Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

Mônica Marin de Souza

Investigação de efeitos sazonais na incidência de fraturas femorais no Estado de São Paulo

Maringá - PR
2019

Mônica Marin de Souza

Investigação de efeitos sazonais na incidência de fraturas femorais no Estado de São Paulo

Dissertação apresentada ao Programa de Pós-graduação em Bioestatística do centro de ciências exatas da Universidade Estadual de Maringá como requisito para obtenção do título de mestre em Bioestatística.

Universidade Estadual de Maringá

Programa de Pós-Graduação em Bioestatística

Orientador: Prof. Dr. Edson Zangiacomi Martinez

Coorientador: Profa. Dra. Eniuce Menezes de Souza

Maringá - PR

2019

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá – PR., Brasil)

Souza, Mônica Marin de
S729i Investigação de efeitos sazonais na incidência de fraturas femorais no Estado de São Paulo. / Mônica Marin de Souza. -- Maringá, 2019.
68 f. : il. col., figs., tabs.

Orientador: Prof. Dr. Edson Zangiacomi Martinez.
Co-orientadora: Profª. Drª. Eniuce Menezes de Souza.
Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2019

1. Análise de séries temporais. 2. Inferência Bayesiana. 3. Regressão linear. 4. Modelos aditivos generalizados. I. Martinez, Edson Zangiacomi, orient. II. Souza, Eniuce Menezes de, coorient. III. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. IV. Título.

CDD 21.ed. 519.542

AHS-CRB-9/1065

MÔNICA MARIN DE SOUZA

Investigação de efeitos sazonais na incidência de fraturas femorais no Estado de São Paulo

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

Prof. Dr. Edson Zangiacomi Martinez
Universidade Estadual de São Paulo – FMRP/USP/PBE

Prof. Dr. Robson Marcelo Rossi
Universidade Estadual de Maringá - UEM

Prof. Dr. Altacílio Aparecido Nunes
Universidade Estadual de São Paulo – FMRP/USP

Maringá, 29 de maio de 2019.

Agradecimentos

Agradeço aos professores do Programa de Pós-Graduação em Bioestatística, em especial a Prof. Dra. Isolde Previdelli, meu orientador Prof. Dr. Edson Zangiacomi Martinez, e minha coorientadora Prof. Dra. Eniuce Menezes de Souza, e também meus colegas Aline, Stéfane, Naiara, Patrícia, Jéssica, Vinícius, Everton e Talita.

Em especial, agradeço minha família e amigos.

E à CAPES pelo apoio financeiro.

Lista de ilustrações

Figura 1 – Gráficos das funções de autocorrelação parcial dos resíduos obtidos após ajustar o modelo definido pelas equações 2.7, 2.8 e 2.9 aos dados descritos no Capítulo 3 24
Figure 2 – Observed and predicted time series describing the monthly average number of events of femoral fractures per day, from January 2008 to December 2017, in the State of São Paulo, Brazil. 35
Figure 3 – Graphs for the seasonality function $S(t)$ of the time series, according to the months of the occurrence of the femoral fractures. The gray areas correspond to 95% credible regions. 36
Figure 4 – Graphs for the seasonality function $S(t)$ of the time series, according to the months of the occurrence of the femoral fractures, stratified by age groups and considering the male population. The gray areas correspond to 95% credible regions. 37
Figure 5 – Graphs for the seasonality function $S(t)$ of the time series, according to the months of the occurrence of the femoral fractures, stratified by age groups and considering the female population. The gray areas correspond to 95% credible regions. 38
Figure 6 – Monthly average number of femoral fractures per day. The left panel shows the values for cases involving men while the right panel depicts those occurring with women. In both panels, the circles represent the observed values, and the continuous lines are the values predicted by the generalized additive models. 42

Figure 7 – Cross-validation error as a function of the smoothing penalty λ. The continuous lines show the cross-validation error estimated in a leave-one-out cross-validation strategy to determine the optimal value for the smoothing penalty λ. The optimal $\lambda = \lambda^*$ is the one minimizing the cross-validation error, as indicated in the plots. The left panel shows the results obtained for the time series associated with men femoral fractures, while the right one shows the results for the women data. 44
Figure 8 – Partial autocorrelation functions of the residuals of the adjusted models. The vertical lines show the correlation values and shaded regions represent the 95% confidence interval for a completely random process. We notice that practically all data points are within the random confidence band, indicating that the residuals are uncorrelated. The left panel displays the results for men data, and the right panel shows the same for women data. . 44
Figure 9 – Effects of the month and year on the incidence of femoral fractures. The left panel depicts the estimated smoothing term accounting for the effect of the month $[g_1(\mathit{month}_i)]$ on the fractures for men (purple curve) and women (green curve) of all ages.. The right panel shows the estimated smoothing functions of the year $[g_2(\mathit{year}_i)]$ for men (purple curve) and women (green curve) of all ages. In all plots, the shaded regions represent 95% confidence intervals. 45
Figure 10 – Effects of the month on the incidence of femoral fractures across different age groups. Each panel shows the estimated smoothing term accounting for the effect of the month $[g_1(\mathit{month}_i)]$ on the fractures for men (purple curve) and women (green curve) for a particular age group (indicated in the plots). In all plots, the shaded regions represent 95% confidence intervals. 46
Figure 11 – Effects of the year on the incidence of femoral fractures across different age groups. Each panel shows the estimated smoothing term accounting for the effect of the year $[g_2(\mathit{year}_i)]$ for men (purple curve) and women (green curve) for a particular age group (indicated in the plots). In all plots, the shaded regions represent 95% confidence intervals. 48
Figure 12 – Effects of the number of holidays on the incidence of femoral fractures between two different age groups. The left panel shows the results for men (purple curves) and women (green curves) of ages between 15 and 19 years, while the right panel shows the same for the age group 60-69 yo. We notice the number of holidays has a significant impact on the fractures of men of ages between 15 and 19 yo. In all plots, the shaded regions represent 95% confidence intervals. 49

Lista de tabelas

Table 1 – Incidence of femoral fractures in man and women per 100,000 population per year, State of São Paulo, Brazil, years of 2008, 2011, 2014 and 2017. . 34
Table 2 – Bayesian estimates for the model parameters including the monthly average number of events of femoral fractures per day, period from January 2008 to December 2017, State of São Paulo, Brazil. 35

Sumário

1 Introdução **11**

2 Metodologia **14**

2.1 Séries temporais e modelo de regressão 14

2.2 Inferência Bayesiana 17

2.3 Amostragem de Gibbs 20

2.4 Testes de adequação do modelo de séries temporais 20

2.4.1 Teste de Dickey-Fuller aumentado 21

2.4.2 Função de autocorrelação 23

2.5 Modelos aditivos generalizados 24

3 Seasonal variation of femoral fractures in the State of São Paulo, Southeast Brazil **31**

3.1 Introduction 31

3.2 Methods 32

3.3 Results 33

3.4 Discussion 36

3.5 Conclusion 39

4 Identifying patterns in the incidence of femoral fractures across gender and age in Brazil **40**

4.1 Introduction 40

4.2 Methods 42

4.2.1 Data 42

4.2.2 Generalized additive models 42

4.3 Results and Discussion 43

4.4 Conclusions 48

5 Considerações finais **51**

Referências **53**

Anexos **57**

ANEXO A Normas da Revista de Saúde Pública **58**

Resumo

A ocorrência de fraturas femorais é um importante problema de saúde pública. Esse tipo de condição médica tem se tornado cada vez mais frequente devido, principalmente, ao envelhecimento da população. Sendo assim, é importante a construção de modelos de séries temporais para extrair possíveis componentes de tendência de longo prazo e sazonalidade nos padrões de incidência de fraturas femorais, além da incerteza associada à essas componentes. Nessa dissertação, utilizou-se de duas abordagens para estudar a incidência de fraturas femorais no Estado de São Paulo no período de janeiro de 2008 a dezembro de 2017. Na primeira, propõe-se a construção de um modelo de regressão para séries temporais sob a perspectiva Bayesiana, em que o termo de sazonalidade é modelado por funções periódicas e os erros pelo método de Prais-Winsten. Na segunda abordagem, utilizou-se de modelos aditivos generalizados para descrever as séries temporais, onde a incidência de fraturas do fêmur foi modelada como um função suavizada do mês e ano. Esses modelos empregam funções não paramétricas expressas em termos de uma base formada por funções interpoladoras para extrair as tendências de longo prazo e sazonais diretamente dos dados. Ambas as abordagens mostram que existe uma clara tendência sazonal na incidência de fraturas femorais, na qual há uma maior quantidade de fraturas nos meses mais frios do ano, com diferenças entre homens e mulheres dependendo da faixa etária.

Palavras-chave: análise de séries temporais, inferência Bayesiana, regressão linear, modelos aditivos generalizados.

Abstract

The incidence of femoral fractures is an important public health issue. This medical condition is becoming increasingly frequent mainly because of the population aging. It is thus essential to create time series models for probing possible long-term trends and seasonal patterns in the incidence of femoral fractures, as well as the uncertainties associated with these patterns. In this work, we employ two frameworks for studying the femoral fractures occurring between January 2008 and December 2017 in the State of São Paulo (Brazil). In the first one, we propose a time series regression model with the Bayesian framework where the seasonal term is described in term of periodic functions and errors are modeled using the Prais-Winsten method. In the second approach, we use generalized additive models for describing the time series. These models use non-parametric functions expressed by a basis of smooth functions for extracting the long-term and seasonal trends directly from data. Both approaches show that there is a clear seasonal trend in the incidence of femoral fractures, in which fractures are more likely to occur during the coldest month of the year, with differences between men and women also depending on the age group.

Key-words: Time series analysis, Bayesian inference, linear regression, generalized additive model.

Capítulo 1

Introdução

O uso de séries temporais para investigar detalhes do comportamento dinâmico de sistemas e fenômenos biológicos, sociais, econômicos e físicos é amplamente difundido na comunidade científica (MORETTIN; TOLOI, 2006). Nesses estudos, conjuntos de dados ordenados no tempo, ou seja, uma série temporal, são analisados visando encontrar variáveis que afetem o comportamento das séries, fazer previsões sobre valores futuros, buscar por padrões associados à dinâmica das séries temporais, entre outras análises. No caso da busca por padrões, uma importante característica que pode estar presente em uma série temporal é a sazonalidade. De maneira simples, sazonalidade pode ser definida como a ocorrência de variações cíclicas nos valores de uma série temporal em intervalos de tempo bem definidos, tais como intervalos anuais, semanais e diários em estudos epidemiológicos. A origem de efeitos sazonais em séries temporais usualmente está relacionada a efeitos climáticos ou comportamentos humanos periódicos como férias escolares e feriados. A presença de efeitos sazonais em uma série temporal pode contribuir para a tomada de decisões relacionadas ao sistema em estudo. Por exemplo, o conhecimento de que um certo tipo de emergência médica ocorre com maior frequência num determinado período do ano pode contribuir para o melhor planejamento de escalas de profissionais da saúde e a melhor alocação de recursos em hospitais.

Nesse contexto, um importante problema de saúde pública é a ocorrência de fraturas femorais. Do ponto de vista médico, uma fratura femoral corresponde a uma fratura do fêmur, o osso mais longo e volumoso do corpo humano localizado na parte superior da perna. Trata-se de uma fratura delicada, que frequentemente requer procedimentos cirúrgicos. Esse tipo de fratura é geralmente causado por traumas decorrentes de acidentes de trânsito (os quais costumam apresentar alto impacto), disparos de arma de fogo e quedas. Essa última causa é mais comum em pessoas idosas, sendo a osteoporose um importante fator de risco (DORE et al., 2013). Além disso, o aumento da expectativa de vida e, consequentemente, da população idosa, torna esse tipo de fratura cada vez mais preocupante. Estima-se que 66 mil idosos foram internados

com esse tipo de fratura entre 1997 e 1998 na Inglaterra e no País de Gales (CLAGUE et al., 2002). Já nos Estados Unidos, acredita-se que 350 mil pessoas por ano sofram fraturas femorais, cujo custo anual envolvendo os cuidados clínicos é de aproximadamente seis bilhões de dólares (HANNAN et al., 2001; SAKAKI et al., 2004). No Brasil ocorreram mais de 181 mil fraturas de fêmur entre os anos de 2008 e 2012 (SOARES et al., 2014), gerando um custo de aproximadamente 60 milhões de reais para o Sistemas Único de Saúde (SUS) apenas no ano de 2008 (CUNHA et al., 2008; SOARES et al., 2014). Em adição ao prejuízo financeiro e social, idosos que sofrem fratura femoral costumam apresentar várias complicações e doenças crônicas associadas, as quais podem conduzir a um aumento significativo do risco de morte desses pacientes (BALEN et al., 2001; SAKAKI et al., 2004).

Dada a importância desse tipo de condição médica, vários pesquisadores têm se dedicado à busca de padrões em séries temporais relacionadas às fraturas femorais, sendo a maior parte desses estudos baseada em dados de países do hemisfério norte (CRAWFORD; PARKER, 2003; JENA; MUDDU; RICHARDSON, 2003; CHIU; NG; CHOW, 1996; JACOBSEN et al., 1991; JACOBSEN et al., 1995; PARKER; MARTIN, 1994; PEDRAZZONI et al., 1993). Nestas regiões, observou-se que fraturas femorais apresentam uma maior ocorrência durante os meses de inverno, que no hemisfério norte compreendem aos meses de dezembro, janeiro e fevereiro. No caso do Reino Unido, observa-se ainda uma maior taxa de mortalidade de pacientes com fraturas femorais que são atendidos por hospitais nos meses de inverno (CRAWFORD; PARKER, 2003). Por outro lado, há menos informação sobre esses efeitos sazonais em países do hemisfério sul, em particular no Brasil. Esse trabalho pretende contribuir para essa questão por meio de um estudo dos efeitos sazonais na incidência de fraturas femorais no Estado de São Paulo, entre janeiro de 2008 e dezembro de 2017. Para isso, propõe-se um estudo de séries temporais associadas ao número de fraturas utilizando-se de duas abordagens diferentes, porém complementares. Na primeira abordagem, constrói-se um modelo de regressão que utiliza termos de sazonalidade e considera a autocorrelação diretamente nos erros, pelo Método de Prais-Winsten. Nessa mesma abordagem, utiliza-se a inferência Bayesiana para a estimativa dos parâmetros, a qual gera bandas de credibilidade da sazonalidade de forma mais simples do que a inferência frequentista. Em uma segunda abordagem, emprega-se um modelo de regressão baseado no método conhecido como modelo aditivo generalizado. Nesse caso, o número de fraturas femorais é modelado por uma função suavizada do mês e ano, a qual é expressa em termos de uma base composta por funções interpoladoras, também chamadas *splines*. Esse procedimento é capaz de extrair os padrões sazonais diretamente dos dados, sem que seja preciso pressupor uma forma funcional particular, como na primeira abordagem.

Para realizar esses dois estudos, inicialmente, foi feita uma revisão sobre conceitos e técnicas relacionados à análise de séries temporais no Capítulo 2. Em particular, a Seção 2.1 desse capítulo apresenta uma breve revisão sobre modelos de regressão em séries temporais, incluindo o método de Prais-Winsten, o qual é empregado em uma das abordagens discutidas

nesse trabalho (Capítulo 3) para modelar séries temporais associadas às fraturas femorais. Em seguida, na Seção 2.2 do Capítulo 2, foi apresentada uma revisão dos principais aspectos envolvendo a inferência Bayesiana, visto que o ajuste dos parâmetros do modelo apresentado na Seção 2.1 aos dados das fraturas femorais foi realizado por esta abordagem. Para esse ajuste utilizou-se o programa *OpenBUGS* (LUNN et al., 2009), no qual a sigla *BUGS* refere-se a *Bayesian inference Using Gibbs Sampling* (em tradução livre: "inferência Bayesiana utilizando amostragem de Gibbs"). Por conta disso, a Seção 2.3 do Capítulo 2 faz uma breve descrição do que vem a ser a técnica de amostragem de Gibbs. Na sequência, a Seção 2.4 do Capítulo 2 apresenta alguns testes usados para verificar a adequação do modelo às séries temporais. Encerrando essa revisão, a Seção 2.5 descreve vários aspectos da técnica denominada modelos aditivos generalizados, a qual será a ferramenta principal utilizada como uma segunda abordagem para descrever as séries temporais de fraturas femorais. Os dois Capítulos (3 e 4) que seguem a revisão apresentam nossa contribuição para o estudo de padrões em séries temporais associadas às fraturas femorais. Especificamente, o Capítulo 3 utiliza um modelo de regressão linear com termos de tendência na forma de senos e cossenos para investigar fraturas femorais no Estado de São Paulo entre janeiro de 2008 e dezembro de 2017. O Capítulo 4, por sua vez, analisa os mesmos dados utilizando modelos aditivos generalizados. Por fim, o Capítulo 5 encerra esse trabalho com as conclusões.

