendência para valores da série próximos à ele; e não utilizando a série como um todo. A análise utilizando parte da série é mais recomendada quando se deseja avaliar apenas uma parte da trajetória ou quando o comportamento da série é muito instável. Nesta situação, melhores projeções devem ser feitas apenas a partir de um passado recente da mesma.

Após a estimativa da tendência, para se analisar os outros componentes é necessário construir uma série "livre de tendência" através das diferenças da mesma (D^dZ_t), onde *d* é o grau do polinômio obtido na análise da tendência. Por exemplo, se a tendência obtida for linear (1º. grau) , bastaria fazer uma diferença da série Z_t (Z_t-Z_{t-1}) para que ela estivesse livre de tendência. Detalhes do processo de diferenças da série podem ser visto em Moretin e Toloi[2].

Sazonalidade

A sazonalidade é um componente da série histórica difícil de ser estimado, pois é necessário compatibilizar a questão conceitual do fenômeno em estudo com a questão estatística. Define-se um fenômeno sazonal como aquele que ocorre regularmente em períodos fixos de tempo[1-3]. Se houver uma sazonalidade dita determinística podem ser utilizados modelos de regressão que incorporem funções do tipo seno ou cosseno à variável tempo. Para se descobrir se existe sazonalidade na série de valores e verificar qual o seu ritmo é importante realizar uma análise espectral[3]. Com este tipo de análise é possível se identificar um padrão sazonal, mesmo dentro de uma variabilidade aleatória. A análise espectral utiliza um conjunto de funções que contêm seno e cosseno e tenta ajustá-las à variância observada em uma série de observações no tempo, levando em conta a amplitude das "ondas", o período em que elas se repetem e a fase em que se iniciam[3].

Para se retirar o efeito da sazonalidade de uma série[2], pode-se usar a média móvel centrada no número de períodos que compõem uma repetição (por exemplo, para sazonalidade anual, seria utilizada a média móvel de 12 meses) ou, então, pode-se trabalhar com a diferença entre a série original (Z_t) e o polinômio estimado para a sazonalidade.

Modelos auto-regressivos

Antes de se conduzir qualquer análise é importante definir se a série é estacionária ou não, para, a partir daí, estabelecer a estrutura do modelo probabilístico que estimará a série. Uma série é considerada estacionária[2,3] quando suas observações ocor-

rem aleatoriamente ao redor de uma média constante, ou seja, não há tendência. Isto significa que $E(Z_t)=E(Z_{t+m})=m$ e $Var(Z_t)=Var(Z_{t+m})=$constante. Para tanto, define-se a função de auto-correlação[2,3] (também chamada de função de correlação serial) que, em cada período j (*lag* j) da série, é calculado o coeficiente de correlação entre as observações t e t+j. Neste caso, se t e t+j são independentes, a correlação entre t e t+j é zero.

O modelo mais simples é obtido para a série histórica estacionária, ou seja, livre de tendência e de sazonalidade. Esta série é conseqüência da variação aleatória do ruído branco ao redor de uma grande média, ao longo do tempo. Ela é escrita[2,3] como a combinação aleatória dos valores anteriores da Z_t ($Z_t=b_1 Z_t\text{-}_1+b_2Z_{t-2}+ ... +b_p Z_{t-p}$) e, por isso, a série toda pode ser função do ruído branco. Essa classe de modelos é conhecida como modelos auto-regressivos-AR (no caso, de ordem $_p$). Este é um processo interativo onde há a identificação da ordem *p* através da função de auto-correlação; a partir daí, faz-se a estimativa de um modelo de previsão bem como a análise dos resíduos para a avaliação da existência de viéses e/ou grandes erros de estimativas. A dificuldade desta técnica é a identificação do modelo, pois é possível que pessoas diferentes identifiquem modelos de ordem diferentes para a mesma série temporal.

Para muitas séries, a melhor solução se encontra em combinar o modelo auto-regressivo (AR) com o de médias móveis (MA)[2,3]. Este é composto pela combinação linear de valores próximos da série (AR de ordem p) com uma combinação linear dos ruídos brancos próximos ao valor da série (MA de ordem q).

