Estatística Computacional e Análise de Dados 120 (2018) 70-8

Listas de conteúdo disponíveis ScienceDirect

Estatística Computacional e Análise de Dados

página inicial do diário www.elsevier.com/locate/csda

Uma nota sobre a validade da validação cruzada para avaliação da previsão da série temporal autoregressiva

Christoph Bergmeir [a,*], Rob J. Hyndman [b], Bonsoo Koo [b]

[a] Faculdade de Tecnologia da Informação, Monash University, Melbourne, Austrália

[b] Departamento de Econometria & Estatísticas de Negócios, Universidade Monash, Melbourne, Austrália

artigo info

Histórico do artigo:
Recebido em 10 de junho de 2016
Recebido em formulário revisado em 30 de outubro de 2017
Aceito 3 novembro 2017 Disponível online
22 novembro 2017

Keywords:
Validação cruzada
Série temporal
Autorressões

abstrair

Um dos procedimentos padrão mais utilizados para avaliação de modelos em classificação e regression é a validação cruzada de dobra k (CV). No entanto, quando se trata de previsão de séries temporândias, devido à correlação serial inerente e potencial não-estacionária dos dados, sua aplicação não é simples e muitas vezes substituída por profissionais em favour de uma avaliação fora da amostra (OOS). Mostra-se que para modelos puramente autoregressivos, o uso do *padrão* K-fold CV é possível desde que os modelos considerados tenham erros não corrigidos. Tal configuração ocorre, por exemplo, quando os modelos aninham um modelo mais apetitoso. Isso é muito comum quando os métodos de Machine Learning são usados para predição, e onde o CV pode controlar para o excesso de adaptação dos dados. São apresentados insights teóricos que apoiam esses argumentos, juntamente com um estudo de simulação e um exame real. É mostrado empiricamente que o *CV* de dobra K tem um desempenho favorável em comparação com a avaliação do OOS e outras técnicas específicas da série de tempo, como a validação cruzada não dependente.

1. Introdução

Validação cruzada (CV) (Stone, 1974; Arlot e Celisse, 2010) é um dos métodos mais utilizados para avaliar a generalizabilidade dos algoritmos na classificação e regressão (Hastie et al., 2009), e está sujeito a pesquisas ativas em andamento (por exemplo, Budka e Gabrys, 2013; Borra e Di Ciaccio, 2010; Bergmeir et al., 2014; Moreno-Torres et al., 2012). No entanto, quando se trata de previsão de séries temporência, os praticantes muitas vezes não têm certeza da melhor maneira de avaliar seus modelos. Muitas vezes há um sentimento de que não poderíamos usar dados futuros para prever o passado. Além disso, a correlação serial nos dados, juntamente com possíveis não-estacionárias, fazem com que o uso do CV pareça problemático, pois não contabiliza essas questões (Bergmeir e Benítez, 2012). Os practitioners geralmente recorrem à avaliação fora da amostra (OOS), onde uma seção do final da série é retida para avaliação. Dessa forma, apenas uma avaliação em um conjunto de testes é considerada, enquanto que com o uso de validação cruzada, são realizadas várias avaliações desse tipo. Assim, usando o OOS, os benefícios do CV, especialmente para pequenos conjuntos de dados, não podem ser explorados. Uma parte importante do problema é que, na literatura tradicional de previsão, a avaliação da OOS é o procedimento de avaliação padrão, parcialmente porque a montagem de modelos padrão como a suavização exponencial (Hyndman et al., 2008) ou modelos ARIMA são totalmente iterativas no sentido de que eles começam a estimar no início da série. Além disso, algumas pesquisas têm demonstrado casos em que o cv standard falha em um contexto de série temporal. Por exemplo, Opsomer et al. (2001) mostrar que o CV padrão subestima a largura de banda em uma estrutura de regressão de estimadores do kernel se a correção automática do erro for alta, de modo que o método se estica os dados. Como um result, várias técnicas cv foram desenvolvidas especialmente para o

* Correspondência para: Faculdade de Tecnologia da Informação, Caixa Postal 63 Monash University, Victoria 3800, Austrália.
Endereço de e-mail: christoph.bergmeir@monash.edu (C. Bergmeir).

https://doi.org/10.1016/j.csda.2017.11.003 0167-9473/

caso dependente (Györfi et al., 1989; Burman e Nolan, 1992; Burman et al., 1994; McQuarrie e Tsai, 1998; Racine, 2000; Kunst, 2008).

Este estudo aborda o problema da seguinte forma. Quando métodos autoregressivos puramente (não lineares, não paramétricos) são aplicados a problemas de previsão, como muitas vezes é o caso (por exemplo, ao usar métodos de Machine Learning), os problemas acima mencionados do CV são em grande parte irrelevantes, e o CV pode e deve ser usado sem modificação, como no caso independente. Pelo que sabemos, este é o primeiro artigo a justificar o uso do CV padrão de dobra K no setting dependente sem modificação. Nosso artigo traça a linha de aplicabilidade do procedimento entre modelos que possuem erros não corrigidos, e modelos que são fortemente mal especificados e, portanto, não produzem erros não corrigidos. Na prática, isso significa que o método pode ser usado sem problemas para detectar o excesso de adaptação, já que modelos superequipados têm erros não corrigidos. A subequiquação deve ser tratada com antecedência, por exemplo, testando resíduos para correlação serial. Fornecemos uma prova teórica e resultados adicionais de experimentos de simulação e um exemplo real para justificar nosso argumento.

2. Validação cruzada para o caso dependente

A validação cruzada para o ambiente dependente tem sido estudada extensivamente na literatura, incluindo Györfi et al. (1989), Burman e Nolan (1992) e Burman et al. (1994). Deixe $\boldsymbol{y} = \{y1,..., yn\}$ ser uma série de tempo. Tradicionalmente, quando *o CV* de dobra K é realizado, *os* números escolhidos aleatoriamente K fora do vetor $\boldsymbol{y}$ são removidos. Essa remoção invalida o CV no ambiente dependente devido à correlação entre erros nos conjuntos de treinamento e teste. Portanto, Burman e Nolan (1992) sugerem correção de viés, enquanto Burman et al. (1994) propor *cv bloco h* pelo qual as observações h anteriores e seguintes à observação são deixadas de fora no conjunto de testes. Chamamos esse procedimento de validação cruzada não dependente, pois deixa de fora as observações possivelmente dependentes e considera apenas pontos de dados que podem ser considerados independentes.

No entanto, tanto a correção de viés quanto o método de CV de bloco h têm suas limitações, incluindo o uso ineficiente dos dados disponíveis.

Vamos agora considerar um modelo puramente autoregressivo de ordem *p*

$$yt = g(_{xt,\varphi}) + \varepsilon_{t}, \tag{1}$$

onde ε t é um termo de choque ou perturbação, φ é um vetor de parâmetro, $\boldsymbol{xt} \in$ Rp consiste em valores defasados de *yt* e $g\,(_{xt,\varphi}) = E\phi[yt \mid \boldsymbol{xt}]$. Aqui $g\,(\cdot)$ pode ser uma função linear ou não linear, ou mesmo uma função não paramétrica. Assim, $g(\cdot)$ poderia ser uma função totalmente não especificada dos valores defasados de *yt* até *pth* order.

Aqui, a ordem de lag do modelo é fixa e a série temporal é *incorporada em* conformidade, gerando uma matriz que é então usada como entrada para um algoritmo de regressão (não paramétrico, não linear). A série de tempo incorporada com pedido *p* e um horizonte de previsão fixa de *h* = 1 é definida da seguinte forma:

$$\begin{bmatrix} y_1 & e2 & \dots & ypyp & +1 \\ \vdots & & & & \vdots \\ y_{t-} & \dots & \dots & \dots & . \\ & & & -1 & \\ \vdots & pyt_{-p+1} & . & . & \\ y_{n-} & . & \cdot\ ytyt & |. \quad (2) & n \\ & \dots & \dots & .... & \end{bmatrix}$$

$$pyn_{-\,p+1} \quad \dots \quad yn_{-\,1y}$$

Assim, cada linha é da forma $[x^{t}, yt]$, e as primeiras colunas *p* da matriz contêm preditores para a última coluna da matriz.

Lembre-se do método CV de dobra K usual, onde os dados de treinamento são particionados em conjuntos separados K, digamos $J = \{J1,..., JK\}$. Defina $Jk_{-} = U_{j0=}kJj$, de modo que para uma avaliação específica *k* na validação cruzada, *Jk* é usado como conjunto de teste, e os conjuntos restantes combinados, $_{Jk-}$, são used para montagem do modelo. Em vez de reduzir o conjunto de treinamento removendo as observações *h* anteriores e seguindo a observação do conjunto de testes, deixamos de fora todo o conjunto de linhas correspondentes a $t \in Jk$ na matriz (2). Fig. 1 ilustra o procedure.