Capítulo 2

Metodologia

2.1 Séries temporais e modelo de regressão

O termo série temporal refere-se a um conjunto de dados representado por uma sequência. Exemplos típicos de séries temporais incluem dados financeiros (preço de ações, cotações de moedas, etc.), geológicos e climatológicos (intensidade de terremotos, velocidade do vento, etc.), nos quais o índice da sequência de dados representa o tempo, mas podem incluir também outros dados sequenciais como cadeias de DNA, nos quais o índice da sequência não está associado ao tempo. É essa propriedade de ordem sequencial/temporal que distingue dados de uma série temporal de outros tipos de dados. Em muitos casos, séries temporais representam um conjunto de observação sucessivas igualmente espaçadas no tempo, por exemplo, pode-se ter séries temporais diárias, semanais, mensais, anuais ou envolvendo outra escala temporal. No caso específico do problema que será estudado no Capítulo 3, a série temporal de interesse representa o número de casos de fraturas femorais em cada mês, dividido pelo número de dias do mês, ocorrido no Estado de São Paulo entre janeiro de 2008 e dezembro de 2017.

Uma série temporal é usualmente decomposta ou modelada por três componentes: tendência, sazonalidade e variação aleatória (MORETTIN; TOLOI, 2006). Essa decomposição é normalmente feita de maneira aditiva, isto é, essas três componentes aparecem na forma de soma ao expressar o modelo matemático para série temporal. Entretanto, também é possível construir modelos multiplicativos, nos quais essas componentes são expressas na forma de produtos. A componente de tendência representa os movimentos de "subida" ou "descida" da série temporal ou, de modo mais geral, representam a mudança de longo prazo do valor médio da série temporal. Essa componente é comumente modelada pelo ajuste de uma função polinomial do tempo (modelos de regressão) e pela análise do comportamento passado da série temporal ao redor de um ponto (modelos autorregressivos). A componente de sazonalidade refere-se às oscilações regulares/determinísticas e periódicas com período bem definido. Em muitos casos pode-se modelar uma sazonalidade por meio de modelos de regressão que

incorporem funções periódicas do tempo (senos e cossenos, por exemplo). Por fim, a componente de variação aleatória ou ruído representa as flutuações aleatórias não relacionadas às componentes de tendência e de sazonalidade. Assim, se y_t, em que $t = 1, 2, \ldots, n$, representa uma observação de uma série temporal no tempo t, pode-se escrever

$$y_t = T(t) + S(t) + e_t \,, \tag{2.1}$$

em que $T(t)$ é a componente de tendência, $S(t)$ a de sazonalidade e e_t a aleatória.

Dependendo da forma de $T(t)$, $S(t)$ e e_t, pode-se recuperar vários modelos tradicionais (ou processos estocásticos) de séries temporais. Por exemplo, se $T(t) = y_{t-1}$, $S(t) = 0$ e $e_t = \varepsilon_t$, com ε_t sendo um ruído completamente aleatório de média nula e variância unitário, tem-se o processo chamado de passeio aleatório, isto é,

$$y_t = y_{t-1} + \varepsilon_t \,. \tag{2.2}$$

Por outro lado, se $T(t) = S(t) = 0$ e $e_t = \varepsilon_t + \beta_1\varepsilon_{t-1} + \cdots + \beta_q\varepsilon_{t-q}$, tem-se o processo de médias móveis de ordem q, usualmente abreviado por MA(q) e descrito por

$$y_t = \varepsilon_t + \beta_1\varepsilon_{t-1} + \cdots + \beta_q\varepsilon_{t-q} \,. \tag{2.3}$$

Outra possibilidade é escolher $T(t) = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p}$, $S(t) = 0$ e $e_t = \varepsilon_t$, como ε_t sendo um ruído completamente aleatório de média nula e variância unitária, conduzindo ao processo

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \varepsilon_t \,, \tag{2.4}$$

denominado autorregressivo de ordem p (usualmente abreviado por AR(p)).

Um outro exemplo que combina os dois processos anteriores é o processo autorregressivo de médias móveis, usualmente abreviado por ARMA(p, q) (sendo p a ordem autorregressiva e q a ordem da média móvel) e descrito como

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \varepsilon_t + \beta_1\varepsilon_{t-1} + \cdots + \beta_q\varepsilon_{t-q} \,. \tag{2.5}$$

Observe que quando $\alpha_1 = \cdots = \alpha_q = 0$ recupera-se o modelo de médias móveis e para $\beta_1 = \cdots = \beta_q = 0$ tem-se o modelo autorregressivo. Nos três casos anteriores, os coeficientes α_i $(i = 1, \ldots, p)$ e β_j $(j = 1, \ldots, q)$ são parâmetros dos modelos que devem ser ajustados conforme a série temporal em estudo. Além disso, os dois exemplos anteriores consideram tendências do tipo autorregressiva.

Existe uma outra possibilidade, de maior interesse para o problema tratado no Capítulo 3, a qual corresponde tomar a componente de tendência $T(t)$ como sendo uma função polinomial do tempo. O exemplo mais simples corresponde a considerar um polinômio de grau 1,

$$T(t) = \beta_0 + \beta_1 t \,, \tag{2.6}$$

sendo β_0 e β_1 parâmetros do modelo, o que usualmente é chamado de modelo de tendência linear. O coeficiente β_1 indica uma tendência de crescimento (se $\beta_1 > 0$) ou descrescimento (se $\beta_1 < 0$) do valor médio de y_t ao longo do tempo. Para esse caso, a expressão geral da Equação 2.1 assume a forma

$$y_t = \beta_0 + \beta_1 t + S(t) + e_t \,, \tag{2.7}$$

que corresponde à classe do modelo que será utilizado para descrever o número mensal de fraturas femorais. Conforme será apresentado no Capítulo 3, o modelo empregado ainda considera uma tendência sazonal com período de 12 meses na forma

$$S(t) = \eta_1 \sin\left(\frac{2\pi t}{12}\right) + \eta_2 \cos\left(\frac{2\pi t}{12}\right) \,, \tag{2.8}$$

sendo η_1 e η_2 parâmetros. Além disso, o termo aleatório é expresso na forma de um processo autorregressivo de ordem 1, isto é,

$$e_t = \rho e_{t-1} + \varepsilon_t \,, \tag{2.9}$$

sendo ρ outro parâmetro e ε_t um número aleatório. Testou-se modelos alternativos para os dados considerando que os termos e t seguem um modelo autorregressivo de ordem 2, mas não observou-se uma melhora significativa do ajuste quando comparado com os modelos anteriores. Essas comparações entre os modelos foram baseadas no Critério de Informação da *Deviance* (DIC), critério baseado na esperança *a posteriori* do logaritmo da distribuição condicional dos dados (SPIEGELHALTER et al., 2002).

Portanto, o modelo apresenta estes parâmetros que devem ser ajustados aos dados:

- β_0 é o chamado intercepto, o qual está associado ao hipotético valor inicial da série temporal (note que se $t = 0$, $y_0 = \beta_0 + \eta_2 + e_0$);
- β_1 é o coeficiente que descreve a tendência de longo prazo da série temporal;
- η_1 e η_2 são parâmetros que determinam a forma da tendência sazonal da série;
- ρ é o coeficiente que indica a intensidade da autocorrelação serial de ordem 1 do ruído;
- σ_ε^2 é o parâmetro de determina a intensidade das flutuações aleatórias da série temporal.

Da maneira como apresentada anteriormente, as equações 2.7 e 2.9 devem ser pensadas como um sistema de equações, ainda que possa parecer que a equação 2.9 não faça parte do modelo diretamente. Porém, é possível combiná-las numa única expressão. Para isso, considere equação 2.7 escrita para o tempo $t - 1$

$$y_{t-1} = \beta_0 + \beta_1(t-1) + S(t-1) + e_{t-1} \,. \tag{2.10}$$

Após isolar a variável e_{t-1}, tem-se

$$e_{t-1} = y_{t-1} - \beta_0 - \beta_1(t-1) - S(t-1)\,. \tag{2.11}$$

Substituindo esse resultado na equação 2.9, obtém-se

$$e_t = \rho y_{t-1} - \rho\beta_0 - \rho\beta_1(t-1) - \rho S(t-1) + \varepsilon_t\,, \tag{2.12}$$

a qual pode ser substituída na equação 2.7, conduzindo a

$$y_t = \rho y_{t-1} + (1-\rho)\beta_0 + [t - \rho(t-1)]\beta_1 + S(t) - \rho S(t-1) + \varepsilon_t\,. \tag{2.13}$$

Essa última expressão é equivalente ao sistema das equações 2.7 e 2.9. Note que o processo autorregressivo no termo de tendência aleatória faz surgir um termo autorregressivo em y_t. Cabe observar ainda que essa dependência desaparece quando $\rho = 0$, como o esperado. Essa formulação é normalmente conhecida como método de Prais-Winsten e é usualmente utilizada em estudos epidemiológicos (GIL-ALANA, 2012; ANTUNES et al., 2015).

No estudo apresentado no Capítulo 3, os parâmetros desse modelo são estimados utilizando inferência Bayesiana, uma vez que esta gera bandas de credibilidade da sazonalidade de forma mais eficiente do que utilizando a inferência frequentista. Por isso, nas duas subseções seguintes, apresenta-se uma breve descrição de conceitos associados à inferência Bayesiana.

2.2 Inferência Bayesiana

Um dos problemas mais fundamentais em Estatística é o da inferência. Nesse contexto, fazer inferências pode ser entendido como o processo de se obter conclusões sobre quantidades não observadas por meio do conhecimento de dados numéricos associados ao problema ou sistema em questão. Essas quantidades não observadas podem representar, por exemplo, parâmetros de um modelo associado ao processo subjacente ao sistema que produz os dados ou possíveis observações futuras do sistema. No estudo de estatística, observa-se que essas questões podem encontrar respostas por meio de duas abordagens: a inferência clássica/frequentista e a inferência Bayesiana. Esse texto concentra-se na segunda abordagem que, entre outras vantagens, facilita a representação mais completa das incertezas associadas aos modelos e valores de parâmetros.

A inferência Bayesiana está apoiada no Teorema de Bayes, o qual é usado como uma espécie de regra para atualizar as probabilidades associadas a alguma hipótese frente ao conhecimento de informação. O Teorema de Bayes (também chamado de regra ou lei de Bayes) fornece uma descrição para encontrar a probabilidade de um evento por meio do conhecimento prévio de condições associadas à ocorrência desse evento. De modo matemático, pode-se escrever

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}\,, \tag{2.14}$$

na qual A e B são eventos probabilísticos e a probabilidade associada a B é diferente de zero ($P(B) \neq 0$). Na expressão 2.14, $P(A|B)$ representa a probabilidade condicional de ocorrência de A dado que B ocorre, $P(B|A)$ é a probabilidade de ocorrência de B dado que A ocorre, $P(A)$ é a probabilidade de ocorrência de A independentemente de B e $P(B)$ é a probabilidade de ocorrência de B independentemente de A. Essas duas últimas também são chamadas de distribuições marginais, já que $P(A) = \int P(A, B)dB$ e $P(B) = \int P(A, B)dA$, sendo $P(A, B)$ a distribuição conjunta de ocorrência dos evento A e B.

A ideia básica da inferência Bayesiana (CONGDON, 2014) consiste em substituir o evento B por um conjunto de n observações $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ e o evento A por um conjunto de parâmetros $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$. Desse modo, a expressão 2.14 se torna

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}\,. \tag{2.15}$$

Aqui, $p(\boldsymbol{\theta})$ é o conjunto das distribuições *a priori* dos parâmetros $\boldsymbol{\theta}$, $p(\boldsymbol{y}|\boldsymbol{\theta})$ é a probabilidade de observar os dados $\boldsymbol{y}$ sujeito ao modelo com parâmetros $\boldsymbol{\theta}$ (também chamada de função de verossimilhança) e $p(\boldsymbol{\theta}|\boldsymbol{y})$ representa as distribuições *a posteriori* dos parâmetros $\boldsymbol{\theta}$. Essa última é a quantidade de interesse na inferência Bayesiana e representa a incerteza associada aos parâmetros $\boldsymbol{\theta}$ após levar em conta sua distribuição *a priori* e dos dados $\boldsymbol{y}$. A quantidade $p(\boldsymbol{y})$ é chamada de distribuição marginal dos dados $\boldsymbol{y}$. Vale notar que essa última quantidade é a mesma para quaisquer valores do conjunto de parâmetros $\boldsymbol{\theta}$, de modo que podemos considerá-la como uma constante de proporcionalidade associada à normalização de $p(\boldsymbol{\theta}|\boldsymbol{y})$, ou seja, apenas as distribuições *a priori* dos parâmetros e função de verossimilhança são necessárias para se obter as distribuições *a posteriori* dos parâmetros. Nesse contexto é comum escrever

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,, \tag{2.16}$$

a qual pode ser ainda expressa de modo mais simplificado como

$$\text{Posterioris} \;\propto\; \text{Verossimilhança} \;\times\; \text{Prioris}\,. \tag{2.17}$$

Assim, em resumo, as principais componentes da inferência Bayesiana são:

- $p(\boldsymbol{\theta})$ – o conjunto das distribuições *a priori* dos parâmetros do modelo;
- $p(\boldsymbol{y}|\boldsymbol{\theta})$ – a função de verossimilhança, na qual os dados e os parâmetros são relacionados;
- $p(\boldsymbol{\theta}|\boldsymbol{y})$ – o conjunto das distribuições *a posteriori* que expressam as incertezas associadas aos parâmetros do modelo, após levar em conta os dados e as distribuições *a priori*.

A escolha das distribuições *a priori* é de grande importância para a inferência Bayesiana e, de fato, em algumas situações diferentes escolhas podem conduzir a resultados diferentes

para uma mesma função de verossimilhança. Além disso, as distribuições *a priori* devem apresentar algumas características como respeitar o espaço paramétrico, levar as distribuições *a posteriori* integráveis (ou próprias), refletir o conhecimento de especialistas acerca do fenômeno em estudo, entre outras. Normalmente, as distribuições *a priori* são divididas em duas categorias: informativas e não informativas. O primeiro caso corresponde a quando existe alguma informação sobre os valores do conjunto de parâmetros $\boldsymbol{\theta}$. Um exemplo típico desse caso ocorre quando o modelo já foi exposto a outros dados previamente, assim pode-se assumir que as distribuições *a priori* sejam iguais às distribuições *a posteriori* obtidas anteriormente. Algo similar também pode ser feito quanto o modelo é exposto a novos dados. No segundo caso referente as distribuições não informativas, tem-se uma informação vaga ou genérica sobre os parâmetros, como os seus intervalos de valores, seus sinais (parâmetro positivo ou negativo), ou ainda algum limite superior para seus valores. Normalmente, os parâmetros associados às distribuições *a priori* são chamados de hiper-parâmetros para não haver confusão com o conjunto dos parâmetros do modelo (isto é, os valors de $\boldsymbol{\theta}$).