Tanto o modelo AR, quanto o MA, quanto o ARMA são utilizados para séries estacionárias. Entretanto, quando o processo é não estacionário homogêneo (ou seja, possui tendência, porém não é explosivo), uma das maneiras de analisá-lo é incorporando um processo de diferenças (D^dZ_t) no modelo ARMA[2,3]. Este é o modelo conhecido como ARIMA (modelo auto-regressivo integrado de médias móveis), onde *d* é a ordem das diferenças necessárias para tirar a tendência da série. Há duas situações em que a série pode ser considerada não estacionária: 1) quando durante um período os pontos oscilam ao redor de uma média e, depois, mudam de patamar (neste caso basta tomar uma diferença da série); e 2) quando a série é não estacionária em relação à tendência (geralmente, para torná-las estacionárias é necessário tomar a segunda diferença). Os modelos ARIMA podem dar conta da sazonalidade quando há *lags* de baixa ordem. Porém, quando a sazonalidade ocorre em múltiplos períodos, é necessário que se considere no modelo um componente de sazonalidade estocástica. Nesta situação, utiliza-se o modelo SARIMA[2,3] que incorpora as funções trigonométricas (preferencialmente, seno e cosseno) ao modelo ARIMA, e a ordem da sazonalidade vai depender da série.

Exemplo da aplicação do modelo SARIMA pode ser visto na trabalho de Otero e colaboradores[9]. Estes autores descreveram a evolução da mortalidade por desnutrição em idosos nas Regiões Metropolitanas dos Estados do Rio de Janeiro (RMRJ) e de São Paulo (RMSP), verificando suas tendências entre 1980 e 1996 e propuseram um modelo para predição dos mesmos. Os resultados apontaram a existência tendências de aumento e revelaram um padrão sazonal no inverno na RMSP (maior número de óbitos em junho e julho) e no verão na RMRJ (aumento em janeiro), concluindo que o modelo SARIMA foi o melhor para fazer previsões.

Modelos lineares generalizados

Neste grupo de modelos estatísticos a variável dependente ou resposta (Y) é um processo de contagem (por exemplo, número de óbitos ou de atendimentos diários) e as variáveis independentes são variáveis candidatas a explicar o comportamento da série ao longo do tempo. Esta classe de modelos é indicada quando as variáveis em estudo não têm aderência à distribuição nor-

mal, principalmente pelo fato de serem processos de contagem (ou seja, são variáveis quantitativas discretas).

Estes modelos compõem um grupo de distribuições de probabilidades conhecido como família exponencial de distribuições que englobam diversas funções aditivas, como a regressão linear, de Poisson, logística, log-linear etc. Os modelos aditivos generalizados são uma extensão desta classe de modelos, nos quais cada variável independente analisada não entra no modelo com o seu valor, mas sim, adotando uma função não paramétrica de forma não especificada, que é estimada a partir de curvas de alisamento. Sendo assim, não é necessário assumir uma relação linear e/ou aditiva entre a variável dependente e a variável independente em estudo. A trajetória alisada proporciona a visualização não somente da forma, mas, também, apresenta as possíveis não linearidades nas relações estudadas, uma vez que não apresenta uma função paramétrica rígida. Maiores detalhes desta classe de modelos podem ser vistos no trabalho de Conceição e colaboradores[10]. Neste artigo há a descrição e comparação dessas duas classes de modelos e os autores mostram, como exemplo de aplicação, a análise da associação entre mortalidade por idosos e poluição atmosférica na cidade de São Paulo, no período de 1994 a 1997.

Um outro exemplo desta técnica é o estudo de Martins e colaboradores[11]. Neste estudo o objetivo foi investigar a associação entre os níveis diários dos poluentes do ar e os atendimentos de idosos com infecções de vias aéreas superiores (IVAS), do Pronto Socorro Médico do Hospital das Clínicas de São Paulo, no período de 1996 a 1998. Foram estimados modelos aditivos generalizados de regressão de Poisson, ajustados por sazonalidade (funções não paramétricas de alisamento), fatores climáticos (funções lineares), variáveis indicadoras de dias da semana, períodos de rodízio e número diário de atendimentos por doenças não respiratórias. Os resultados mostraram que tanto o monóxido de carbono quanto o dióxido de enxofre estiveram diretamente associados a IVAS, sendo esta associação robusta, resistindo à inclusão das variáveis de controle.