Desde que (1) seja verdadeira, as linhas da matriz (2) são condicionalmente não corrigidas porque $yt - g\dots(_{xt,\varphi}) = \varepsilon\dots_{t}$ é simplesmente um erro de regressão, e $\{\varepsilon\dots_{t}\}$ são assintáticamente independentes e distribuídas de forma idêntica (i.i.d.) desde *que g* é estimado adequadamente. Consequentemente, omitir linhas da matriz não afetará o viés ou consistência das estimativas.

Na prática, embora não saibamos o p correto e outros hiper parâmetros do modelo, se o nosso modelo for suficientemente grande e flexível (e, portanto, passível de ser superjustado), os erros não serão corrigidos. No caso de erros correlacionados, o procedimento cv será tendencioso, mas isso pode ser relativamente abordado testando os resíduos para correlação serial.

Vale ressaltar que este método deixa toda a linha relacionada ao conjunto de testes escolhido, em vez de apenas componentes de conjunto de teste. Como resultado, perdemos muito menos informações incorporadas nos dados dessa forma do que no CV do bloco H.

3. O resultado teórico

Que $y1,\ldots, yn$ sejam as observações de um processo estacionário onde cada yt tem distribuição P em R. Consideramos modelos AR(p) puros para discutir a validade do nosso método. Sem a totalidade da generalidade e para a facilidade de notação, focamos no CV de saída porque a generalização do *CV* K-fold é simples.

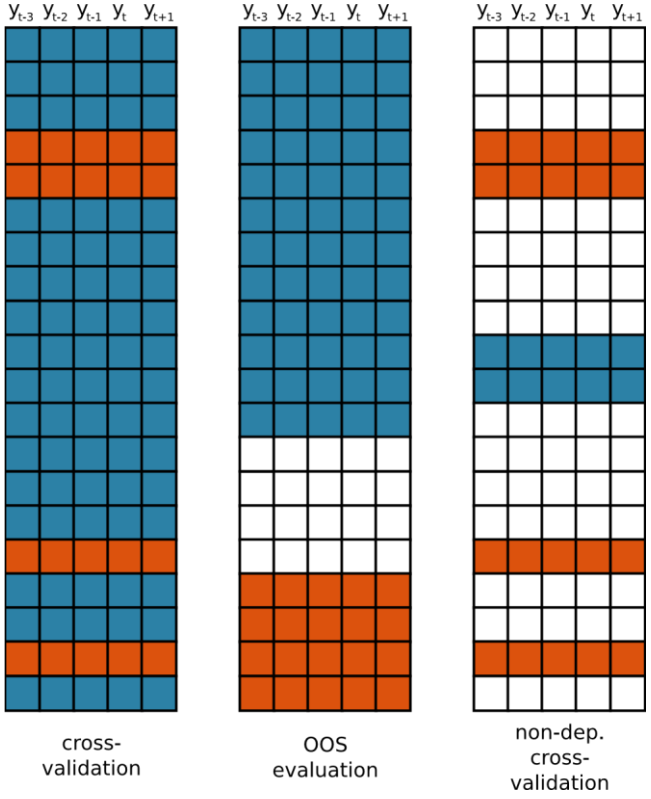

Fig. 1. Conjuntos de treinamento e teste para diferentes procedimentos de validação cruzada para uma série de tempo incorporada. As linhas escolhidas para o treinamento são mostradas em azul, as linhas escolhidas para testes em laranja, as linhas mostradas em branco são omitidas devido a considerações de dependência. O exemplo mostra uma dobra de um CV de 5 vezes e uma incorporação da ordem 4. Assim, para as considerações de dependência, 4 valores antes e depois de um caso de teste não podem ser usados para treinamento. Vemos que o CV não dependente reduz consideravelmente os dados de treinamento disponíveis. (Para interpr etation das referências à cor nesta legenda de figura, o leitor é referido à versão web deste artigo.)

Considere o seguinte modelo de regressão não linear,

$$yt = g(xt,\varphi) + \varepsilon_t, \tag{3}$$

onde $g(\cdot)$ é uma função contínua e diferente com respect to φ para todos $xt = (yt-1, yt-2,\ldots, yt-p)'$ e ε_t é um erro de regressão. A função objetiva relacionada à estimativa de $\varphi\cdots$ é dada como

$$Q(\varphi) = \sum_{t=p+1}^{n} (yt - g(xt,\varphi))^2.$$

Por definição, $\varphi^{\cdots}$ = arg min$\phi \in \Phi Q(\varphi)$. Suponha $\{y' t\}^{mt}{}_{=1}$ é outro processo que tem a mesma distribuição que os dados da amostra $\{yt\}^{nt}{}_{=1}$ e $\boldsymbol{x}'_t = (y' t_{-1}, y t_{-2}, \ldots, y t_{-p})$. Por exemplo, $\{y \dot{I} t\}^{tm}{}_{=1}$ podem ser os dados futuros. Em seguida, o erro de previsão que consideramos nos modelos não lineares é dado como

PE = $E\{y' - g(\boldsymbol{x}', \varphi^{\cdots})\}^2$, onde $\varphi^{\cdots}$ é a solução para o conjunto de equações não lineares

$$\left(\frac{\partial}{\partial} g(\tilde{\boldsymbol{x}}_t, \hat{\boldsymbol{\theta}})\right)\left(\tilde{y}_t - g(\tilde{\boldsymbol{x}}_t, \hat{\boldsymbol{\theta}})\right) = \sum^{n} \left(\frac{\partial}{\partial} g(\tilde{\boldsymbol{x}}_t, \hat{\boldsymbol{\theta}})\right)$$

n

$q(\varphi^{\cdots}) = \sum \varepsilon\, t(\varphi^{\cdots}) = 0, \qquad \varphi\varphi\varphi_{\,t=p+1t} \qquad =p+1$

onde $\varepsilon\, \phi(\varphi) = y'_t - g(\boldsymbol{x}'t, \varphi)$. Por construção, ε tem a mesma distribuição *que* ε_t. Enquanto isso, consideramos estimar PE por validação cruzada. Aqui a amostra de treinamento é $\{(\boldsymbol{xj}, yj);\ j = p + 1, \ldots, n, j º = t\}$ e a amostra de teste é $\{(\boldsymbol{xt}, yt)\}$. Deixamos de fora toda a linha de matriz (2) correspondente ao conjunto de testes. Um estimador de pe using validação cruzada é

$$\widehat{\mathrm{PE}} = \frac{1}{n-p} \sum_{t=p+1}^{n} \{y_t - g(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_{-t})\}^2 \quad \_\_\_\_ \ ,$$

onde $\varphi^{\cdots}{}_t$ é a estimativa de saída de φ, que é a solução para o seguinte conjunto de equações não lineares

$$\left(\frac{\partial}{\partial \boldsymbol{\theta}} g(\boldsymbol{x}_j, \hat{\boldsymbol{\theta}}_{-t})\right)\left(y_j - g(\boldsymbol{x}_j, \hat{\boldsymbol{\theta}}_{-t})\right) = \sum_{j=p+1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\theta}} g(\boldsymbol{x}_j, \hat{\boldsymbol{\theta}}_{-t})\right)$$

n

$q(\varphi \qquad \ldots t) = \sum \varepsilon_j(\varphi \qquad \ldots t) = 0,$

$j=p+1$ $j2=$ $tjes=t$

n

onde $q(\varphi... \quad {}_t)$ é obtido a partir da diferenciação $Q_{-t}(\varphi)(= \sum (yj - g(_{xj,\varphi}))^2)$ em relação a φ.

$j=p+1\ j0$

$=t$

Para fazer o trabalho de validação cruzada, pe deve aproximar-se da PE. Para isso, precisamos de um conjunto de suposições da seguinte forma.

Suposições

Processo AR(p) não linear estacionário) Uma sequência de $\{yt\}^{nt}{}_{=1}$ é representada pelo processo AR(p) não linear como em (3). O processo AR(p) não linear é assumido como estacionário e ergodídico.

A2 (**Consistência dos Quadrados Menos Não Lineares**) *é* um estimador consistente de φ.

A3 (**Erros são MDS**) Os distúrbios $\{\varepsilon_t\}$ satisfazem o seguinte.