Neste trabalho, a *priori*, tem-se a distribuição uniforme para ρ e a distribuição normal para β_0, β_1, η_0 e η_1. As estimativas dos parâmetros foram feitas utilizando o *software OpenBUGS* (LUNN et al., 2009), no qual a sigla *BUGS* refere-se a *Bayesian inference Using Gibbs Sampling* (em tradução livre: "inferência bayesiana utilizando amostragem de Gibbs"). Portanto, na próxima seção tem-se uma breve descrição do que vem a ser a técnica de amostragem de Gibbs. O código do modelo é dado abaixo:

```
model {
 for (t in 2:N) {
   y[t] ~ dnorm(mu[t],tau.delta)
   mu[t] <- b0*(1-rho)+b1*(t-rho*(t-1))+S[t]+rho*y[t-1]

   S[t] <- S1[t] + S2[t]
   S1[t] <- eta[1]*(sin(2*pi*t/12)-rho*sin(2*pi*(t-1)/12))
   S2[t] <- eta[2]*(cos(2*pi*t/12)-rho*cos(2*pi*(t-1)/12)) }

#tau.e <- 1/pow(sigma.e,2)
#sigma.e <- sigma.delta/sqrt(1-pow(rho,2))
sigma.delta ~dunif(0,100)
tau.delta <- 1/pow(sigma.delta,2)
rho ~ dunif(-1,1)
b0 ~ dnorm(0,0.0001)
b1 ~ dnorm(0,0.0001)
eta[1] ~ dnorm(0,0.0001)
eta[2] ~ dnorm(0,0.0001) }
```

2.3 Amostragem de Gibbs

Como deve estar claro, o objetivo da inferência Bayesiana é encontrar as distribuições *a posteriori* relacionadas a um conjunto de dados. Ainda que seja possível obter soluções analíticas/exatas para essas distribuições em alguns casos, para a maioria dos modelos, o cálculo das distribuições *a posteriori* envolve integrais que são intratáveis do ponto de vista analítico. Desse modo, o uso de métodos numéricos para construir amostras das distribuições *a posteriori* é praticamente inevitável na maioria dos problemas envolvendo inferência Bayesiana. Neste contexto, uma classe de algoritmos capazes de executar esta amostragem é denominada de *Markov Chain Monte Carlo*, frequentemente abreviada por MCMC. Exemplos de métodos MCMC incluem o algoritmo de Metropolis–Hastings, a *slice sampling* e a amostragem de Gibbs. Esse texto concentra-se em apresentar a amostragem de Gibbs, o qual figura entre os algoritmos mais populares para realizar inferência Bayesiana, principalmente por não depender de parâmetros ajustáveis e por produzir amostras a partir do conhecimento de distribuições condicionais (CASELLA; GEORGE, 1992).

Essa técnica será apresentada por meio de um exemplo. Para tal, considere uma distribuição de probabilidade $p(x, y)$ que seja complicada de se obter uma amostra. Suponha, entretanto, que seja possível obter amostras das distribuições condicionais $p(x|y)$ e $p(y|x)$. O processo de amostragem de Gibbs inicia a partir de dois valores (x_0, y_0) para construir um novo estado (x_1, y_0), no qual x_1 é amostrado de $p(x|y_0)$. Em seguida, o estado é modificado para (x_1, y_1), sendo y_1 amostrado de $p(y|x_1)$. Esse processo é repetido um número M de vezes, produzindo um sequência de variáveis aleatórias da forma

$$(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots, (x_M, y_M)\,, \tag{2.18}$$

cuja distribuição tende a se aproximar de $p(x, y)$. É ainda muito comum descartar as primeiras B observações (o chamado *burn-in period*) para garantir a estacionariedade das observações. Tipicamente o número de realizações M é muito grande, tornando a escolha de B não muito crucial .

2.4 Testes de adequação do modelo de séries temporais

A adequação do modelo para descrever os dados inclui uma análise residual, sendo que os resíduos são a diferença entre os valores observados para as séries e os valores correspondentes preditos pelo modelo. Esses resíduos devem compor uma série temporal estacionária, aleatória, com correlação não significativa entre seus valores sucessivos, ou seja, os resíduos devem se desenvolver no tempo ao redor de uma média próxima de zero e não devem ser

autocorrelacionados. Existem vários métodos para verificar a estacionariedade dos resíduos (como os testes de Dickey-Fuller Aumentado e Phillips-Perron) e também para verificar a autocorrelação dos resíduos (como o gráfico de resíduos, o teste de Durbin-Watson e a função de autocorrelação). Na aplicação apresentada no Capítulo 3, foram utilizados o teste de Dickey-Fuller aumentado e a análise dos gráficos da função de autocorrelação amostral parcial (facp), métodos que serão apresentados nas próximas subseções.

2.4.1 Teste de Dickey-Fuller aumentado

Como dito anteriormente, na verificação do pressuposto de estacionariedade de uma série temporal, pode-se construir testes de hipótese, tais como os testes de Dickey-Fuller aumentado e Phillips-Perron, dentre outros. Como não existe um teste que verifica a estacionariedade em seu sentido mais amplo, é comum construir tais testes a partir do princípio da presença da raiz unitária, que indicaria que o processo envolvido seria um passeio aleatório, tal como a Equação 2.2, o qual não é estacionário.

Considerando o processo autorregressivo de ordem p

$$e_t = \alpha_1 e_{t-1} + \alpha_2 e_{t-2} + \cdots + \alpha_p e_{t-p} + \varepsilon_t, \tag{2.19}$$

tem-se

$$e_t = \alpha_1 B e_t + \alpha_2 B^2 e_t + \cdots + \alpha_p B^p e_t + \varepsilon_t, \tag{2.20}$$

em que $B^j e_j = e_{t-j}$.

Dessa forma, para avaliar a estacionariedade analisa-se as raízes do polinômio característico

$$1 - \alpha_1 B + \alpha_2 B^2 + \cdots + \alpha_p B^p = 0. \tag{2.21}$$

Sabe-se que se $B = 1$ é raiz da equação 2.21, ou seja, se existe pelo menos uma raiz unitária, o processo não é estacionário (CRYER; CHAN, 2008). Considerando um processo autorregressivo de ordem 1, ou seja, $p = 1$ e $\alpha_1 = \rho$, tem-se

$$e_t = \rho e_{t-1} + \varepsilon_t, \tag{2.22}$$

de onde tem-se

$$(1 - \rho B) e_t = \varepsilon_t, \tag{2.23}$$

em que e_t é não estacionária se a raiz de

$$1 - \rho B = 0 \tag{2.24}$$

é

$$B = \frac{1}{\rho} = 1, \tag{2.25}$$

o que implica em $\rho = 1$. Logo, sob a hipótese nula de $\rho = 1$, tem-se $e_t = e_t - e_{t-1}$. Denotando $\delta = \rho - 1$, obtém-se

$$\begin{aligned} e_t - e_{t-1} &= (\rho - 1)e_{t-1} + \varepsilon_t \\ &= \delta e_{t-1} + \varepsilon_t \end{aligned}$$

Assim, tem-se a equivalência em testar a hipótese nula $\rho = 1$ ou $\delta = 0$, ou seja, tem-se as seguintes hipóteses

$$\begin{aligned} H_0 :&\ \ \delta = 0 \text{ (nesse caso, } e_t \text{ não é estacionário)} \\ H_1 :&\ \ \delta < 0 \text{ (nesse caso, } e_t \text{ é estacionário).} \end{aligned}$$

Este teste é conhecido como teste de Dickey-Fuller e pode ser estendido para considerar defasagens maiores do que 1, além de termos de tendência.

Este caso é denominado teste de Dickey-Fuller aumentado e pode ser expresso como a seguinte regressão

$$\Delta e_t = \beta_1 + \beta_2 t + \delta e_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta e_{t-i} + u_t, \tag{2.26}$$

em que β_1 é o intercepto, β_2 é o coeficiente de tendência, δ é o coeficiente de presença de raiz unitária, m é o número de defasagens tomadas na série e α_i são os coeficientes constantes.

O teste de Dickey-Fuller aumentado testa as mesmas hipóteses apresentadas anteriormente e está baseado na estatística T definida por

$$T = \frac{\hat{\delta}}{se(\hat{\delta})}, \tag{2.27}$$

em que $\hat{\delta}$ é estimador para δ e $se(\hat{\delta})$ é estimador para o desvio padrão de δ. O valor T é um número negativo e seus valores críticos foram tabelados por Dickey e Fuller (DICKEY; FULLER, 1979).

Na aplicação apresentada no Capítulo 3, o teste de Dickey-Fuller aumentado foi calculado utilizando a função `adf.test` do pacote *fpp* (HYNDMAN, 2013) no programa R.

2.4.2 Função de autocorrelação

Considere um conjunto qualquer T e a família $Z = \{Z(t), t \in T\}$. Se para cada t, Z é uma variável aleatória, então Z é um processo estocástico. Seja x_t uma série temporal estacionária de média zero na qual $t \in Z$. A função de autocovariância (facv) é definida por

$$\gamma_\tau = E\{x_t x_{t+\tau}\} \tag{2.28}$$

e satisfaz as seguintes propriedades (MORETTIN; TOLOI, 2006):

(i) $\gamma_0 > 0$,

(ii) $\gamma_{-\tau} = \gamma_\tau$,

(iii) $|\gamma_\tau| \leq \gamma_0$,

(iv) γ_τ é não negativa definida, ou seja, $\sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k \gamma_{\tau_j - \tau_k} \geq 0$, para todo número real a_i, $i = 1, 2, \cdots, n$ e $\tau_1, \tau_2, \cdots, \tau_k \in Z$.

A função de autocorrelação (fac) de x_t é definida por

$$\rho_\tau = \frac{\gamma_\tau}{\gamma_0}, \tag{2.29}$$

na qual $\tau \in Z$. Tem-se que $\rho_0 = 1$ e as propriedades (ii), (iii) e (iv) de γ_τ são válidas para ρ_τ (MORETTIN; TOLOI, 2006).

Agora, seja ϕ_{kj} o j-ésimo coeficiente de um modelo autoregressivo de ordem k da forma

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \cdots + \phi_{kk}\rho_{j-k}\,, \tag{2.30}$$

no qual $j = 1, 2, \cdots, k$. A partir da equação 2.30, obtem-se as equações de Yule-Walker (MORETTIN; TOLOI, 2006), que podem ser expressas na seguinte forma matricial

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_2 & \cdots & \rho_{k-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \phi_{k3} \\ \vdots \\ \phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_k \end{bmatrix}. \tag{2.31}$$

Utilizando a regra de Cramer em 2.31, obtém-se a função de autocorrelação parcial que é definida por

$$\phi_{kk} = \frac{|A_k^*|}{|A_k|}\,, \tag{2.32}$$

no qual A_k é a matriz de autocorrelações e A_k^* é uma matriz simétrica com diagonal principal unitária, semelhante à matriz A_k porém com a última coluna substituída pelo vetor de autocorrelações.

A Figura 1 mostra exemplos da função autocorrelação parcial utilizados para analisar os dados do Capítulo 3, todas geradas utilizando a função `pacf` (R Core Team, 2017) do programa R. Essas facp referem-se ao resíduos do ajuste do modelo definido pelas equações 2.7, 2.8 e 2.9, , quando ajustado aos dados após agrupá-los por sexo. Observando os gráficos da Figura 1, nota-se que em ambos os casos os resíduos não são autocorrelacionados, pois se as FACP encontram-se praticamente dentro dos intervalos de 95% de confiança, os quais são construídos assumindo que os resíduos sejam normalmente distribuídos. Observa-se também a presença de poucos pontos das facp fora dos intervalos de confiança, o que que compatível com o fato que a cada 40 *lags*, em média, dois possam estar fora do intervalo (MORETTIN; TOLOI, 2006).

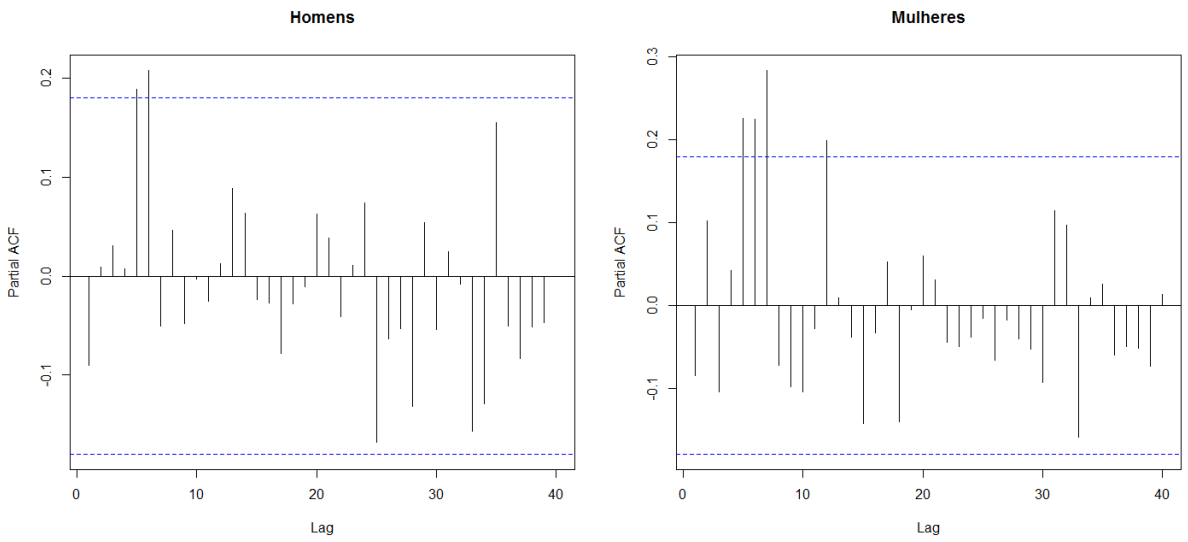

Figura 1 – Gráficos das funções de autocorrelação parcial dos resíduos obtidos após ajustar o modelo definido pelas equações 2.7, 2.8 e 2.9 aos dados descritos no Capítulo 3

2.5 Modelos aditivos generalizados

O modelo de regressão discutido na Seção 2.1, expresso nas equações 2.7, 2.8 e 2.9, e que produz os resultados do Capítulo 3, assume um termo de tendência linear e um termo de sazonalidade com período de 12 meses expresso como uma soma de funções trigonométricas. Embora essas hipóteses sejam plausíveis e razoáveis, em geral, hipóteses como essas não

podem ser diretamente observadas a partir nos dados e costumam refletir, em algum grau, a intuição do pesquisador sobre a série temporal em estudo. Uma possibilidade interessante é tentar extrair esse tipo de informação diretamente dos dados. Por exemplo, no lugar de considerar um termo de tendência linear poderia-se utilizar uma função polinomial do tempo de grau d, ou seja,

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_j t^j + \cdots + \beta_d t^d + e_t \,, \tag{2.33}$$

em que β_j representa j-ésimo coeficiente da expansão polinomial. Nesse caso, o ajuste desse tipo de modelo para um d suficientemente grande deve ser capaz de extrair possíveis padrões não lineares subjacentes à serie temporal. Esse tipo de abordagem é conhecida como regressão polinomial (JAMES et al., 2013) e os valores de β_j podem ser estimados utilizando mínimos quadrados, visto que é possível pensar esse modelo como uma regressão linear na qual as variáveis preditoras são t, t^2, ..., t^n. Na prática, porém, é muito incomum usar d muito maior que 3 ou 4 (JAMES et al., 2013) pois nesses casos o ajuste pode ser muito flexível e produzir formas funcionais estranhas, principalmente no início e no final da série temporal. Sendo assim, esse tipo de modelagem não é conveniente para descrever séries temporais que tenham algum tipo de sazonalidade.

Uma tentativa para contornar esse problema apresentado pela regressão polinomial é considerar a chamada regressão polinomial segmentada (JAMES et al., 2013) (ou *piecewise polynomial regression*). Nesse caso, ao invés de ajustar um polinômio de grau alto ao dados, ajusta-se vários polinômios de graus menores em diferentes partes da série temporal. Os pontos que definem as diferentes regiões da série são usualmente denominados por "nós" e correspondem aos pontos nos quais ocorrem as mudanças nos valores dos coeficientes dos polinômios. Por exemplo, uma regressão polinomial cúbica com um único nó localizado no ponto c pode ser expressa como

$$y_t = \begin{cases} \beta_{01} + \beta_{11} t + \beta_{21} t^2 + \beta_{31} t^3 + e_t & \text{se } t < c \\ \beta_{02} + \beta_{12} t + \beta_{22} t^2 + \beta_{32} t^3 + e_t & \text{se } t \geq c \end{cases} \,, \tag{2.34}$$

na qual β_{01}, β_{11}, β_{21} e β_{31} são os coeficientes da primeira função polinomial e β_{02}, β_{12}, β_{22} e β_{32} representam o mesmo para a segunda função polinomial. De modo geral, quanto maior o número de nós usados, maior a flexibilidade do ajuste; em particular, ao usar K nós ao longo da série temporal, ajusta-se $K + 1$ polinômios de grau d a ela. Nos pontos de nó a função ajustada ao dado é usualmente descontínua, um comportamento que não é natural para muitas séries temporais. Para remover essas descontinuidades, costuma-se adicionar vínculos entre as partes polinomiais de modo a impor a continuidade da função ajustada, bem como de suas derivadas, nos pontos de nó. Na prática, esses vínculos impõem que a função ajustada seja contínua e bastante suave. Além disso, os vínculos impostos reduzem o número de parâmetros a ser ajustado e a complexidade do modelo. Um exemplo muito utilizado desse procedimento é a chamada regressão *cubic spline* (JAMES et al., 2013). Nesse caso, ajusta-se um polinômio de

terceiro grau em $K+1$ partições da série, impondo que nos pontos de nó as partes polinomiais e suas derivadas de primeira e segunda ordem sejam contínuas. No caso geral, denomina-se de regressão *spline* de grau d o ajuste de $K+1$ polinômios de grau d, no qual as partes polinomiais são contínuas nos pontos de nó, assim como suas derivadas de ordem $d-1$ (JAMES et al., 2013).