Considerações finais

Este número da Revista Brasileira de Epidemiologia tem como enfoque principal a utilização de análise de séries temporais em Epidemiologia. A análise de séries históricas é comum em estudos descritivos, porém muitas vezes observa-se a falta da metodologia estatística adequada. Comparam-se, visualmente, dois ou mais pontos, porém não é levado em conta que existe uma oscilação devido ao acaso. Daí decorre a necessidade de avaliação de séries históricas através de modelos estatísticos diversos, cada um mais adequado a uma específica trajetória no tempo. Neste texto introdutório foram apresentados de maneira sumária apenas os modelos estatísticos mais utilizados em análise de séries temporais. Detalhes de modelagem podem ser vistos nos livros textos indicados nas referências bibliográficas, e diversos exemplos de suas aplicações podem ser vistos ao longo deste número especial da Revista. Espera-se, com isso, mostrar as inúmeras aplicações da análise de séries temporais, estimulando-se a busca das técnicas estatísticas adequadas. A disponibilidade de diversos pacotes estatísticos permite aos pesquisadores a utilização da melhor técnica estatística, porém o fascínio pela técnica não deve obscurecer o desenvolvimento das hipóteses, pois as análises estatísticas ganham pleno sentido quando orientadas por hipóteses epidemiologicamente relevantes.

Referências

1. EveritT BS. ***The Cambridge dictionary of statistics in the medical sciences***. Cambridge: Cambridge University Press; 1995.

2. Morettin PA, Toloi CMC. ***Previsão de séries temporais.*** 2ª. ed**.** São Paulo:Atual Editora; 1985.

3. Diggle PJ. ***Time series: a biostatistical introduction.*** Oxford:Oxford University Press; 1992.

4. Draper NR, Smith S. ***Applied regression analysis.*** New York: John Wiley and Sons; 1981. p:141-69, 250-65. (Wiley Series in Probability and Mathematical Statistics).

5. Latorre MRDO. ***Câncer em Goiânia: análise da incidência e da mortalidade no período de 1988 a 1997.*** [Tese de Livre Docência]: São Paulo: Faculdade de Saúde Pública da USP; 2001.

6. Tomé EA, Latorre MRD de O. Tendências da mortalidade infantil no Município de Guarulhos: análise do período de 1971 a 1998. ***Rev Bras Epidemiol.***, 2001; 4:153-67.

7. Hallal ALC, Gotlieb SL, Latorre MRD de O. Evolução da mortalidade por neoplasias malignas no Rio Grande do Sul, 1979-1995. ***Rev Bras Epidemiol.***, 2001; 4:168-77.

8. Bastos MSCBO, Latorre MRD de O, Waldman EA. Tendência da epidemia de AIDS em usuários de drogas injetáveis no Município de São Paulo de 1985 a 1997. ***Rev Bras Epidemiol.***, 2001; 4:178-90.

9. Otero UB, Rozenfeld S, Gadelha AJ. Óbitos por desnutrição em idosos em São Paulo e Rio de Janeiro: análise de séries temporais- 1980-1996. ***Rev Bras Epidemiol.***, 2001; 4:191-205.

10. Conceição GMS, Saldiva PHN, Singer JM. Modelos GLM e GAM: uma tradução para leigos e aplicação a um estudo de mortalidade e poluição atmosférica na cidade de São Paulo. ***Rev Bras Epidemiol.***, 2001; 4:206-19.

11. Martins LC, Latorre MRD de O, Braga ALF, Saldiva PHN. Relação entre poluição atmosférica e atendimentos por infecção de vias aéreas superiores no Município de São Paulo: avaliação do rodízio de veículos. ***Rev Bras Epidemiol.***, 2001; 4:220-9.