(i) $\{\varepsilon t, Ft\}$ formar uma sequência de diferenças de martingale (MDS) onde $_{Ft}$ é o campo sigma gerado por $\{\varepsilon_{s,\ ys};\ s \le t\}$. Note que os erros de I.I.D são MDS.

(ii) var $(\varepsilon_t \mid \boldsymbol{xt}) = \sigma^2$ e $E|\ \varepsilon_{\ l}{}^{4+\delta} < K$ para alguns $K < \infty$ e $\delta > 0$.

(iii) $\{\varepsilon_t\}$ têm distribuição absolutamente contínua em relação à medida lebesgue. Essa suposição está satisfeita com erros normalmente utilizados, incluindo erros normalmente distribuídos.

Assunção **A1** afirma nosso modelo ao qual o CV é aplicado, e garante que nenhuma especificação de erro seja considerada em nossa configuração. Observe que (3) é um processo AR(p) padrão e nossa configuração abrange aplicações comuns de CV para dados dependentes. Além disso, sob certas condições de regularidade, a sequência $\{yt\}$ satisfaz a condição de mistura forte below (ver Mokkadem, 1988, para mais detalhes):

a $|P(A \cap B) - P(A)P(B)| \leq \alpha(k)P(A)$, para todos $A \in {}_{Ft}$ e $B \in Ft_{+k}$ para qualquer t e k, **(b)** $\sum_{k\geq 1}$ $\alpha(k) < \infty$.

Suposição **A2** e **A3** são necessárias para garantir que o estimador seja consistente e que a estrutura de erro possa ser tal que o método CV proposto possa funcionar. Observe que devido à stationaridade de $\{yt\}^{nt}{}_{=1}$, as condições para a consistência de $\varphi^{\cdots}$ e $\varphi^{\cdots}{}_{-t}$ são equivalentes. Em relação ao **A2**, para que o estimador de licença-um-out seja consistente, há uma variedade de condições sugeridas na literatura, por exemplo, Andrews (1987). Por exemplo, o seguinte conjunto de condições garante a consistência de $\varphi^{\cdots}{}_{t}$.

(i) O espaço do parâmetro **Φ** no qual φ pertence é um subconjunto compacto de Rk. (ii) $g(\cdot, \varphi)$ é contínuo e diferenciado em φ para todos (yt, xt). (iii) Definir $Q_{-t}(\varphi)$ como

$$Q_{-t}(\varphi) = \sum_{\substack{n \\ j=p+1 \\ j0 \\ =t}} (yj - g(xj, \varphi))^2.$$

Em seguida, ${}_{n}\text{———}{}^{-1p}{}_{-1Q-t}(\varphi) \rightarrow_{p} Q(\varphi)$ uniformemente em φ, onde $Q(\varphi)$ é uma função não-ostática que atinge um mínimo único em $\varphi = \varphi 0$.

A3.(i) garante que os erros não sejam corrigidos em série, o que será o componente chave para o sucesso da validação cruzada. Vale a pena notar que **A2** e **A3** não são assumptions fortes, mas sim são padrão, por exemplo, quando os métodos de Machine Learning são usados para previsão. Por exemplo, os erros i.i.d. estão aninhados na sequência de diferença martingale. No entanto, erros com valores discretos, digamos $\varepsilon_t \in \{-1, 1\}$ não satisfazem **A3**.

Teorema 1. *Suponha que as suposições **A1-A3** se mantenham. Então*

$$\text{EP}... \rightarrow_{p} EP. \tag{4}$$

Prova. Veja o apêndice. ■

Além disso, consideramos o que aconteceria se as suposições acima mencionadas fossem violadas. Para começar, como o método CV empregado faz uso de características de resíduos, os resíduos devem emular termos de erro para o sucesso deste tipo de CV. Suponha que **a A1** seja violada, ou seja, o modelo é mal especificado. Os resíduos não terão mais correlação serial zero. Em seguida, a violação do **A3** segue naturalmente, uma vez *que* ε_t não é mais um MDS, o que torna Lemma 1 no apêndice inválido.
Portanto, o CV não funciona mais. Isso é declarado implicitamente em nossa prova, uma vez que assumimos que $yt = g(_{xt,\varphi}) + \varepsilon_t$

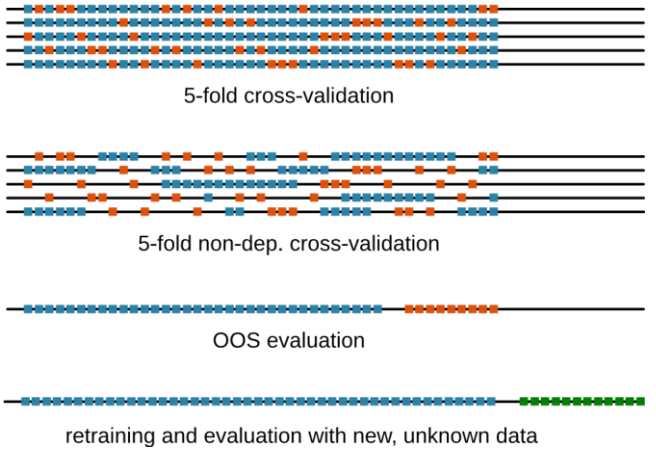

Fig. 2. Treinamento e conjuntos de testes usados para os experimentos. Os pontos azul e laranja representam valores no conjunto de treinamento e teste, respectivamente. Os pontos verdes representam dados futuros não disponíveis no momento da construção do modelo. (Para interpretação das referências t o cor nesta legenda de figura, o leitor é referido à versão web deste artigo.)

onde ε_t é um MDS. Como **a A2** garante que a PE se aproxime da PE de perto através da consistência do estimador OLS de ϕ, sua violação invalida o CV proposto não apenas para a configuração dependente, mas também para a configuração independente.

Por fim, também notamos que o modelo (3) é bastante geral. Desde que os erros não sejam corrigidos em série, Teorema 1 e sua prova garantem que o CV proposto seja válido para uma regressão paramétrica não linear linear. Sua validade é estendida a modelos considerados ''não paramétricos'' como linhas de regressão, splines de suavização, regressão local e redes neurais, que podem ser representadas por (3).

4. Estudo experimental de Monte Carlo

Realizamos experimentos de Monte Carlo ilustrando as consequências do Teorema 1. Na sequência, discutimos a configuração geral de nossos experimentos, bem como as medidas de erro, procedimentos de seleção de modelos, algoritmos de previsão e processos de geração de dados empregados. Os experimentos são realizados utilizando a linguagem de programação R (R Core Team, 2014).

4.1. Configuração geral dos experimentos

O design experimental é baseado na configuração de Bergmeir et al. (2014). Cada série de tempo é particionada into um conjunto disponível para o preditor, chamado *de in-set*, e uma parte do final da série não disponível nesta fase (o *out-set*), que é considerado o futuro desconhecido. Em nossos experimentos, usamos séries com um comprimento total de 200 valores, e usamos 70% dos dados (140 observações) como no conjunto, o resto (60 observações) é retido como o out-set. O conjunto é particionado de acordo com o procedimento de seleção do modelo, os modelos são construídos e avaliados, e a PE é calculada. Desta forma, temos uma estimativa do erro no in-set. Em seguida, os modelos são construídos usando todos os dados do conjunto in-set e avaliados nos dados out-set. O erro nos dados out-set é considerado o verdadeiro erro PE, que estamos estimando por PE. A Fig. 2 ilustra o procedimento. Análogo à prova teórica, podemos calcular PE como um erro quadrado médio (MSE) da seguinte forma:

$$\mathrm{PE}(\varphi, x_{T-20}) = \frac{1}{n}\sum_{t=1}^{n}\{y_{T+t} - \boldsymbol{g}(\boldsymbol{x}_{T+t}, \varphi)\}^2, \qquad (5)$$

onde T é o fim do conjunto in-set e, portanto, $\{y_{T+t}, \boldsymbol{x}_{T+t}\}$, $t = 1, 2, \ldots, n$, compreende o out-set.