Uma maneira mais adequada de expressar esse tipo de regressão *spline* é considerar que y_t é expresso como uma expansão utilizando uma base composta por funções apropriadamente escolhidas. Em particular, no caso da regressão *cubic spline* com K nós, pode-se escrever

$$y_t = \beta_0 + \beta_1 b_1(t) + \beta_2 b_2(t) + \cdots + \beta_j b_j(t) + \cdots + \beta_{K+3} b_{K+3}(t) + e_t \tag{2.35}$$

na qual a $b_1(t)$, $b_2(t)$, $\ldots$, $b_{K+3}(t)$ representam a base de funções. Uma das maneiras mais diretas para expressar essa base é considerar as funções $b_1(t) = t$, $b_2(t) = t^2$, $b_3(t) = t^3$ e

$$b_j(t) = \begin{cases} (t - c_i)^3 & \text{se } t > c_i \\ 0 & \text{caso contrário} \end{cases}, \tag{2.36}$$

para todo $j > 3$, em que c_i $(i = 1, 2, \ldots, K)$ representa o i-ésimo nó. Dessa maneira, considerando que os K nós estejam fixados, a regressão *cubic spline* pode ser vista como um modelo linear no qual as variáveis preditoras são os valores das funções $b_j(t)$ e os coeficientes lineares são os β_j $(\forall j = 1, 2, \ldots, K+3)$.

É possível observar a plausibilidade dessa representação verificando que o número de parâmetros numa regressão *cubic spline* com dois nós expressa por

$$y_t = \begin{cases} \beta_{01} + \beta_{11} t + \beta_{21} t^2 + \beta_{31} t^3 + e_t & \text{se } t < c_1 \\ \beta_{02} + \beta_{12} t + \beta_{22} t^2 + \beta_{32} t^3 + e_t & \text{se } c_1 \leq t \leq c_2 \\ \beta_{03} + \beta_{13} t + \beta_{23} t^2 + \beta_{33} t^3 + e_t & \text{se } t > c_2 \end{cases} \tag{2.37}$$

é o mesmo da representação

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 b_4(t) + \beta_5 b_5(t) + e_t\,. \tag{2.38}$$

Nessa última equação, o número de parâmetros é seis. Já na equação 2.37, tem-se doze coeficientes, porém, deve-se considerar ainda os vínculos associados às condições de continuidade dessa função e de suas derivadas de primeira e segunda ordens nos dois pontos de nó (c_1 e c_2). Essas condições geram 3 vínculos em cada ponto de nó, reduzindo o número de coeficientes independentes da equação 2.37 de doze para seis, ou seja, o mesmo número observado na representação da equação 2.38.

Seja na forma da equação 2.37 ou 2.38, um ponto importante desse tipo de regressão é a quantidade e a localização de nós. Naturalmente, quanto maior o número de nós, maior é a flexibilidade do modelo. Sendo assim, uma possibilidade é adicionar um número maior de

nós em regiões nas quais o valor de y_t varia mais rapidamente e um número menor nas regiões de baixa variabilidade de y_t. Embora seja uma ideia razoável, na prática costuma-se distribuir os K nós de maneira uniforme (JAMES et al., 2013) ao longo dos valores de t. Desse modo, a questão pode ser reduzida a determinar o número K de nós. Deve-se observar que um número muito grande de nós pode produzir uma função demasiadamente oscilante, enquanto um número pequeno pode conduzir a um ajuste muito suavizado que não descreve todas nuances do comportamento de y_t. Ainda que existam procedimentos voltados para encontrar um valor ótimo para o número de nós (JAMES et al., 2013), uma estratégia muito usada na literatura é utilizar um grande número de nós e introduzir uma maneira de penalizar o ajuste para evitar funções muito oscilantes.

Para descrever melhor esse procedimento, observa-se que o modelo da equação 2.35 pode ser reescrito como

$$y_t = g(t) + e_t \tag{2.39}$$

sendo

$$g(t) = \sum_{j=0}^{K+3} \beta_j b_j(t)\,, \tag{2.40}$$

com $b_0(t) = 1$. Desse modo, pode-se imaginar que o objetivo da regressão é encontrar uma função $g(t)$ que descreva bem os dados y_t no sentido de minimizar o valor da soma dos quadrados dos resíduos

$$\sum_{t=1}^{n} [y_t - g(t)]^2\,. \tag{2.41}$$

Porém, conforme discutido anteriormente, um grande número de nós pode conduzir a uma soma dos quadrados dos resíduos tão pequena quanto se queira caso não exista nenhum vínculo. Nesse contexto, uma possibilidade para evitar essa situação é buscar por uma função $g(t)$ que minimize a quantidade (JAMES et al., 2013)

$$\sum_{t=1}^{n} [y_t - g(t)]^2 + \lambda \int g''(t)^2 dt\,, \tag{2.42}$$

na qual λ é um parâmetro não negativo, $g''(t)$ representa a derivada segunda de $g(t)$ e a integral é realizada nos limites apropriados relacionados aos dados.

Assim, busca-se uma função $g(t)$ que ajuste bem o dado (ou seja, que leve a um valor pequeno para $\sum_{t=1}^{n} [y_t - g(t)]^2$), mas que também não seja demasiadamente oscilante. Essa última restrição está relacionada ao termo $\lambda \int g''(t)^2 dt$. Em particular, sabe-se que a derivada primeira $g'(t)$ mede a inclinação de $g(t)$ no ponto t e a derivada segunda $g''(t)$ fornece uma estimativa de como essa inclinação varia ao redor do ponto t. De modo simplificado, pode-se imaginar que $g''(t)$ mede a rugosidade de $g(t)$ no ponto t. Assim, a quantidade $\int g''(t)^2 dt$, que pode ser pensada como uma soma ao longo do intervalo de t, mede em certo sentido a rugosidade total de $g(t)$ no intervalo de tempo. Se $g(t)$ for bastante suave, $g'(t)$ deve ser bastante próxima de uma função constante e, consequentemente, $g''(t)$ vai assumir valores

muito pequenos e levar a um valor pequeno para $\int g''(t)^2 dt$. Por outro lado, se $g(t)$ for muito oscilante, $g'(t)$ deve variar muito também, conduzindo a valores grandes de $g''(t)$ e, consequentemente, um grande valor para a integral $\int g''(t)^2 dt$. Portando, grandes valores λ na equação 2.42 devem produzir uma função ajustada $g(t)$ muito suave, em particular, no limite de $\lambda \to \infty$ teremos uma função "perfeitamente suave", ou seja, um reta (regressão linear). Já no caso de pequenos valores de λ, o termo $\lambda \int g''(t)^2 dt$ tem pequeno efeito no ajuste e a função $g(t)$ pode se tornar muito oscilante.

Nesse ponto, mesmo com introdução do termo $\lambda \int g''(t)^2 dt$, pode parecer que a regressão *spline* apresenta muitos graus de liberdade para o ajuste. Ocorre que o parâmetro λ controla não apenas a rugosidade de $g(t)$ mas também uma quantidade chamada de número efetivo de graus de liberdade df_λ (JAMES et al., 2013). Em particular, para a regressão *cubic spline*, a medida de λ varia de zero a infinito, df_λ diminui de $K+4$ (ou seja, o número de parâmetros na equação 2.40) para dois (o número de parâmetros numa regressão linear simples). De modo geral, o número de graus de liberdade de um modelo costuma ser definido como o número de parâmetros livres que ele apresenta. Entretanto, no caso da regressão *spline* que busca minimizar a expressão 2.42 (muitas vezes chamada de *smoothing spline*), a variabilidade que pode ser produzida por seus parâmetros ($K+4$ no caso de *cubic smoothing splines*) é fortemente limitada pelo termo $\lambda \int g''(t)^2 dt$ (muitas vezes chamado de *penalty term*).

Para definir o número efetivo de graus de liberdade df_λ formalmente, suponha que a função de $g(t)$ que minimiza a expressão 2.42 para um dado λ possa ser escrita como (JAMES et al., 2013)

$$\mathbf{g}_\lambda = \mathbf{S}_\lambda \mathbf{y}\,, \tag{2.43}$$

na qual $\mathbf{g}_\lambda$ é um vetor coluna com os parâmetros de $g(t)$ (cujo número é assumido ser igual ao número de elementos na série temporal), $\mathbf{y}$ é um vetor coluna contendo toda observações de y_t e $\mathbf{S}_\lambda$ é uma matriz quadrada. Nesse contexto, define-se o número efetivo de graus de liberdade df_λ como o traço da matriz $\mathbf{S}_\lambda$ (ou seja, a soma dos elementos da diagonal de $\mathbf{S}_\lambda$). Embora essa definição seja bastante técnica, a interpretação da quantidade df_λ é relativamente intuitiva no sentido que pode ser imaginada como o número de parâmetros que fato estão livres para produzir variações na forma de $g(t)$ perante as restrições impostas pela equação 2.42. Assim, num caso hipotético de uma regressão *cubic smoothing spline* como 5 nós (portanto, 9 parâmetros), encontrar um valor de df_λ bem inferior a 9 é um indicativo de que os 5 nós foram suficientes para descrever os dados. Nesse mesmo caso, um valor de df_λ muito próximo de 9 indica que o 5 nós podem não ser suficientes para modelar adequadamente os dados.

Enquando o valor de df_λ fornece algum indicativo para escolher o número de nós usados na regressão, ele não contribui muito para encontrar um valor ótimo para o termo λ. Para resolver esse problema é comum empregar uma técnica denominada *leave-one-out cross-validation* (JAMES et al., 2013). Esse procedimento consiste em ajustar o modelo aos dados para um dado valor de λ excluindo o i-ésimo termo da série temporal e, em seguida,

calcular o chamado *cross-validation error* definido por

$$\text{RSS}_{cv}(\lambda) = \sum_{t=1}^{n} [y_t - g_\lambda^{\dagger i}(t)]^2 \,, \tag{2.44}$$

na qual $g_\lambda^{\dagger i}$ representa o modelo ajustado ao excluir o i-ésimo dado e n e número de termos da série (abordagem usualmente denomida por *leave-one-out cross-validation*). Dentro dessa abordagem, o valor ótimo de $\lambda = \lambda^*$ é definido como aquele que minimiza o *cross-validation error* RSS_{cv}.

Resumidamente, imagine que deseja-se modelar o comportamento de um série temporal y_t $(t = 1, \ldots, 100)$ utilizando uma regressão *cubic smoothing splines*, ou seja, encontrar uma $g(t) = \sum_{j=0}^{K+3} \beta_j b_j(t)$ em

$$y_t = g(t) + e_t \tag{2.45}$$

que se adeque bem aos dados. Para isso, posiciona-se, por exemplo, $K = 50$ nós ao longo dos valores de t e ajusta-se o modelo anterior aos dados procurando pelos coeficientes β_j que minimizem a quantidade definida na expressão 2.42 para um dado valor de λ. Faz-se uma varredura nos valores de λ (escolhendo, por exemplo, 1000 valores igualmente espaçados entre 0 e 100) para encontrar o $\lambda = \lambda^*$ que minimiza a quantidade RSS_{cv}. Finalmente, os coeficientes β_j ajustados aos dados que minimizam o RSS_{cv} são aqueles que compõem a expansão de $g(t)$ e melhor se adequam aos dados da série temporal. Nota-se, portanto, que a forma de $g(t)$ é completamente obtida a partir dos dados, supondo apenas que seu comportamento é contínuo.

Uma extensão natural da abordagem de regressão *cubic smoothing spline* é considerar o caso em que a variável a ser predita é função aditiva de mais de uma variável preditora. Nesse caso, pode-se imaginar que a série temporal y_t seja modelada não apenas como uma função do tempo, mas que possa depender de algum outro fator. Um exemplo pertinente para o estudo que será apresentado no capítulo 4, é imaginar que y_t representa o número de fraturas femorais por mês, e que essa quantidade é suposta ser função do ano, mês e número de feriados correspondes ao valor de y_t. Em uma notação mais simplificada, pode-se escrever

$$\text{fraturas} = \beta_0 + g_1(\text{ano}) + g_2(\text{mês}) + g_3(\text{feriados}) + e_t \,, \tag{2.46}$$

na qual β_0 é uma constante, $g_1(\text{ano})$ representa como ano afeta a número de fraturas, $g_2(\text{mês})$ indica o efeito do mês no número de fraturas, $g_3(\text{feriados})$ é o impacto do número de feriados sobre o número de fraturas e e_t representa o termo de erro. Nesse caso, cada uma das funções g_1, g_2 e g_3 pode ser expressa utilizando uma expansão como a mostrada na equação 2.40. O modelo descrito pela equação 2.46 é o que se costuma denominar de modelo aditivo generalizado (JAMES et al., 2013; WOOD, 2017) e pode ser pensado como um extensão natural dos modelos de regressão linear múltipla, na qual a dependência linear nas variáveis preditoras é substituída por relações não lineares modeladas por *smoothing splines*. Nesse caso, o nome aditivo se deve ao fato de que pode-se calcular o efeito de cada variável preditora separadamente e então adicioná-las todas para produzir o seu efeito conjunto. Essa característica

torna os modelos generalizados aditivos de fácil interpretação, tal qual numa regressão linear múltipla, porém com a vantagem serem capazes de modelar relações não lineares de forma não paramétrica.

Costuma-se dizer que esse tipo de modelagem pode "aprender" diversos padrões existentes nos dados. Por exemplo, no caso do modelo linear generalizado da equação 2.46, após ajustado, pode-se verificar uma possível componente sazonal no número de fraturas analisando o comportamento de função $g_2(\text{mês})$. Caso haja, por exemplo, uma incidência maior de fraturas nos meses de inverno do que nos de verão, a função $g_2(\text{mês})$ deve apresentar um máximo ao redor do valor 6 (correspondente ao mês de junho) e decrescer a media que número correspondente ao mês aumenta ou diminui. Similarmente, se a quantidade de feriados no mês impactar positivamente o número de fraturas, espera-se que a função $g_3(\text{feriados})$ seja crescente com número de feriados. Assim, a grande vantagem de um modelo como o 2.46 em comparação com aquele descrito nas equação 2.33 é ser não paramétrico no sentido em que, por exemplo, no lugar de termos uma função perfeitamente periódica como a expressa na equação 2.8, pode-se encontrar uma função $g_2(\text{mês})$ como alguma assimetria como relação ao meio do ano. O estudo apresentado no capítulo 4 foi feito utilizando o pacote *PyGAM* (SERVéN et al., 2018).

Para encerrar essa breve descrição dos modelos generalizados aditivos, cabe mencionar que existe uma relação entre esse tipo de modelagem e a inferência Bayesiana que discutimos anteriormente (LARSEN, 2015). Em particular, ocorre que esse tipo de modelo é equivalente a um modelo linear simples no qual os coeficientes têm uma distribuição normal *a priori*. Ainda que não apresente-se um demonstração desse resultado, essa conexão parece fazer sentido uma vez que ao ajustar um modelo generalizado aditivo impõe-se a crença *a priori* de que as funções das variáveis preditoras sejam funções suaves expressas por meio de uma expansão numa base conveniente. Conforme será visto no capítulo 4, modelos como o expresso na equação 2.46 podem ser muito úteis para compreender o comportamento no número de fraturas femorais e revelar algumas de suas características, como a existência de tendências sazonais distintas entre homens e mulheres, assim como uma dependência como a faixa etária das pessoas com esse tipo de condição médica.