A questão em análise é o quão bem a PE estima PE. Avaliamos isso avaliando o erro entre a PE e a PE, em todas as séries envolvidas nos experimentos. Usamos um erro absoluto médio (MAE) para avaliar o tamanho do efeito e chamar essa medida de ''erro de precisão preditiva absoluta'' (MAPAE). É calculado da seguinte forma:

$$\mathrm{MAPAE} = \frac{1}{m}\sum_{j=1}^{m}\left|\widehat{\mathrm{PE}}_j(\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{x}}) - \mathrm{PE}_j(\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{x}})\right| \qquad \ldots \qquad \boldsymbol{x}t),$$

onde m é o número de séries no estudo Monte-Carlo. Além disso, para ver se algum viés é presemente, usamos a média do erro de precisão preditiva (MPAE), definido análogamente como

$$\mathrm{MPAE} = \frac{1}{m}\sum_{j=1}^{m}\left(\widehat{\mathrm{PE}}_j(\hat{\boldsymbol{\theta}}, \boldsymbol{x}t) - \mathrm{PE}_j(\varphi, 2.x\,').\right.$$

4.2. Medidas de erro

Para a prova teórica era conveniente usar o MSE, mas nos experimentos usamos o erro quadrado médio raiz (RMSE), definido como

$$\mathrm{PERMSE}(\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{x}}) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\{y_{T+t} - \boldsymbol{g}(\boldsymbol{x}_{T+t}, \hat{\boldsymbol{\theta}})\}^2}. \qquad (6)$$

O RMSE é mais comum em aplicações, pois opera na mesma escala que os dados, e como a raiz quadrada é uma função bijetiva nos números reais não negativos, a teoria desenvolvida também se mantém para a RMSE. Além disso, também realizamos experimentos

com o MAE, para explorar se os achados teóricos podem se aplicar a essa medida de erro como well. Neste contexto, o MAE é definido como

$$\text{PEMAE}(\Phi\cdots, \boldsymbol{x}') = \frac{\sum_{1}^{n} \left| y_{T+t} - \boldsymbol{g}(\boldsymbol{x}_{T+t}, \hat{\boldsymbol{\theta}}) \right|}{t=1} \quad (7)$$

4.3. Procedimentos de seleção de modelos

Os seguintes procedimentos de seleção de modelos são usados em nossos experimentos:

CV de 5 vezes Isso denota cv 5 vezes normal, para o qual as linhas de uma série de tempo incorporada são atribuídas aleatoriamente a dobras.

LOOCV Deixe-se um cv. Isso é muito semelhante ao procedimento *cv de 5 vezes* , mas o número de dobras é igual ao número de linhas na matriz incorporada, de modo que cada dobra consiste em apenas uma linha da matriz.

nãoDepCV CV não dependente. As mesmas dobras são usadas como para o *CV de 5 vezes*, mas as linhas são removidas do conjunto de training se tiverem uma distância de lag menor que *p* de uma fileira no conjunto de testes, onde *p* é a ordem do modelo máximo (5 em nossos experimentos).

LESTE Avaliação clássica fora da amostra, onde um bloco de dados do final da série é usado para avaliação.

4.4. Algoritmos de previsão

Nos experimentos, considera-se uma previsão de um passo à frente. Encaixamos as sessões lineares automáticas com até 5 valores defasados, ou seja, ar(1) para modelos AR(5). Além disso, como método não linear, usamos um modelo padrão de rede neural perceptron (MLP) de várias camadas, disponível em R no pacote nnet. Ele usa o algoritmo BFGS para montagem de modelo, e definimos os parâmetros da network para um tamanho de 5 unidades ocultas e uma decadência de peso de 0,00316. Para o MLP, são utilizados até 5 valores defasados.

Usamos esses métodos relativamente simples porque são suficientes para mostrar que o CV funciona nesse contexto. Os experimentos poderiam ser facilmente expandidos para outros métodos populares, como regressão vetorial de suporte, Florestas Aleatórias, etc. Vale ressaltar que os resultados não mudariam muito. Isso porque, independentemente dos modelos e métodos utilizados para a previsão, nossa abordagem está preocupada com os resíduos (ver Assunção **A3** na Seção 3), e, portanto, o termo de perturbação ou ruído, que têm um conjunto universal de características como zero média e zero correlações entre erros. Isso é importante para que essas características garantam a justificativa de usar o standard K-fold CV sem modificação.

4.5. Processos geradores de dados

Implementamos três experimentos diferentes, cada um envolvendo 1000 testes de Monte Carlo. Nos dois primeiros experimentos, geramos dados de processos estacionários de AR(3) e processos invertíveis de MA(1), respectivamente. Usamos o design estocástico de Bergmeir et al. (2014), para que para cada ensaio de Monte Carlo, novos coeficientes para o processo de geração de dados (DGP) sejam gerados, e possamos explorar áreas maiores do espaço parâmetro e alcançar resultados mais gerais, não relacionados aos parâmetros/coeficientes de um DGP particular.

O objetivo do Experimento 1 com processos AR(3) é ilustrar como os métodos funcionam quando o modelo verdadeiro, ou modelos muito semelhantes ao DGP, são usados para previsão. O Experimento 2 mostra uma situação em que o verdadeiro modelo não está entre os modelos de forecasting, mas os modelos ainda podem se encaixar razoavelmente bem nos dados. Este é o caso de um processo de MA que pode aproximar um processo AR com um grande número de lags. Na prática, geralmente um número relativamente baixo de lags AR é suficiente para modelar tais dados.

O terceiro experimento é um contraexamplo; ou seja, uma situação em que os procedimentos de CV se desembaram. Utilizamos um processo AR sazonal como o DGP com um lag 12 significativo (lag sazonal 1). Como os modelos levados em conta só usam até os primeiros cinco lags, os modelos não devem ser capazes de encaixar bem tais dados. Obtemos os parâmetros para o DGP, encaixando um modelo AR sazonal a uma série temporal que mostra totais mensais de mortes acidentais nos EUA, de 1973 a 1978 (Brockwell e Davis, 1991). Este conjunto de dados está incluído em uma instalação padrão de R. É ilustrado em Fig. 3. Usamos o modelo AR sazonal como um DGP nos experimentos de Monte Carlo.

Todas as séries nos experimentos são totalmente positivas subtraindo o mínimo e adicionando 1.

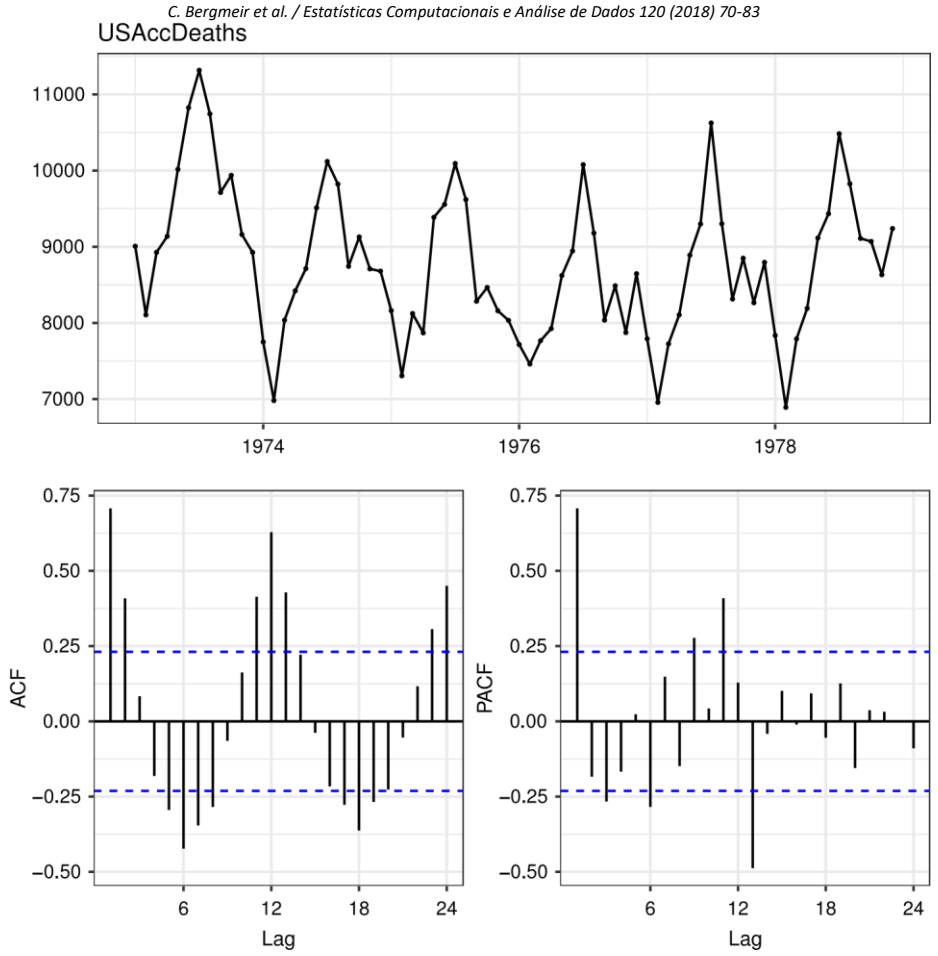

Fig. 3. Série usada para obter um DGP para os experimentos de Monte Carlo. As parcelas ACF e PACF mostram claramente a sazonalidade mensal nos dados.

5. Resultados e discussão

A seguir, apresentamos resultados e oferecemos discussões sobre os achados.

5.1. Resultados para a montagem do modelo linear

O painel superior da Tabela 1 mostra os resultados do Experimento 1, onde os processos AR(3) são usados como DGPs. Vemos que, para a RMSE, os procedimentos *de CV* e *LOOCV* 5 vezes alcançam valores em torno de 0,09 para mapae, enquanto o procedimento *OOS* tem valores mais elevados em torno de 0,16, para que os procedimentos cv atinjam estimativas de erro mais precisas. O procedimento *não DeepCV* tem um desempenho consideravelmente pior, o que se deve ao fato de que os modelos equipados são menos precisos, pois são equipados com menos dados. Em relação ao MPAE, o *LOOCV* alcança estimativas consistentemente baixas com valores absolutos inferiores a 0,003, enquanto *o OOS* e *o CV 5 vezes* têm valores absolutos de até 0,01 e 0,007, respectivamente, comparáveis entre si. Os achados são semelhantes para o MAE.

O painel middle da Tabela 1 mostra os resultados do Experimento 2, onde usamos processos MA(1) como os DGPs. Em relação ao MAPAE, vemos resultados semelhantes como no Experimento 1; ou seja, os procedimentos cv produzem estimativas de erro mais precisas do que o procedimento *OOS*. Em relação ao MPAE, o *OOS* agora supera ligeiramente os procedimentos cv com valores absolutos entre 0,003 e 0,01. Os procedimentos cv têm valores entre 0,008 e 0,013 e 0,005 e 0,011, respectivamente. Achados semelhantes são para a medida de erro MAE. O procedure *não-FepCV* novamente não é competitivo.

Finalmente, o painel inferior da Tabela 1 mostra os resultados do Experimento 3. Neste experimento, onde todos os modelos são fortemente mal especificados, vemos que a vantagem dos procedimentos de CV w.r.t. MAPAE quase desapareceu, e as estimativas de CV são mais tendenciosas do que as estimativas obtidas com o procedimento *OOS*.

Tabela 1

Modelo montado: AR linear. Comprimento da série: 200.

		#	RMSE		MAE	
			MAPAE	MPAE	MAPAE	MPAE
DGP: AR(3)						
5 vezes	CVAR(1)	0.098-				
0.0000.084–		AR(2)	0.089	0.004	0.078	0.000
		AR(3)	0.090	0.006	0.077	0.001
		AR(4)	0.092	0.006	0.078	0.001
		AR(5)	0.094	0.007	0.080	0.001
		AR(1)	0.098	−0,002	0.084	−0,005
		AR(2)	0.089	0.002	0.077	−0,002
		AR(3)	0.090	0.002	0.077	−0,002
		AR(4)	0.091	0.001	0.078	−0,003
		AR(5)	0.093	0.001	0.079	−0,003
		AR(1)	0.423	0.411	0.283	0.271
		AR(2)	0.510	0.505	0.341	0.336
		AR(3)	0.630	0.628	0.419	0.418
		AR(4)	1.014	1.014	0.620	0.619
		AR(5)	6.137	6.137	2.580	2.580
					0.004	
LOOCV						
nãoDepCV						
LESTE		AR(1)	0.170	−0,010	0.143	−0,002
		AR(2)	0.157	−0,004	0.134	0.002
		AR(3)	0.158	−0,002	0.135	0.003
		AR(4)	0.160	−0,002	0.136	0.003
		AR(5)	0.163	−0,001	0.139	0.003
DGP: MA(1)						
5 vezes	CVAR(1)	0.1130.0130.0960.007				
		AR(2)	0.106	0.008	0.090	0.003
		AR(3)	0.102	0.012	0.087	0.007
		AR(4)	0.100	0.009	0.085	0.005
		AR(5)	0.100	0.012	0.085	0.006
LOOCV		AR(1)	0.113	0.011	0.096	0.005
		AR(2)	0.105	0.005	0.090	0.001
		AR(3)	0.101	0.009	0.086	0.004
		AR(4)	0.099	0.005	0.084	0.001
		AR(5)	0.098	0.007	0.084	0.002
nãoDepCV		AR(1)	0.264	0.244	0.201	0.179
		AR(2)	0.359	0.352	0.266	0.258
		AR(3)	0.482	0.480	0.343	0.340
		AR(4)	0.862	0.861	0.545	0.543
		AR(5)	10.225	10.224	4.038	4.037
LESTE		AR(1)	0.192	−0,010	0.161	−0,002
		AR(2)	0.181	−0,006	0.153	−0,001
		AR(3)	0.173	−0,003	0.145	0.003

	AR(4)	0.171	−0,003	0.144	0.004
	AR(5)	0.171	−0,005	0.143	0.001
DGP: AR(12)					
5 vezes	CVAR(1)150.890-43.549128.103-41.810				
	AR(2)	154.210	−50.193	129.217	−46.432
	AR(3)	158.004	−59.821	132.424	−54.444
	AR(4)	166.364	−80.904	139.967	−71.794
	AR(5)	172.824	−95.194	145.233	−83.965
	AR(1)	150.661	−44.063	127.822	−42.250
	AR(2)	154.152	−52.161	129.122	−47.950
	AR(3)	157.858	−62.983	132.123	−56.868
	AR(4)	166.496	−84.866	139.968	−74.659
	AR(5)	173.410	−100.682	145.731	−88.044
LOOCV					
	não DepCVAR (1)206.934126.776159.91684.506			AR(2)	245.753187.501181.995126.540
	AR(3)332.597292.792223.966184.539				
	AR(4)690.263664.060382.009354.904				
	AR(5)8101.9538090.2373090.7193077.161				

(*continuou na próxima página*)

Tabela 1 (*continuada*)

	#	RMSE		MAE	
		MAPAE	MPAE	MAPAE	MPAE
LESTE	AR(1)	157.690	−25.556	135.948	−16.516
	AR(2)	161.896	−28.484	138.869	−18.247
	AR(3)	165.107	−34.500	140.313	−22.344
	AR(4)	171.390	−40.088	145.932	−25.820
	AR(5)	177.577	−42.417	152.985	−28.486

5.2. Resultados para a montagem do modelo MLP

A Tabela 2 mostra os resultados análogos onde as redes neurais têm sido usadas para a previsão. Os experimentos confirmam essencialmente os achados dos modelos lineares. Se apenas um valor defasado for utilizado, o procedimento de montagem do modelo tem dificuldades e os modelos resultantes não são competitivos, rendendo altos valores tanto do MAPAE quanto do MPAE em todos os procedimentos e experimentos de seleção do modelo.

Para o primeiro experimento, os métodos cv mostram vantagens no sentido de que produzem estimativas de erro mais precisas (mapae inferior) e um viés comparável (medido pelo MPAE) em comparação com o procedimento *OOS*.

Essas vantagens também são vistas no Experimento 2. Para o Experimento 3, onde os modelos são fortemente mal especificados, as vantagens dos procedimentos de CV para mapae desapareceram principalmente e as vantagens de alto viés prevalecem.

6. Exemplo de dados do mundo real

Além dos experimentos de Monte Carlo, realizamos nesta seção um estudo com uma série de dados do mundo real. Usamos a conhecida série anual de manchas solares, disponível, por exemplo, de Tong (1993) e in R como mancha solar em série.ano no pacote base. Contém as quantidades de manchas solares anuales de 1700 a 1988, com 289 observações no total. Seguindo Tong (1993), transformamos os dados como uma etapa de pré-processamento da seguinte forma:

$$_{yt} = 2(1 + xt)^{0.5} - 1.$$

Aqui, *xt* é a série temporal original e *yt* é a versão transformada, mostrada em Fig. 4.

Seguindo a configuração de nossos experimentos em Monte Carlo, usamos 30% dos dados como out-set, executamos *cv 5 vezes* e executamos a avaliação *OOS* de acordo com 20% dos dados in-set. Ou seja, o out-set contém os últimos 86 valores da série, e o conjunto de testes *OOS* contém os últimos 40 valores do conjunto in-set.

Tong (1993) descreve que o critério AIC para um modelo linear escolhe uma ordem de lag de 9, e que os métodos de pré-idade não lineares superam os lineares para esta série temporal. Em particular, ele se encaixa em um modelo ar limiar auto-emocionante da ordem 11.