Capítulo 3

Seasonal variation of femoral fractures in the State of São Paulo, Southeast Brazil

In this section, we present an article accepted in the Revista de Saúde Pública. This article was written aiming to present a practical application of the time series models discussed in the previous chapters and we have found interesting results. Revista de Saúde Pública is ranked in the B1 stratum of the Interdisciplinary Area by the Qualis CAPES. In 2017, Revista de Saúde Pública had an impact factor of 1.911 in the ISI Web of Science.

3.1 Introduction

Femoral fractures are a common and important cause of mortality and functional loss, mainly among elderly people (RAJU et al., 2017; KHARAZMI et al., 2016). Osteoporosis is an important risk factor for these fractures, which justifies its greater occurrence among women aged 60 years and over (DORE et al., 2013). Other known risk factors are falls, early menopause and sedentary lifestyle (RAJU et al., 2017; KHARAZMI et al., 2016; DORE et al., 2013; HAENTJENS et al., 2010). It is estimated that the life expectancy of patients suffering from this type of fracture is reduced by 15 to 20% (HAENTJENS et al., 2010). A review study showed that in Brazil, 5.5% of patients with fracture in the proximal third of the femur die during the hospitalization and 4.7% die at the end of one month of follow-up (SAKAKI et al., 2004). The cost per femoral fracture surgery is significant to the Brazilian health system, since that the direct cost of hospitalization is estimated between 4,100 and 70,000 dollars. Furthermore, the hospital medical resources used in the postoperative period accounted for 5.7% of the direct cost of hospitalization (ARNDT; TELLES; KOWALSKI, 2011).

A number of authors has studied the effects of seasonality on femoral fracture incidence (CRAWFORD; PARKER, 2003; JENA; MUDDU; RICHARDSON, 2003; CHIU; NG; CHOW, 1996; JACOBSEN et al., 1991; JACOBSEN et al., 1995; LAU et al., 1995; PARKER; MARTIN, 1994; PEDRAZZONI et al., 1993; RALIS, 1981). In a prospective study including 3,034 consecutive hip fracture patients in the United Kingdom, more fractures occurred during the winter compared to summer, and a tendency to a higher mortality for those patients admitted in the winter months (CRAWFORD; PARKER, 2003). However, this study did not showed significant difference in patient characteristics between the winter and summer seasons (CRAWFORD; PARKER, 2003). Other study (JENA; MUDDU; RICHARDSON, 2003), considering the monthly number of operated proximal femoral fractures across 31 hospitals of North-west England and Scotland, described an increase in the incidence of hip fractures in the months of December and January (winter months in the northern hemisphere). Douglas et al. (2000) studied the seasonal variation of hip fracture admissions at three different latitudes: Scotland, Shatin (Hong Kong) and Auckland (New Zealand). These authors also concluded that there is a higher incidence of fractures in the coldest months.

Considering the fact that most of the articles on the effects of seasonality on fracture incidence includes populations from countries of the Northern Hemisphere, the objective of the present article is to describe the seasonal variation of femoral fractures in the State of São Paulo, Southeast Brazil.

3.2 Methods

This is an ecological study, conducted over a period from January 2008 to December 2017. A consecutive series of $216,348$ reports of hospital admissions caused by femoral fractures was considered. This number corresponds to approximately one quarter of the admissions reported throughout the country in this period. The State of São Paulo is the most industrialized in Brazil, with a population of approximately 45 million inhabitants in an area of 248,209 square kilometers (based on official statistics of the Brazilian Institute of Geography and Statistics, available in <https://cidades.ibge.gov.br/brasil/sp/panorama>). The territory is between the latitudes 19°46'S and 25°18'S. Data on the fracture incidence were obtained from the Hospital Information System (SIH), available from the DATASUS (Database of the Brazilian Unified Health System). The SIH includes all hospitalizations within the Unified Health System (SUS). Yearly population projections based on census data were obtained by the TABNET system, also available from the DATASUS (<http://www2.datasus.gov.br/DATASUS/index.php>).

This study was conducted exclusively with secondary data, of public access, without identification of subjects, and its procedures are in accordance with the principles of ethics in research involving human beings. Thus, the study was waived from formal review and informed

consent by the institutional Research Ethical Committee. However, the study was conducted according to the Helsinki declaration and good clinical practice.

Let y_t be the number of cases of femoral fractures recorded at the month t in the State of São Paulo, divided by the number of days of the corresponding month. Thus, y_t refers to the monthly average number of events of femoral fractures per day. Considering the period from January 2008 to December 2017, we have $t = 1, 2, \ldots, 120$. The statistical data analysis was based on the time series model given by the expression

$$y_t = \beta_0 + \beta_1 t + S(t) + e_t \,, \tag{3.1}$$

where β_0 is an intercept, β_1 describes the trend component of the time series data, $S(t)$ is a function related to the seasonality of the time series and the terms e_t follow an autoregressive structure of order 1, given by $e_t = \rho e_{t-1} + \varepsilon_t$. We assumed that the terms follow a normal distribution with mean zero and variance σ^2_ε. The seasonality of the time series in given by

$$S(t) = \eta_1 \sin\left(\frac{2\pi t}{12}\right) + \eta_2 \cos\left(\frac{2\pi t}{12}\right) \tag{3.2}$$

Thus, the significance of the terms η_1 and η_2 determine the significance of the seasonal pattern of the time series and its amplitude can be obtained from $\sqrt{\eta_1^2 + \eta_2^2}$. This model formulation is based on the well-known Prais-Winsten method, commonly used in epidemiological studies (GIL-ALANA, 2012; ANTUNES et al., 2015). Estimates of the parameters of interest were obtained by Bayesian methods using the Gibbs sampler. Non-informative prior distributions were used, and posterior distributions of the parameters were constructed from 500,000 simulations after a burn-in period of 10,000 simulations (initial iterations to allow simulations to stabilize).

The adequacy of the model to describe the data includes a residual analysis, where the residuals (differences between the observed values for y_t and the correspondent values predicted by the model) should be a stationary time series, with non-significant serial correlation between their successive values. These assumptions were verified by graphs of the partial autocorrelation function (not showed in this article) and using the augmented Dickey-Fuller test. All statistical analysis was performed with OpenBUGS software (freely available at <http://openbugs.net>) and the R language version 3.4.3 (freely available at <https://www.r-project.org/>) was used to generate graphs and tables (CRAWLEY, 2005).

3.3 Results

The incidence of femoral fractures are shown according to age groups and sex in Table 1, considering the years of 2008, 2011, 2014 and 2017, with the rates per 100,000 population. These rates were obtained directly from the data and without the use of the

proposed regression model. We can observe that the fracture incidence is higher among men than women, considering the age groups up to 60 years. However, after the age of 70 years, the incidence is much higher among women than among men. For each age group, there are no important differences between the incidences when compared among all four years described in Table 1.

Table 1 – Incidence of femoral fractures in man and women per 100,000 population per year, State of São Paulo, Brazil, years of 2008, 2011, 2014 and 2017.

Age groups (years)	**Men**				**Women**			
	2008	**2011**	**2014**	**2017**	**2008**	**2011**	**2014**	**2017**
0-4	19.5	19.1	21.8	19.8	10.9	9.2	10.3	9.4
5-9	14.8	12.4	13.9	11.6	7.7	8.0	6.9	6.9
10-14	22.0	22.0	18.9	17.5	8.9	9.2	6.3	7.2
15-19	45.7	54.2	53.5	46.8	12.9	12.0	12.1	12.1
20-29	56.9	59.7	57.4	54.4	9.2	11.9	10.3	11.0
30-39	32.5	39.0	37.6	34.0	6.3	6.5	6.2	8.1
40-49	36.8	36.9	36.0	37.5	8.9	9.5	7.5	8.7
50-59	48.0	49.8	50.8	53.6	23.2	25.9	21.0	23.3
60-69	90.0	86.2	74.9	86.8	83.7	81.3	72.1	82.1
70-79	200.0	202.8	179.9	188.3	306.0	306.5	286.8	300.8
80 or more	594.4	550.7	496.2	553.1	1,007.5	981.5	901.0	941.7
Total	**50.0**	**53.2**	**52.0**	**54.3**	**43.0**	**46.8**	**45.9**	**53.4**

Panels (a) and (b) of the Figure 2 compare the observed time series for the monthly average number of events of femoral fractures per day and the values estimated from the proposed Bayesian model. The model was separately fitted to data for both sexes. We can observe that the predicted values are closed to the observed values, suggesting a reasonable fit of the model to the data in both cases. We also fitted alternative models for the data considering that the terms e_t follow an autoregressive structure of order 2, but we did not observed an expressive improvement of the fit compared with the previous models (comparisons between models was based on the deviance information criterion).

Bayesian estimates for the model parameters are showed in Table 2, with their correspondent 95% credible intervals. The 95% credible intervals are the Bayesian equivalent of standard confidence intervals, representing the central 95% of the range of values that are credible for the parameter estimated. While the values for the parameter β_0 are interpreted as hypothetical means when $t = 0$ (a "starting point" for the series), the values for β_1 are positives for both sexes, indicating an increasing trend over time. This increase is probably a result of the process of aging of the population of the State of São Paulo during the studied period. As intuitive visual inclinations in Figure 1, the estimated increasing trend for women is a bit stronger than for men. The 95% credible intervals for η_1 and η_2 do not include zero, indicating a significant seasonal effect for both time series. Credible intervals for the autore-

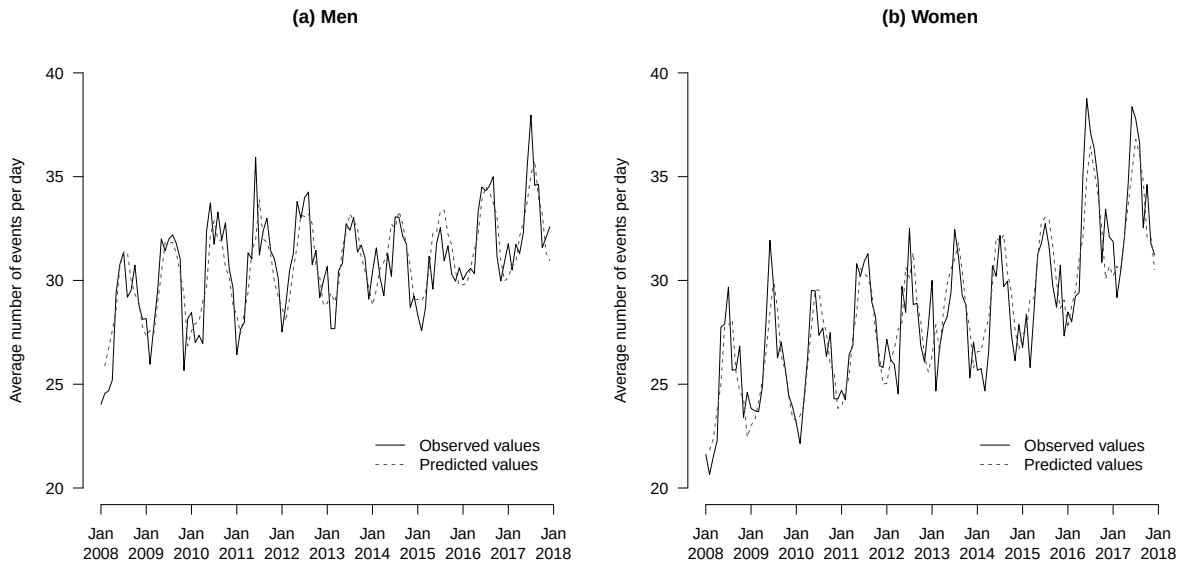

Figure 2 – Observed and predicted time series describing the monthly average number of events of femoral fractures per day, from January 2008 to December 2017, in the State of São Paulo, Brazil.

gressive coefficient ρ also do not include zero, indicating a significant and positive correlation between successive values of e_t in the time series.

Table 2 – Bayesian estimates for the model parameters including the monthly average number of events of femoral fractures per day, period from January 2008 to December 2017, State of São Paulo, Brazil.

Parameter	**Men**		**Women**	
	Estimate	**95% credible interval**	**Estimate**	**95% credible interval**
β_0	28.98	(28.16 , 29.80)	24.07	(22.96 , 25.16)
β_1	0.029	(0.017 , 0.041)	0.073	(0.057 , 0.088)
η_1	-1.540	(-2.026 , -1.036)	-1.502	(-2.127 , -0.881)
η_2	-1.655	(-2.145 , -1.164)	-2.463	(-3.084 , -1.853)
ρ	0.356	(0.175 , 0.536)	0.415	(0.241 , 0.593)

Panels (a) and (b) of Figure 3 shows graphs for the seasonality function $S(t)$ of the time series, according to the months of the occurrence of the femoral fractures. The graphs evidence that more cases of fractures were reported between June and August, the coldest months in the Southern Hemisphere. The gray areas in the graphs correspond to Bayesian credible regions, which is the interval in which we are 95% sure the true curves lies.

Figure 4 shows plots of the seasonal effects given by the function $S(t)$ stratified by age groups and considering the male population. In each graph, when the horizontal line crossing the zero is whole within the Bayesian credible region, we also have an indicative that the seasonal effect is not significant. Note that it was not possible to standardize the y-axes scales of these plots, which would impair the visualization of the behavior of the curves. Therefore,

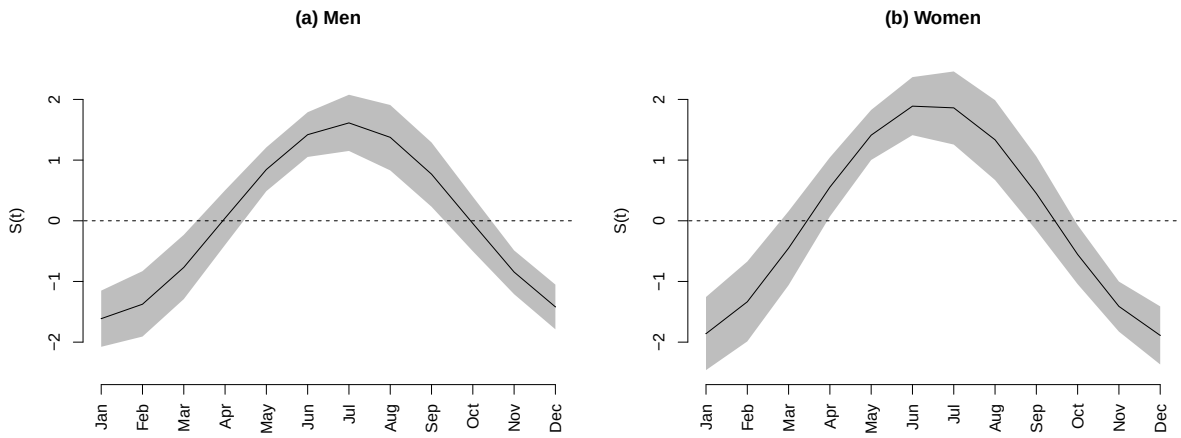

Figure 3 – Graphs for the seasonality function $S(t)$ of the time series, according to the months of the occurrence of the femoral fractures. The gray areas correspond to 95% credible regions.

we have not significant seasonal effects for the age groups 0-14 and 15-19 years. Unlike other age groups, where more cases of fractures were reported between June and August, it was observed a higher incidence between August and September among men aged 40-49 and 50-59 years.

Figure 5 shows plots of the seasonal effects given by the function $S(t)$ stratified by age groups and considering the female population. We observe significant seasonal effects only for older women, aged 60 years or more. In the age group 60-69, more cases of fractures were reported between July and September, while it was observed a higher incidence between May and August among women aged 70-79 and 80 years or more.

3.4 Discussion

Results in Table 1 show that the incidence of femoral fractures among men increased steadily with age, up to the age group of 20-29 years (54.4 cases per 100.000 in 2017). It is possible that men of this age group are more subject to risks due to work activities. The incidence among men decreased in the age group of 30-39 years and again increased steadily with age, up to the age group of 80 years or more. Incidence of femoral fractures was lower in women than in men before the age group of 60-69 years. In the age group of 60-69 years the incidence is similar between men and women, and after the age of 70 years, the incidence is much higher among women than among men. The number of fractures that are due to osteoporosis can explain these differences.