Seguindo a sugestão de que modelos não lineares são apropriados para esta série, focamos aqui no modelo de rede neural como usado em nossos experimentos de Monte Carlo. Essa escolha também é motivada pelo fato de não estarmos interessados principalmente em performance modelo, mas queremos avaliar os procedimentos de seleção do modelo, e com um modelo com mais hiperparômetros permitirá melhores insights. Usamos a seguinte grade de parâmetros, que está em consonância com os experimentos de Monte Carlo e já foi usada em experimentos semelhantes antes (Bergmeir e Benítez, 2012). A grade do parâmetro é construída a partir do tamanho = {3, 4, 5,..., 15}, decaimento = {0, 0. 00316, 0. 0147, 0. 05, 0. 075, 0. 1}, e ordem de lag máxima *p* = {1, 2, 3,..., 20}. Como modelo de seleção, usamos CV e *OOS 5 vezes* como nos experimentos de Monte Carlo.

A validação cruzada nos dá uma previsão genuína fora da amostra para cada ponto de dados. Usamos as previsões em todas as dobras para construir uma série residual e aplicamos o teste Ljung-Box (Ljung e Box, 1978 implementado em R em função Box.test) para esta série. Como estamos considerando previsões fora da amostra, estabelecemos graus de liberdade para 0. Como mencionado anteriormente, nos modelos de literatura com uma ordem máxima de lag de 11 são montados nesta série, por isso definimos a quantidade de lags no teste Ljung-Box para 20, para capturar a autocorrelação restante relevante.

Todas as combinações possíveis da grade de hiperdímetros definidas nos dão 1560 configurações de modelos diferentes no total. Para cada uma dessas configurações de modelo, temos uma série residual descrita acima que é verificada com o teste Ljung-Box para correlação serial. No total, as séries residuais de 763 configurações de modelo passam pelo teste. A partir dessas configurações de modelo, escolhemos a configuração que resulta no RMSE validado transversal minimal, e aquele que resulta no *Mínimo OOS* RMSE. Os resultados são mostrados na Tabela 3.

O modelo escolhido por *CV 5 vezes* tem parâmetros *p* = 7 e tamanho = 3, e resulta em um erro de 2.247. Compare isso com o modelo selecionado pelo procedimento *OOS*, que tem parâmetros *p* = 3 e tamanho = 9, e um erro ligeiramente maior de 2.281. Notamos que a configuração do último modelo é a configuração que dá o melhor erro *OOS* não apenas entre as configurações que passam no teste Ljung-Box, mas entre todas as configurações utilizadas nos experimentos.

Vemos que o procedimento de validação cruzada escolhe uma configuração de modelo que tenha uma estrutura de lag razoável e esteja com apenas 3 unidades escondidas não superequipadas. O modelo selecionado usando *OOS* uses menos lags e unidades mais ocultas, o que parece ser uma escolha inferior, no entanto, ambas as configurações do modelo têm desempenho semelhante no out-set, com o modelo escolhido pelo *OOS* cedendo a um erro um pouco maior de out set.

Tabela 2
Modelo montado: redes neurais. Comprimento de series: 200.

	#	RMSE		MAE	
		MAPAE	MPAE	MAPAE	MPAE
DGP: AR(3)					
CV de 5 vezes	AR(1)	0.769	−0,724	0.665	−0,632
	AR(2)	0.151	0.027	0.107	0.009
	AR(3)	0.195	0.040	0.133	0.017
	AR(4)	0.207	0.057	0.135	0.028
	AR(5)	0.258	0.075	0.159	0.040

LOOCV	AR(1)	0.769	−0,727	0.663	−0,632
	AR(2)	0.152	0.009	0.109	−0,005
	AR(3)	0.170	0.028	0.118	0.005
	AR(4)	0.201	0.033	0.129	0.012
	AR(5)	0.205	0.018	0.135	−0,004
nãoDepCV	AR(1)	0.580	−0,109	0.505	−0,207
	AR(2)	0.844	0.838	0.606	0.605
	AR(3)	0.885	0.862	0.659	0.643
	AR(4)	0.842	0.826	0.643	0.639
	AR(5)	0.771	0.750	0.603	0.597
	AR(1)	0.296	−0,265	0.258	−0,237
	AR(2)	0.157	0.012	0.115	0.002
	AR(3)	0.178	0.019	0.122	0.007
	AR(4)	0.185	0.030	0.124	0.017
	AR(5)	0.176	0.000	0.120	−0,012
	AR(1)	0.352	0.215	0.256	0.107
	AR(2)	0.712	0.704	0.533	0.531
	AR(3)	0.761	0.755	0.589	0.587
	AR(4)	0.727	0.719	0.578	0.575
	AR(5)	0.659	0.649	0.534	0.532
LESTE	AR(1)	0.818	−0,729	0.693	−0,619
	AR(2)	0.232	0.008	0.180	0.013
	AR(3)	0.262	0.011	0.198	0.016
	AR(4)	0.295	0.002	0.215	0.013
	AR(5)	0.326	0.013	0.240	0.024
DGP: MA(1)					
CV de 5 vezes	AR(1)	0.289	−0,253	0.252	−0,228
	AR(2)	0.159	0.027	0.114	0.012
	AR(3)	0.182	0.068	0.125	0.043
	AR(4)	0.187	0.061	0.127	0.038
	AR(5)	0.204	0.065	0.132	0.040
LOOCV					
nãoDepCV					

LESTE	AR(1)	0.368	−0,285	0.305	−0,239
	AR(2)	0.247	0.005	0.191	0.013
	AR(3)	0.267	0.015	0.198	0.021
	AR(4)	0.276	0.006	0.206	0.016
	AR(5)	0.284	0.042	0.217	0.047
DGP: AR(12)					
CV de 5 vezes	AR(1)	154.137	−37.919	130.981	−37.587
	AR(2)	157.743	−49.764	132.260	−45.950
	AR(3)	152.300	−38.585	127.494	−38.171
	AR(4)	155.298	−55.475	130.786	−51.674
	AR(5)	158.104	−62.537	132.781	−58.896
LOOCV	AR(1)	153.873	−38.041	130.153	−37.107
	AR(2)	162.565	−63.654	135.013	−56.950
	AR(3)	169.791	−82.204	140.290	−70.038
	AR(4)	163.306	−69.063	136.467	−61.480
	AR(5)	164.931	−79.233	136.795	−70.160
nãoDepCV	AR(1)	195.267	103.594	156.112	68.756
	AR(2)	195.145	96.722	155.378	64.679
	AR(3)	204.659	125.448	158.247	84.718
	AR(4)	208.368	136.821	162.852	96.170
	AR(5)	205.205	125.914	162.812	89.028

(*continuou na próxima página*)

Tabela 2 (*continuada*)

	#	RMSE		MAE	
		MAPAE	MPAE	MAPAE	MPAE
LESTE	AR(1)	159.824	−23.745	137.788	−15.963
	AR(2)	163.665	−33.175	139.471	−21.946
	AR(3)	168.314	−17.899	141.165	−9.545
	AR(4)	173.594	−28.360	144.881	−17.344
	AR(5)	179.041	−29.841	150.671	−18.436

Tabela 3
Configurações de modelo escolhidas e erros de saída nos dados da mancha solar anual, usando avaliação de CV e OOS 5 vezes.