In the time series analysis using the Bayesian model, results in Table 2 and Figure 3 shows an important seasonal variation in the incidence of femoral fractures with a higher incidence during the coldest months compared to the summer period. This supports findings

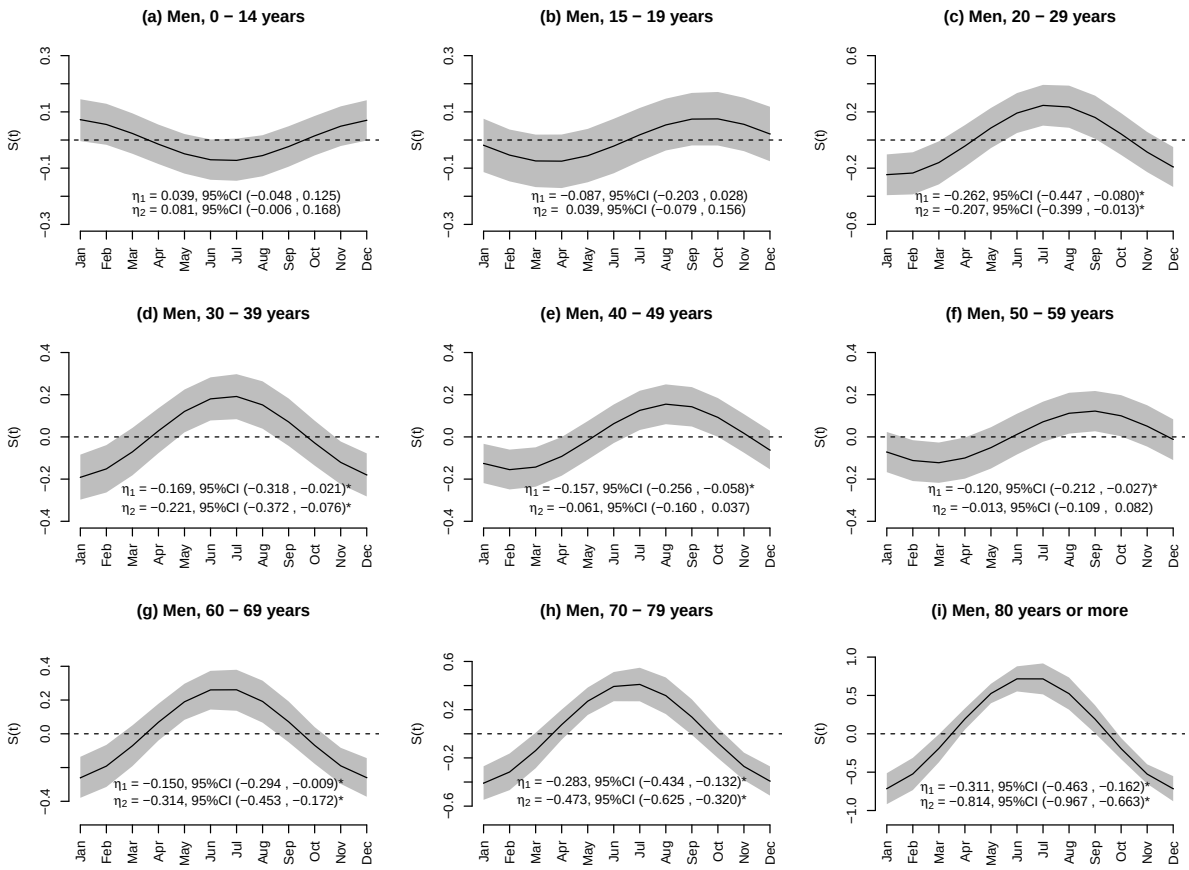

Figure 4 – Graphs for the seasonality function $S(t)$ of the time series, according to the months of the occurrence of the femoral fractures, stratified by age groups and considering the male population. The gray areas correspond to 95% credible regions.

from several other studies (CRAWFORD; PARKER, 2003; JENA; MUDDU; RICHARDSON, 2003; CHIU; NG; CHOW, 1996; JACOBSEN et al., 1991; JACOBSEN et al., 1995; LAU et al., 1995; PARKER; MARTIN, 1994; PEDRAZZONI et al., 1993; RALIS, 1981) although reasons for this variation remain uncertain. In the literature, there is no clear explanation for the increased number of fractures in winter. Among possible hypothesis, Douglas et al. (2000) remember that winters are darker as well as colder, and this may add to the number of falls in this season. While Ralis (1981) attributed the fractures to slipping on ice and snow, other authors describe that many cases in the winter time occur in indoor environments. Bastow, Rawlings e Allison (1983) proposed that undernutrition of elderly people is a risk factor for hypothermia in cold weather, and hypothermia could lead to impairment of judgement and co-ordination, and consequently, to injury. Chiu, Ng e Chow (1996) postulated that the hindrance of free movements and increased clumsiness of the patients with many layers of clothes put on during the cold weather might be associated to an increase of the incidence of fractures. The study from Mazzucchelli et al. (2018) showed a short-term association between several indicators of air pollution and daily hospital admissions due to osteoporotic fracture, which can also in part explain the seasonal variations in the numbers of femur fractures. Other articles

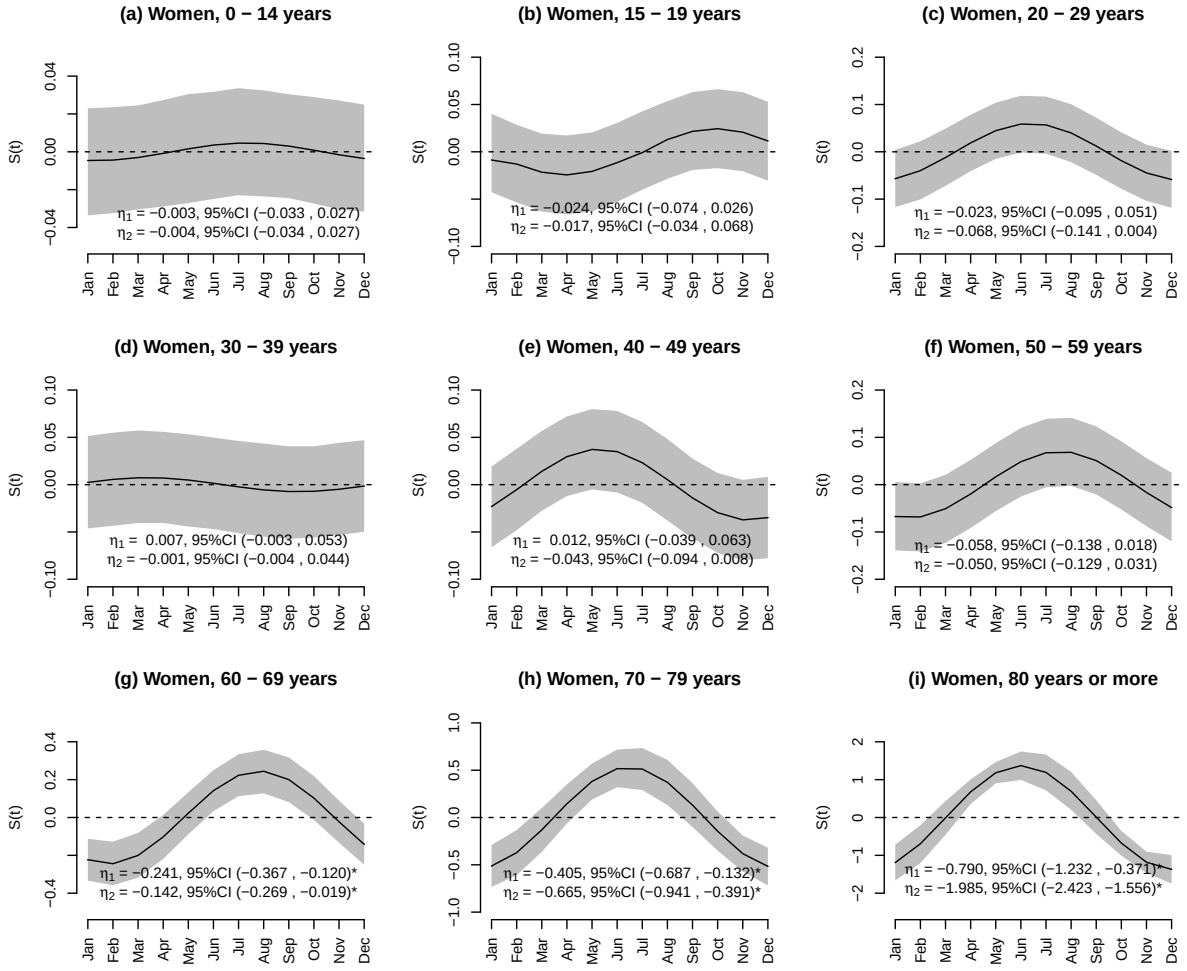

Figure 5 – Graphs for the seasonality function $S(t)$ of the time series, according to the months of the occurrence of the femoral fractures, stratified by age groups and considering the female population. The gray areas correspond to 95% credible regions.

support the role of the low exposure to sunshine in the vitamin D deficiency (ANTUNES et al., 2015). In this model, the low serum concentration of 25-hydroxyvitamin D leads to a low 1,25-dihydroxyvitamin D concentration and then to a higher serum parathyroid hormone concentration, principally in the winter season (ANTUNES et al., 2015). On the other hand, a low level of parathyroid activity can be associated with high bone turnover, leading to cortical bone loss and low density bone. Finally, this effect may lead to fracture (ANTUNES et al., 2015).

Results in Figures 4 and 5 demonstrate the strong effect of sex and age groups on the seasonal patterns of the time series. Figure 4 shows that among younger men (aged less than 20 years) there is not a clear seasonal effect, but among the other age groups seems to have a higher number of cases of femoral fractures in the coldest months of the year. On the other hand, results in Figure 5 show that women sustaining femoral fractures in the coldest months were older (aged more than 60 years) and probably less active. Again, it is unclear what mechanism may exist to elicit this phenomenon, certainly, that many of the probable reasons

for femoral fracture, mentioned in the previous paragraph, may also act in this scenario. So, future research conducted in specific patient populations may include more variables that best describe the cases, such as the environment in which the fracture occurred (outdoor or indoor), the form of the accident (falls, traffic accident, sports accidents, stumbling on an object, or other), the severity of the trauma, among others.

If we observe the results in Table 1, Table 2, and Figure 2 (Panels (a) and (b)) for man and woman, the estimated average of fractures per day at the beginning of the analysis (January 2008) was almost 29 (95% credible interval from 28.16 to 29.80) and about 24 (95% credible interval from 22.96 to 25.16). After that, a stronger trend for woman (0.073 which means an increase of about 1 case each 14 days) than for man (0.029 meaning an increase of about 1 case each 35 days) has conducted this group to reach even high rates of fractures per day in the last two years. Furthermore, the contribution of the seasonal term, expressed by its amplitude, is slightly stronger for woman (2.88) than for man (2.26), although, on average, that is less evident, as showed in Figure 5.

A potential limitation of the present study is the underreporting of the events and the possibility of miscoding of femoral fractures. The Hospital Information System (SIH) has incomplete coverage and there are uncertainties about the reliability of its data (BITTENCOURT; CAMACHO; LEAL, 2006). It covers only records obtained from the hospitals affiliated to the Brazilian Unified Health System (SUS), excluding the admissions from the Private Health System (PHS) and the Supplementary Health System (SHS), which are administrated by private healthcare insurance companies or health cooperatives. Moreover, due to its ecological design, we cannot to draw causal inferences about the effects of the different seasons of the year on the fractures incidence. However, despite its descriptive nature, findings of our study are important to a better understanding of how fracture occurrences are distributed over time. This information is useful for the future planning of health care.

3.5 Conclusion

In the State of São Paulo, Brazil, more cases of fractures occurred in the coldest months of the year, but men and women have different patterns of incidence according to the age groups.

Capítulo 4

Identifying patterns in the incidence of femoral fractures across gender and age in Brazil

The results discussed in this article are based on modeling the same dataset studied in the previous chapter, but now we rely on generalized additive models. Because of this methodological difference, the results reported in this chapter allows for the proper identification of non-linearities in long-term trends of the incidence of femoral fractures, and also make possible to model the seasonality of these time series in terms of more general functions. Thus, the results presented here are somehow complementary to those previously discussed.

4.1 Introduction

Revealing and understanding patterns about the incidence of different bone fractures is very important for planning and implementing preventive strategies and also for better allocating medical resources (COOPER et al., 2011). In more severe conditions such as hip or femur fractures, the aggregated costs go far beyond hospitalization expenses and are particularly impacted by often long recovery times faced by people suffering from these problems. These serious fractures have become more common every year mainly because of the increase in life expectancy and the consequent growth of elderly population around the world (COOPER et al., 2011). COOPER; CAMPION; MELTON have estimated that the number of hip fractures worldwide will jump from 1.6 million cases reported in 2000 to over 6.3 million cases in 2050. The same authors estimate that the annual costs of osteoporotic fractures exceed US $20 billion only in the United States and as much as € 30 billion in the European Union (COOPER; CAMPION; MELTON, 1992). Underlying these numbers, osteoporosis and low bone mineral density represent important risk factors (CLARK, 2002), and the former is considered a ma-

jor public health concern because of its association with age-related fractures, particularly hip and proximal femur fractures (EVANS; SEAGROATT; GOLDACRE, 1997; CUMMINGS; MELTON, 2002; CLARK, 2002; HAENTJENS et al., 2010).

Due to the impact these fractures have on people's life and their life quality, several works have been focused on identifying, modeling, and trying to better understand possible underlying patterns about the incidence of severe fractures in the population of different places across the globe (CRAWFORD; PARKER, 2003; JENA; MUDDU; RICHARDSON, 2003; CHIU; NG; CHOW, 1996; JACOBSEN et al., 1991; JACOBSEN et al., 1995; LAU et al., 1995; PARKER; MARTIN, 1994; PEDRAZZONI et al., 1993; RALIS, 1981; EVANS; SEAGROATT; GOLDACRE, 1997; CUMMINGS; MELTON, 2002; COOPER; CAMPION; MELTON, 1992). Among these studies, there is a consensus about the existence of a seasonal component in time series related to the incidence of femoral and hip fractures, in which not only more injuries are reported during winter time, but there is also a higher mortality rate of patients admitted during this time of the year (CRAWFORD; PARKER, 2003). The majority of these works are however based on data from developed countries in the Northern Hemisphere, and much less is known about how these findings generalize to developing nations located below the equator such as Brazil. Understanding whether the universality of these findings holds true or not is essential not only from an academic viewpoint but may also help to inform policy-makers and hospital managers about the particularities of the places where they live and work, improving the quality of their decisions on resource management.

Here we contribute to filling this gap by presenting an investigation on the incidence of femoral fractures in São Paulo, a Brazilian state located in Southeast Region of the country between the latitudes 19°46'S and 25°18'S. Our dataset comprises over 200 thousand reports grouped by gender and age of hospital admissions, spanning the period between January 2008 and December 2017. We show that time series of monthly rates of femoral fractures are well-described by generalized additive models in which the predictors are smooth functions of a variable indexing time and another variable representing the months of the year. By analyzing the profile of the adjusted smooth functions, we find that both men and women fractures are more likely to occur between June and September, a period that coincides with the winter season in the Southern Hemisphere. However, the effect of the month on the fracture incidence peaks around mid-June for women, while the same happens a bit later (in early July) and less intensly for men. Our results also reveal a growth over time of the incidence of femoral fractures both for men and women but at a more accerated rate for women after mid-2015. We further investigate these time series after grouping the reported cases by age. This analysis shows that the monthly seasonal structure is more intense in the oldest age groups, particularly for senior women. We also observe that while men and women fractures display a similar yearly increasing trend for the age group 60-69 yo, in older groups this increasing trend is much more intense for women than men. Finally, we have found that the number of holidays in a month

impacts on the incidence of femoral fractures among young men, but has no effect on women of the same age and on both genders of other age groups.

The remainder of this work is as follows. The next section presents our data and describes the models used for investigating the fractures time series. Section 3 depicts our findings and discussions. Section 4 summarizes our results and concludes the article.

4.2 Methods

4.2.1 Data

We have collected data about the incidence of femoral fractures freely available from DATASUS (Database of the Brazilian Unified Health System - <http://www2.datasus.gov.br/DATASUS/index.php>). This dataset comprises 216,348 reports of hospital admissions caused by femoral fractures occurring between January 2008 and December 2017 in the State of São Paulo (Brazil). Our analyses are based on time series of the monthly average number of reported cases per day grouped by gender (120 observations). These time series are further investigated by considering 9 age groups (0-14 yo, 15-19 yo, 20-29 yo, 30-39 yo, 40-49 yo, 50-59 yo, 60-69 yo, 70-79 yo, and over 80 yo). Figure 6 illustrates the time behavior of the femoral fractures for men and women of all ages. Since our study is based on freely available data, it does not require ethical approval of our institutional Research Ethics Committee.

4.2.2 Generalized additive models

Let us denote by y_i the monthly average number of femoral fractures reported per day for the i-th observation ($i = 1, 2, \ldots, 120$). We further consider the variables *year*$_i \in$ $(2008, 2018)$

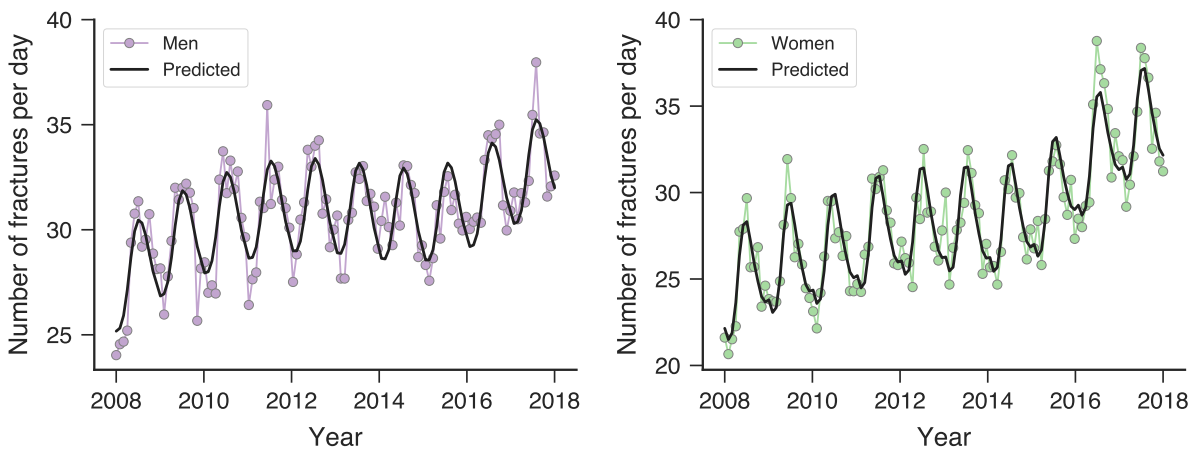

Figure 6 – Monthly average number of femoral fractures per day. The left panel shows the values for cases involving men while the right panel depicts those occurring with women. In both panels, the circles represent the observed values, and the continuous lines are the values predicted by the generalized additive models.

to be the corresponding time and $month_i \in (1, 2, \ldots, 12)$ as the corresponding month associated with the i-th observation. Within the framework of the generalized additive models (WOOD, 2017; JAMES et al., 2013), we model the response variable y_i as smooth functions of the $year_i$ and $month_i$, that is,

$$y_i = \beta_0 + g_1(month_i) + g_2(year_i) + e_i \,, \tag{4.1}$$

where β_0 is the intercept, $g_1(\ldots)$ is a non-parametric smooth function of $month_i$, $g_2(\ldots)$ is a non-parametric smooth function of $year_i$, and e_i is the error term.

The model of Eq. 4.1 is particularly useful for studying these time series because it allows the proper evaluation of possible non-linear effects that the predictors $month_i$ and $year_i$ may have on the response variable y_i. It is also worth noticing that no particular functional forms are assumed for $g_1(\ldots)$ and $g_2(\ldots)$ and that their profiles are directly inferred from data. In all our results, each smooth function is built using 20 cubic splines. In order to avoid overfitting, the model parameters are estimated by minimizing residual sum of squares plus a penalty term proportional to the integrated square of the second derivatives of the smooth functions, where λ is the proportionality constant. The higher the λ, the more the wiggliness of the interpolating functions are penalized, reducing the effective degrees of freedom of the model. The optimal value for the regularization parameter λ is obtained by minimizing the cross-validation error in a leave-one-out cross-validation strategy. This process is illustrated in Figure 7 when fitting the model of Eq. 4.1 to data shown in Figure 6. We notice that the cross-validation error displays a minimum value for $\lambda^* = 17.58$ for the men time series, while for the women data this minimum occurs for $\lambda^* = 3.91$. A similar situation is found when fitting all the other models we shall discuss over this manuscript.

Figure 6 depicts the adjusted models (continuous lines) for men and women data, where a good agreement is found between observed and predicted values in both cases. To further verify the adequacy of the model of Eq. 4.1 in describing our time series, we have estimated the partial autocorrelation functions of the residuals, as shown in Figure 8. We notice that no significant serial correlation is observed in both cases. We have further tested the normality of the residuals with the Kolmogorov-Smirnov test, and the null hypothesis cannot be rejected in both cases (p-values are 0.26 and 0.14 respectively for men and women data). A quite similar situation is observed for all models we have used in this manuscript.

4.3 Results and Discussion

We start by investigating the profiles of the estimated smoothing functions [$g_1(month_i)$ and $g_2(year_i)$] for men and women considering data from all ages. The left plot of Figure 9 shows the smoothing term accounting for the effect of the month [$g_1(month_i)$] on the incidence of femoral fractures for men (purple curve) and women (green curve). We first observe that the

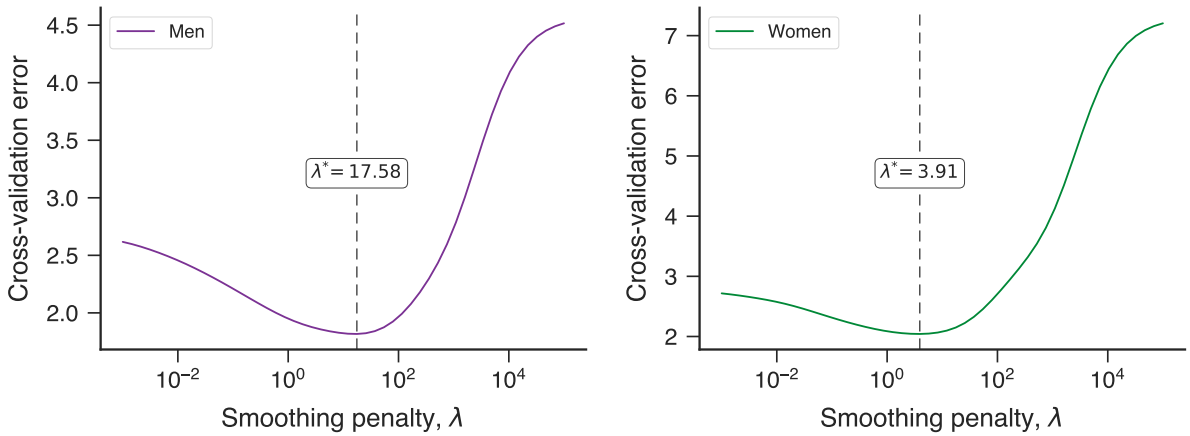

Figure 7 – Cross-validation error as a function of the smoothing penalty λ. The continuous lines show the cross-validation error estimated in a leave-one-out cross-validation strategy to determine the optimal value for the smoothing penalty λ. The optimal $\lambda = \lambda^*$ is the one minimizing the cross-validation error, as indicated in the plots. The left panel shows the results obtained for the time series associated with men femoral fractures, while the right one shows the results for the women data.

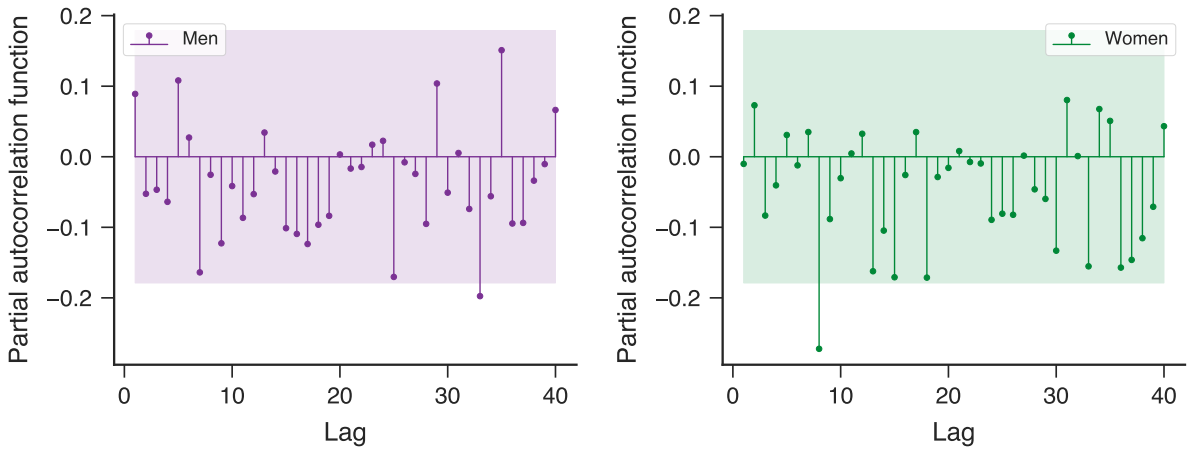

Figure 8 – Partial autocorrelation functions of the residuals of the adjusted models. The vertical lines show the correlation values and shaded regions represent the 95% confidence interval for a completely random process. We notice that practically all data points are within the random confidence band, indicating that the residuals are uncorrelated. The left panel displays the results for men data, and the right panel shows the same for women data.

effect of the month is significant for describing the fracture incidence of both genders. We also notice that the effect of the month is greater between June and September, achieving a peak in mid-June for women and early July for men. This result is in line with those reported for places in the Northern Hemisphere, in which a increase in the incidence of femoral fractures is found during the coldest months of the year (CRAWFORD; PARKER, 2003; JENA; MUDDU; RICHARDSON, 2003; CHIU; NG; CHOW, 1996; JACOBSEN et al., 1991; JACOBSEN et al., 1995; LAU et al., 1995; PARKER; MARTIN, 1994; PEDRAZZONI et al., 1993; RALIS,

1981). It also agrees with previous results based on the same dataset but obtained through a different approach (SOUZA et al., 2019). However, different from the results reported by SOUZA et al., where a parametric periodic form (in term of sines and cosines with a period of 12 months) was explicitly assumed as the seasonal term, we remark that ours are obtained in a non-parametric way. The model of Eq. 4.1 allows a much more flexible estimation of the seasonality of the time series, which in turn reveals features that would remain hidden in the approach of SOUZA et al.. That is the case of the asymmetric profile observed in the left panel of Figure 9 and also the more intense effect of the month detected for women, features that do not appear when using the approach described in SOUZA et al. (2019).

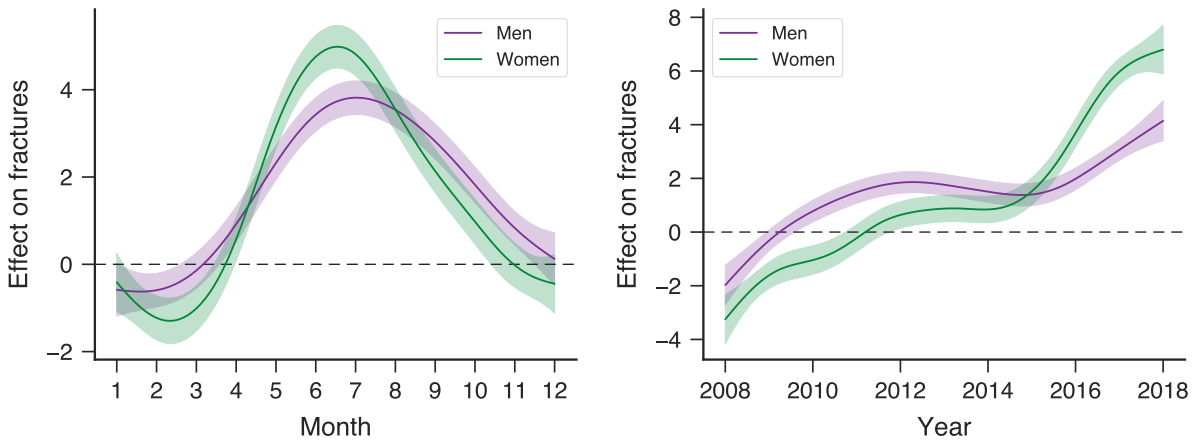

Figure 9 – Effects of the month and year on the incidence of femoral fractures. The left panel depicts the estimated smoothing term accounting for the effect of the month $[g_1(month_i)]$ on the fractures for men (purple curve) and women (green curve) of all ages.. The right panel shows the estimated smoothing functions of the year $[g_2(year_i)]$ for men (purple curve) and women (green curve) of all ages. In all plots, the shaded regions represent 95% confidence intervals.

We have also investigated the profiles of the smoothing terms accounting for the effect of the year $[g_2(year_i)]$. This function is shown in the right plot of Figure 9 when considering the time series for men (purple curve) and women (green curve) of all ages. We initially observe that the effect of the year on the femoral fractures have become more intense in more recent days, revealing an increasing trend in the incidence of femoral fractures over the years for both men and women. However, we do note that this increasing trend has a non-linear behavior that differs between men and women. There seems to exist a tipping point around the year 2015, from which the increasing trend becomes more intense for women than men. It is worth remarking that a linearly increasing trend in the incidence of femoral fractures was also found in the analysis of SOUZA et al. (2019). However, that linear regression was unable to reveal the non-linearities of the increasing trend reported in Figure 9.

Another important question related to our previous findings is whether they are dependent on the age of the people suffering from this medical condition. To address this question,

we have disaggregated our observations into 9 age groups (0-14 yo, 15-19 yo, 20-29 yo, 30-39 yo, 40-49 yo, 50-59 yo, 60-69 yo, 70-79 yo, and over 80 yo) and for each one we have applied the model of Eq. 4.1 and used the same approach when all ages were grouped together. Figure 10 shows the profiles of the smoothing terms accounting for the effect of the month for each age group and both for men and women. We observe that the impact of the month is either not significant or very weak for men and women under the age of 49 (that is, for the first six age groups in our dataset). The significance of the variable month starts to emerge only for age groups containing people older than 50 years. In particular, we notice a small but significant trend of higher incidence of femoral fractures for winter months for the age group containing people between 50 and 59 yo. As we move attention to older age groups, we start noticing an increase in the effect of the month on the fractures for both men and women. However, we do detect a gender discrepancy where the impact of the month on the fractures is higher for women than men. This is very clear for men and women older than 80 years, where the impact of mid-year months is about twice as higher for women. It is worth

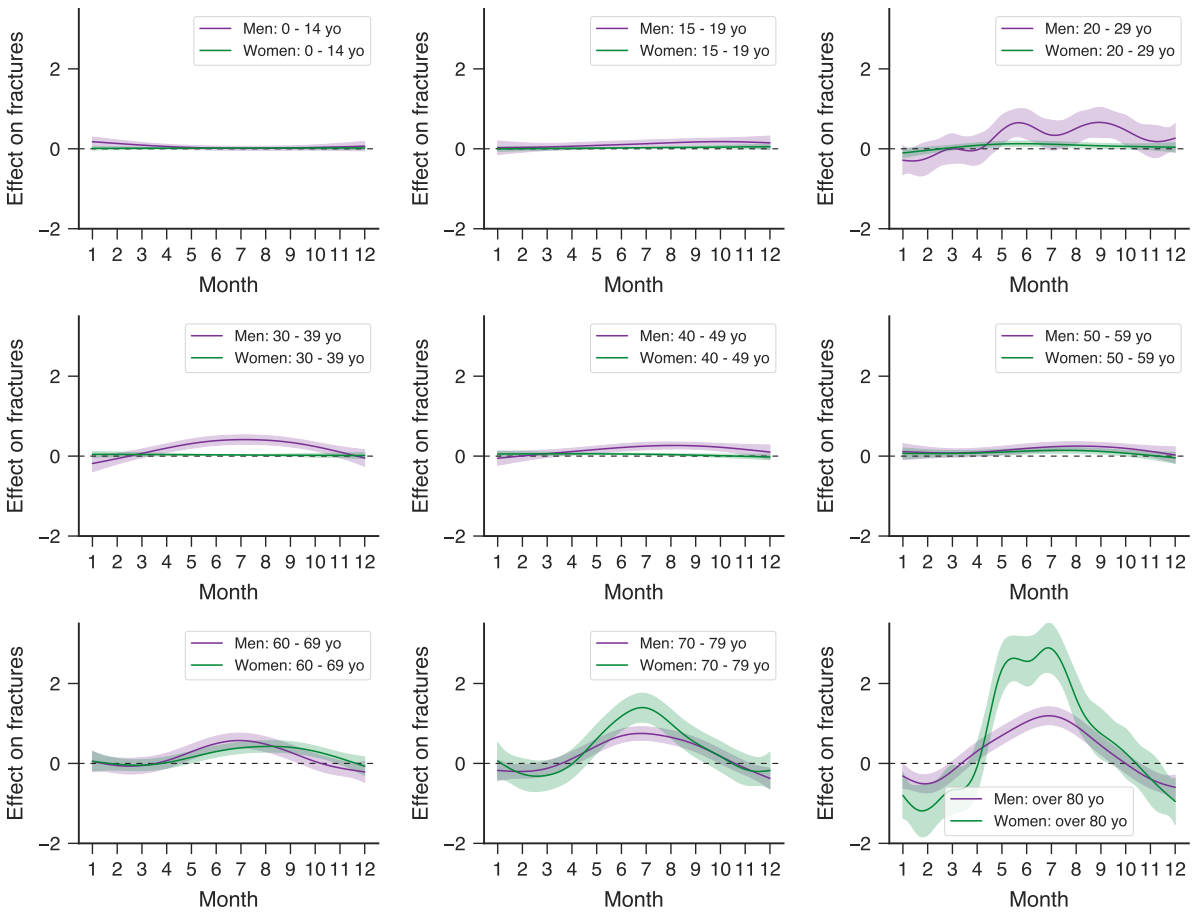

Figure 10 – Effects of the month on the incidence of femoral fractures across different age groups. Each panel shows the estimated smoothing term accounting for the effect of the month $[g_1(month_i)]$ on the fractures for men (purple curve) and women (green curve) for a particular age group (indicated in the plots). In all plots, the shaded regions represent 95% confidence intervals.

noticing that similar results regarding the age groups were reported by SOUZA et al. (2019), but the imperfect periodic nature of the effect of the month on the fractures captured by the generalized additive models indicates that the differences between senior men and senior women are higher than previously thought.