	Parâmetros			RMSE
	p	Tamanho	Decair	Out-set
CV	7	3	0.1	2.247
LESTE	3	9	0	2.281

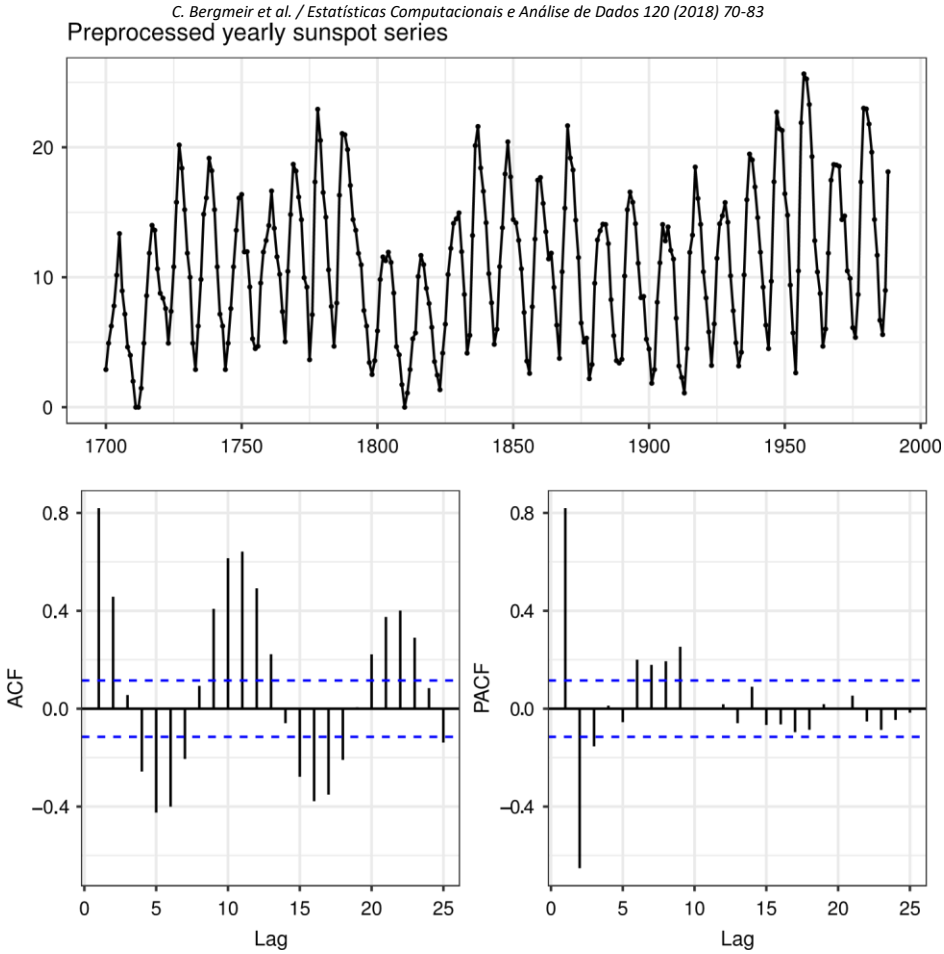

Fig. 4. Série anual de manchas solares após o pré-processamento.

7. Conclusões

Neste trabalho temos investigado o uso de procedimentos de validação cruzada para avaliação de previsões de séries temporais quando são utilizados modelos puramente autoregressivos, situação muito comum; por exemplo, usando procedimentos de Machine Learning para previsão de séries temporáticas. Em uma prova teórica, mostramos que um procedimento normal de validação cruzada k-fold pode ser usado se os resíduos do nosso modelo não forem corrigidos, o que é especialmente o caso se o modelo nests um modelo apropriado. Nos experimentos de Monte Carlo, mostramos empiricamente que, mesmo que a estrutura de lag não esteja correta, desde que os dados sejam bem encaixados pelo modelo, a validação cruzada sem qualquer modificação é uma escolha melhor do que a evaluuação OOS. Temos, então, em um exemplo de dados do mundo real mostrado como esses achados podem ser usados em uma situação prática. A validação cruzada pode controlar adequadamente o excesso de adequação nesta aplicação, e somente se os modelos se adequarem aos dados e levarem a erros fortemente correlacionados, são os procedimentos de validação cruzada a serem evitados, pois em tal caso podem produzir uma subestimação sistemática do erro. No entanto, este caso pode ser facilmente detectado verificando os resíduos para correlação serial, por exemplo, usando o teste Ljung-Box.

Reconhecimento

Koo reconhece a ajuda financeira do Australian Research Council Discovery Grant No. DP150104292.

Apêndice. Prova de Teorema 1

Prova do Teorema 1. Suponha que **a A1** esteja satisfeita e, portanto, a ordem do processo autoregressivo é p. Seguindo Burman e Nolan (1992),

$$\begin{aligned} \text{PE} &= \int [\quad g(\boldsymbol{x}', \varphi\ldots) - e\ 2dF \\ &\approx \int \left[g(\tilde{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}) - g(\tilde{\boldsymbol{x}}, \boldsymbol{\theta})\right]^2 dF + \int \tilde{\varepsilon}^2 \\ &= \int \left[g(\tilde{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}) - \mathrm{E}g(\tilde{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}) + \mathrm{E}g(\tilde{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}) - g(\tilde{\boldsymbol{x}}, \boldsymbol{\theta})\right]^2 dF + \int \tilde{\varepsilon}^2 \end{aligned}$$

onde F é a distribuição do processo $\{y'_k\}^{nk}_{=1}$. Portanto, com um pouco de álgebra, PE se torna

$$g(\boldsymbol{x}', \phi\ldots) - Eg(\boldsymbol{x}\ \phi, \varphi)\ dF + \quad \mathrm{E}g(\boldsymbol{x}\ \varphi), - g(\boldsymbol{x}\ \phi, \varphi)\ dF + \quad \varepsilon\ 2dF, \tag{A.1}$$

enquanto, em uma veia semelhante, PE corresponde a ...

$$\int [\qquad\qquad\qquad]_2 \qquad \int [\qquad\qquad\qquad\qquad]_2 \qquad \int$$

$$\frac{1}{n-p} \sum_{t=p+1}^{n} \left[g(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_{-t}) - \mathrm{E}g(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_{-t})\right]^2 + \int \left[\mathrm{E}g(\boldsymbol{x}, \hat{\boldsymbol{\theta}}_{-t}) - g(\boldsymbol{x}, \boldsymbol{\theta})\right]^2 dF_n + \int \varepsilon^2 \qquad dFn, \tag{A.2}$$

onde Fn é a distribuição empírica da amostra de teste. O segundo e terceiro termos das duas equações acima, (A.1) e (A.2) são assintáticomente idênticos. Então podemos nos concentrar no primeiro termo de cada equação. O primeiro termo de (A.1) torna-se

$$\frac{1}{(n-p)} 2 \quad \sum\sum \varepsilon\ \mathrm{j}(\varphi\ldots)\varepsilon\ \mathrm{t}(\varphi\ldots) = n \quad \sum \tilde{\varepsilon}_j^2(\hat{\boldsymbol{\theta}}) + \frac{1}{(n-p)(n-p-1)} \sum \sum \tilde{\varepsilon}_j(\hat{\boldsymbol{\theta}})\tilde{\varepsilon}_t(\hat{\boldsymbol{\theta}}) \quad -1\ p \quad , \tag{A.3}$$

$$t=p+1\ j=p+1\ j=p+1\ j=p+1\ \mathrm{j=p+}_1\ t=p+1\ ji=t$$

que o primeiro termo (A.2) é $(n-p)^{-1}\sum_{j=p+1}^{n}\varepsilon_j^2(\hat{\boldsymbol{\theta}}_{-t})$. Intuitivamente, o limite de probabilidade do primeiro termo em (A.3) é assintáticamente equivalente ao limite de probabilidade do primeiro termo de (A.2), e, devido a deixar toda a linha de fora, o segundo termo de (A.3) é uma sequência de diferença martingale que converge a zero na probabilidade. Isso será mostrado em Lemma 1. Vale ressaltar que isso é compatível com a condição que Burman e Nolan (1992) fornecem sob sua configuração; ou seja, para qualquer $< j$,

$$\mathrm{E}\left[\varepsilon_t \varepsilon_j \mid \boldsymbol{x1}, \ldots, \boldsymbol{xj}\right] = 0. \tag{A.4}$$

Ou seja, o uso da validação cruzada padrão em nossa configuração é válido desde que

$$\mathrm{E}\sum_{t=p+1}^{n}\sum_{j=p+1\mathrm{j}}^{n} \tilde{\varepsilon}_j(\hat{\boldsymbol{\theta}})\tilde{\varepsilon}_t(\hat{\boldsymbol{\theta}}) \approx \mathrm{E}\sum_{=p+1} \varepsilon_j^2(\hat{\boldsymbol{\theta}}_{-t}), \tag{A.5}$$

que segue de Lemma 1. ■

Lemma 1. *Suponha que* ***a A1-A3*** *espere. Então, existe uma constante arbitrariamente pequena c >* 0 *de tal forma que*

$$\sum_{t=p+1}^{n}\left[\sum_{j=p+1}^{n}\sum_{t=p+1}^{n}\tilde{\varepsilon}_j(\hat{\boldsymbol{\theta}})\tilde{\varepsilon}_t(\hat{\boldsymbol{\theta}})-\sum_{j=p+1}^{n}\varepsilon_j^2(\hat{\boldsymbol{\theta}}_{-t})\right]^4 \le cn^4 \quad E.$$

Prova de Lemma 1. Note que

$$\sum_{t=p+1}^{nnn}\sum_{j=p+1}^{1j}\varepsilon_j(\varphi')\varepsilon_t(\varphi)) - \sum_{=p+1}^{}\varepsilon_j^2(\hat{\boldsymbol{\theta}}_{-t})$$

$$= \underbrace{\sum_{j=p+1}\tilde{\varepsilon}_j^2(\hat{\boldsymbol{\theta}})}_{\text{A. 6. 1}} - \underbrace{\sum\varepsilon_j^2(\hat{\boldsymbol{\theta}}_{-t})}_{\text{A. 6. algarismo}} + \underbrace{\sum_{j=p+1}\sum_{j=p+1}\tilde{\varepsilon}_j(\hat{\boldsymbol{\theta}})\tilde{\varepsilon}_t(\hat{\boldsymbol{\theta}})}_{t=p+1\,/\,=t}^{nnnn} \quad (A.6)$$

= *A.* 6. 1 + *A.* 6. 2.

Por *A.* 6. 1,

$$\sum_{=p+1}\tilde{\varepsilon}_j^2(\hat{\boldsymbol{\theta}}) - \sum_{j}\varepsilon_j^2(\hat{\boldsymbol{\theta}}_{-t}) = \underbrace{\sum_{=\ +}\left(\tilde{\varepsilon}_j^2(\hat{\boldsymbol{\theta}}) - \mathrm{E}\tilde{\varepsilon}_j^2(\hat{\boldsymbol{\theta}})\right)}_{\text{A. 6. 1. 1A}} - \underbrace{\sum_{=\ +}\left(\varepsilon_j^2(\hat{\boldsymbol{\theta}}_{-t}) - \mathrm{E}\varepsilon_j^2(\hat{\boldsymbol{\theta}}_{-t})\right)}_{\text{. 6. 1. algarismo}}^{nnnn}$$

$$+\sum_{j=p+1}^{n}\left(\mathrm{E}\tilde{\varepsilon}_j^2(\hat{\boldsymbol{\theta}}) - \mathrm{E}\varepsilon_j^2(\hat{\boldsymbol{\theta}}_{-t})\right)$$

A. 6. 1. 3

Observe que $\varepsilon_j(\varphi)$ e $\varepsilon_j(\varphi^{\cdots}{}_{-t})$ têm a mesma distribuição, desde que tanto $\varphi\ \varphi$ e φ_{-t} sejam consistentes devido a **A1** e **A2**. Portanto, *A.* 6. 1. 3 é de menor ordem que *A.* 6. 1. 1 e *A.* 6. 1. 2. Para *A.* 6. 1. 1, o somado e são sequências de diferença martingale e, portanto, usando a desigualdade de Burkholder, nós obtemos,

$$\left|\sum_{j=p+1}^{n}\left(\tilde{\varepsilon}_j^2(\hat{\boldsymbol{\theta}}) - \mathrm{E}\varepsilon_j^2(\hat{\boldsymbol{\theta}})\right)\right|^4 \le c\mathrm{E}\left|\sum_{t=1}^{n}\left(\tilde{\varepsilon}_j^2(\hat{\boldsymbol{\theta}}) - \mathrm{E}\tilde{\varepsilon}_j^2(\hat{\boldsymbol{\theta}})\right)^2\right|^2 \le cn^2 \quad \text{E.}$$

Por *A.* 6. 1. 2, é análogo. Vamos voltar nossa atenção para *A.* 6. 2. Para *A.* 6. 2, devido à simetria de uma matriz de covariância,

$$\sum_{j=p+1}^{nnnt}\sum_{t=p+1}^{<j}\varepsilon_j(\varphi)\varepsilon_t(\varphi...) = 2\sum_{j=p+1}^{}\varepsilon_j(\varphi)\sum_{t=p+2}^{}\varepsilon_t(\varphi...).$$

Portanto, a prova se resume a mostrar que o impacto dos termos de covariância, ou seja, a soma dos elementos off-diagonal da matriz variância-covariância torna-se insignificante, ou seja.

$$\left[\sum_{j=p+1}^{n}\sum_{t=1}^{n}\varepsilon_j(\hat{\boldsymbol{\theta}})\varepsilon_t(\hat{\boldsymbol{\theta}})\right]^2 = 2\mathrm{E}\left[\sum_{=p+2}^{n}\tilde{\varepsilon}_j(\hat{\boldsymbol{\theta}})\sum_{t=p+1}^{t<j}\tilde{\varepsilon}_t(\hat{\boldsymbol{\theta}})\right]^2 \le cn^4 \quad \text{E.} \quad (A.7)$$

Observe que $\{\varepsilon_t\}$ e $\{\varepsilon_t\}$ são sequências de diferença de martingale estacionárias e qualquer combinação linear de martingales definida na mesma filtragem também é um martingale. O método CV proposto trabalha exatamente para garantir a falta de correção entre os resíduos. Uma vez omitindo linhas of a matriz e aplicando a lei da expectativa iterada, (A.7) mantém-se da mesma forma como na prova de Lemma 5.3 como em Birman e Nolan (1992). ■

Referências

Andrews, D.W., 1987. Consistência em modelos econométricos não lineares: Uma lei uniforme genérica de grandes números. Econometrica 1465-1471.

Arlot, S., Celisse, A., 2010. Um levantamento dos procedimentos de validação cruzada para seleção de modelos. Surv. 4, 40-79.

Bergmeir, C., Benítez, J.M., 2012. Sobre o uso de validação cruzada para avaliação do preditor de séries temporais. Informe Sci. 191, 192-213.

Bergmeir, C., Costantini, M., Benítez, J.M., 2014. Sobre a utilidade da validação cruzada para avaliação de previsão direcional. Computação. Estatal. Data Anal. 76, 132-143.

Borra, S., Di Ciaccio, A., 2010. Medindo o erro de previsão. Um comparison de métodos de validação cruzada, bootstrap e covariância. Computação. Estatal. Data Anal. 54 (12), 2976-2989 .

Brockwell, P.J., Davis, R.A., 1991. Séries temporísticas: Teoria e Métodos. Springer, New York.

Budka, M., Gabrys, B., 2013. Amostragem de preservação de densidade: Robust e alternativa eficiente à validação cruzada para estimativa de erro. IEEE Trans. Neural Netw. Syst. 24 (1), 22-34.

Burman, P., Chow, E., Nolan, D., 1994. Um método multi-validatório para dados dependentes. Biometrika 81 (2), 351-358.

Burman, P., Nolan, D., 1992. Estimativa dependente de dados das funções de previsão. séries da época. 13 (3), 189-207.

Györfi, L., Härdle, W., Sarda, P., Vieu, P., 1989. Estimativa de curva não paramétrica da Série Temporal. Springer Verlag, Berlin.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Elementos da Aprendizagem Estatística. Springer, New York.

Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D., 2008. Previsão with Exponential Smoothing: The State Space Approach. Springer, Berlim, URL http://www.exponentialsmoothing.net.

Kunst, R., 2008. Validação cruzada de modelos de predação para séries temporáticas por bootstrapping paramétrico. Austral. J. Statist. 37, 271-284.

Ljung, G.M., Box, G.E.P., 1978. Em uma medida de falta de ajuste em modelos de séries temporâneas. Biometrika 297-303.

McQuarrie, A.D.R., Tsai, C.-L., 1998. Regressão e seleção de modelos de séries temporcial. Editora Científica Mundial.

Mokkadem, A., 1988. Misturando propriedades dos processos ARMA. Processo estocástico. Appl. 29 (2), 309-315.

Moreno-Torres, J., Saez, J., Herrera, F., 2012. Estudo sobre o impacto do conjunto de dados induzido por partição na validação cruzada k-fold. IEEE Trans. Neural Netw. Syst. 23 (8), 1304-1312 .

Opsomer, J., Wang, Y., Yang, Y., 2001. Regressão não paramétrica com erros correlacionados. Statista. 16 (2), 134-153.

R Core Team, 2014. R: Uma Linguagem e Ambiente para Computação Estatística. Fundação R para Computação Estatística, Viena, Áustria. URL http://www. R project.org/.

Racine, J., 2000. Seleção de modelos transversais consistentes para dados dependentes: validação cruzada de blocos hv. J. Econometria 99 (1), 39-61.

Stone, M., 1974. Escolha transversal e avaliação de previsões estatísticas. J.R. Stat. Soc. Soc. Ser.B Stat. Methodol. 36 (2), 111-147.

Tong, H., 1993. Série temporal não linear: uma abordagem do sistema dinâmico. Clarendon Press.