We have further examined the profiles of the smoothing terms accounting for the effect of the year on the fractures when considering different age groups. These functions are depicted in Figure 11, where we immediately observe that the long-term trends in the incidence of femoral fractures are quite distinct among different age groups and also between the two genders. Our results indicate that practically no long-term trend emerges for age groups containing people younger than 50 years. The effect of the year on the fractures displays a significant increasing trend only for age groups containing people older than 50 years. We also note that the older the group, the more intense is the effect of the year on the incidence of fractures. This is particularly evident for the group containing people older than 80 years. When analyzing the differences between the two genders, we observe that the effect of the year is more intense for older women, most noticeably for women older than 80 years.

Our results so far have been based on modeling the incidence of femoral fractures using month and year as predictors (that is, on the model of Eq. 4.1). Naturally, other variables may also play a role in the incidence of this medical condition. An interesting possibility is related to the number of holidays in the month associated with the observed number of femoral fractures. It is likely that the number of holidays may have a significant impact on modeling fractures among younger age groups since these events are often associated with traffic accidents. To test for this possibility, we have extended the model of Eq. 4.1 to include the number of holidays as a predictor variable, that is,

$$y_i = \beta_0 + g_1(\textit{month}_i) + g_2(\textit{year}_i) + g_3(\textit{holidays}_i) + e_i\,, \tag{4.2}$$

where $g_3(\dots)$ is a non-parametric smooth function of the number of holidays ($\textit{holidays}_i$) associated with the i-th observation.

We have fitted the model of Eq. 4.2 to our data considering 9 age groups for men and women. These fits are obtained through the same approach described in Methods Section 4.2.2 and the inclusion of this additional term related to the number of holidays does not change the results reported in Figure 10 and Figure 11. Figure 12 illustrates the profile of the smoothing term $g_3(\textit{holidays}_i)$ for the age groups 15-19 yo and 60-69 yo. We note that the number of holidays has a significant impact on the incidence of fractures among men of the group 15-19 yo, particularly when the number of holidays is larger than two. Practically no effect is observed for women aged between 15 and 19 years. Moreover, the number of holidays has no effect on the incidence of fractures of men and women of the age group 60-69 yo. Indeed, the profile of $g_3(\textit{holidays}_i)$ observed for the age group 60-69 yo is quite similar for all other age groups not shown in Figure 12, indicating that the number of holidays has no effect on the incidence

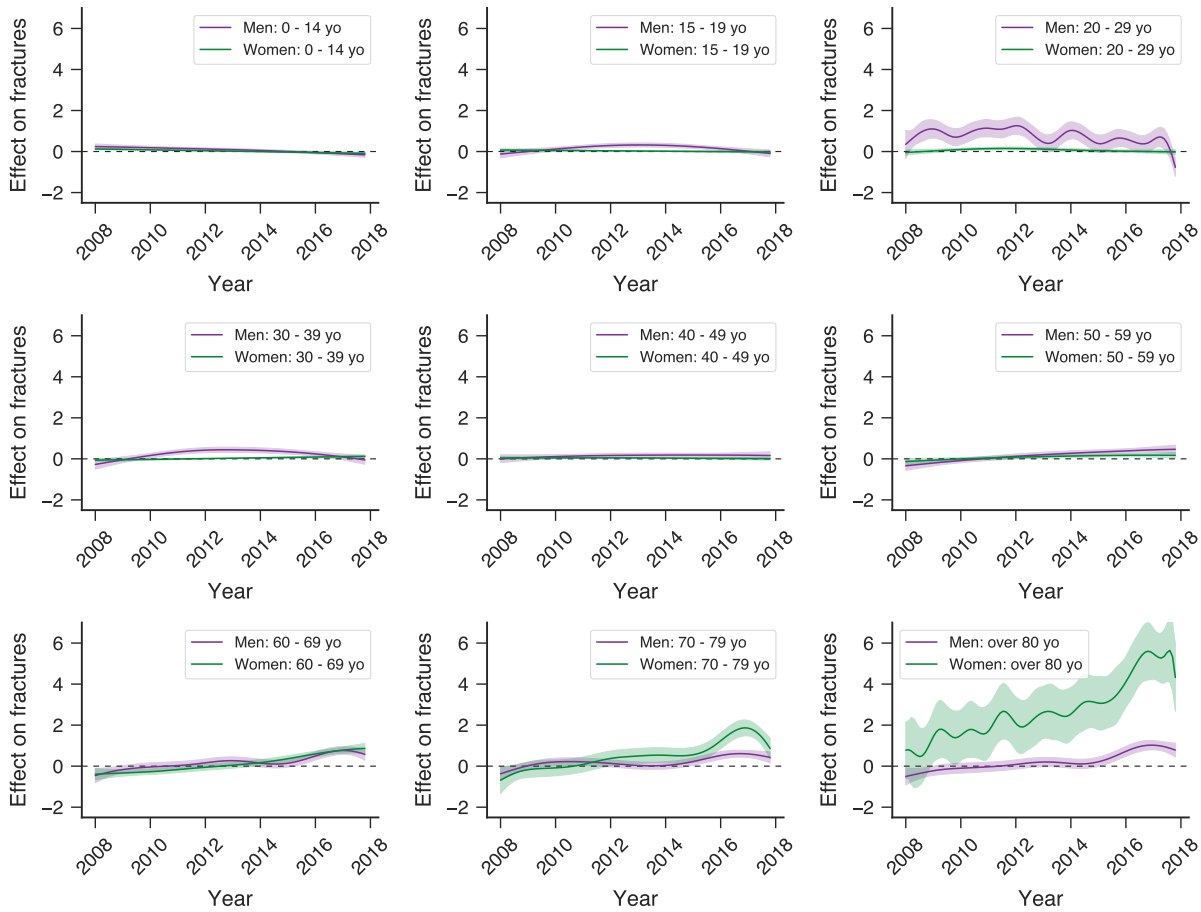

Figure 11 – Effects of the year on the incidence of femoral fractures across different age groups. Each panel shows the estimated smoothing term accounting for the effect of the year $[g_2(year_i)]$ for men (purple curve) and women (green curve) for a particular age group (indicated in the plots). In all plots, the shaded regions represent 95% confidence intervals.

of fractures among all age groups, except for men in the age group 15-19 yo. This result thus corroborates with the hypothesis that the incidence of fractures among young men is likely to be affected by the number of holidays in a month, probably because they tend to travel more during holidays which, in turn, make them more likely to get involved in traffic accidents.

4.4 Conclusions

We have presented a study on the incidence of femoral fractures in Brazil (State of São Paulo) along a period of ten years (2008-2017) by grouping data among gender and age group. For our analyses, we have relied on generalized additive models in which year, month, and number of holidays in a month are the predictors of the monthly average number of femoral fractures. Our results are somehow in line with those reported by SOUZA et al. (2019), where a linear model with a periodic seasonal term was used to describe the same dataset. However, our findings add knowledge to this problem because our approach allows for

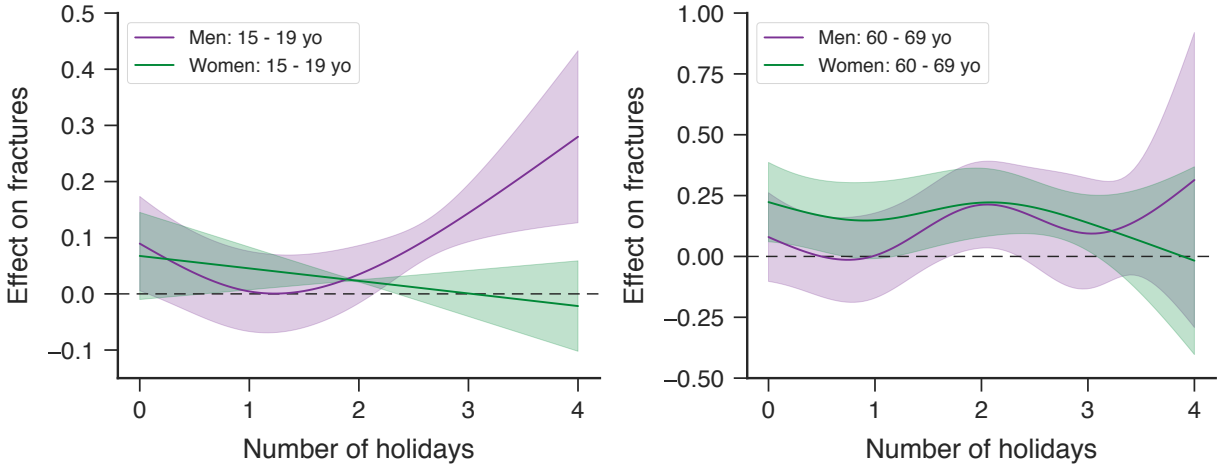

Figure 12 – Effects of the number of holidays on the incidence of femoral fractures between two different age groups. The left panel shows the results for men (purple curves) and women (green curves) of ages between 15 and 19 years, while the right panel shows the same for the age group 60-69 yo. We notice the number of holidays has a significant impact on the fractures of men of ages between 15 and 19 yo. In all plots, the shaded regions represent 95% confidence intervals.

the proper identification of non-linear effects of the predictors. This methodological difference was essential to reveal additional findings that would be missing with standard linear regression approaches. Particularly, our results have shown that the month has a significant impact on the occurrence of femoral fractures, where the effect of this predictor is more intense during winter time, reaching a maximum in mid-June for women and in early July for men. We have also found that the effect of the year has become more intense over time, but this long-term trend displays a non-linear behavior that differs for men and women. Indeed, there seems to exist a tipping point around the year 2015, from which this long-term increasing trend becomes much more intense for women.

When disaggregating our observations into 9 age groups, we have noticed that the incidence of fractures is remarkably different across the age ranges. Our findings indicate that the effect of the month is significant only among senior men and senior women. However, gender discrepancy increases with the age of the group up to the point that the effect of mid-year months is about twice as higher for women older than 80 years than for men of the same age. The long-term trends are also different among age groups, emerging as significant only for senior men and senior women. Again, the difference between men and women are exacerbated with age, particularly among the very elderly women (over 80 years). Conversely, other predictors such as the number of holidays in a month are likely to have a more significant impact on the incidence of femoral fractures among young men. In our analysis, we have found that the number of holidays impacts on the incidence femoral fractures of men aged between 15 and 19 years, but no effect is observed for women of the same age nor for men and women of all other age groups. Finally, we believe our work helps in reducing the shortage of studies about the incidence of femoral fractures in Southern Hemisphere places, and that our

non-linear approach may inspire new investigations on previously established findings in other areas.

Capítulo 5

Considerações finais

Esse trabalho apresentou dois estudos sobre padrões em séries temporais relacionadas à incidência de fraturas femorais na população do estado de São Paulo ao longo do período de janeiro de 2008 a dezembro de 2017. Para isso, inicialmente realizou-se uma revisão sobre conceitos de séries temporais, modelos de regressão, inferência Bayesiana e modelos aditivos generalizados. No primeiro estudo, descrito no Capítulo 3, a incidência de fraturas femorais foi modelada utilizando um modelo de regressão linear, no qual um termo sazonal (no forma de senos e cossenos) foi explicitamente suposto, bem como um erro na forma de um processo autoregressivo de ordem 1. Esse modelo foi ajustado aos dados utilizando métodos Bayesianos, os quais permitiram a construção de intervalos de credibilidade de forma mais simples e direta. Ao analisar a forma do termo sazonal ajustado aos dados, observou-se que a maioria dos casos ocorre nos meses mais frios do ano (junho, julho e agosto), com diferentes padrões para homens e mulheres, dependendo ainda da faixa etária. Em particular, verificou-se que a incidência de fraturas femorais é maior entre homens até os 60 anos do que entre mulheres da mesma idade. Porém, a partir dos 70 anos a incidência é muito maior entre mulheres do que entre homens. Esses resultados foram publicados na Revista de Saúde Pública (SOUZA et al., 2019).

Já o segundo estudo, descrito no Capítulo 4, modela os mesmos dados utilizando a abordagem conhecida como modelos aditivos generalizados. Nesse contexto, a incidência de fraturas do fêmur foi modelada como um função suavizada do mês e ano. Diferentemente dos resultados no Capítulo 3, essa abordagem não pressupõe uma forma funcional específica para o termo da sazonalidade nem para o termo de tendência de longo prazo. De fato, essas duas quantidades são "aprendidas" diretamente dos dados, o que proporciona uma maior flexibilidade ao modelo. Os resultados obtidos utilizando os modelos aditivos generalizados corroboraram todos aqueles que foram encontrados no Capítulo 3; porém, a forma mais geral da segunda abordagem permitiu a identificação de alguns comportamentos que não estavam presentes da análise baseada em regressão linear do Capítulo 3. Entre esses comportamentos,

verificou-se que o mês com maior impacto na ocorrência de fraturas é ligeiramente diferente entre homens e mulheres (meados de junho para mulheres e início de julho para homens). Observou-se também um crescimento não linear da tendência de longo prazo da incidência de fraturas, a qual passou a ser muito mais intensa para as mulheres a partir do ano de 2015. Além disso, mostrou-se que o número de feriados no mês tem impacto maior na incidência de fraturas entre jovens do sexo masculino do que os efeitos de sazonalidade. Esses resultados estão sendo finalizados para posterior submissão para publicação.

De modo geral, acredita-se que esse trabalho contribui para o melhor entendimento dos padrões temporais de ocorrência de fraturas de fêmur em regiões localizadas no hemisfério sul. Conforme foi discutido anteriormente, a grande maioria dos estudos existentes na literatura está baseada em dados de países desenvolvidos localizados no hemisfério norte. São poucos os trabalhos que buscaram analisar a incidência de fraturas femorais no Brasil e outros países do hemisfério sul. Sendo assim, acredita-se que a presente dissertação colabora em reduzir um pouco a escassez desse tipo de trabalho. Em adição à componente acadêmica desses estudos, o conhecimento das particularidades de cada região no tocante ao incidência de fraturas pode ajudar também profissionais da saúde e administradores de hospitais a tomarem melhores decisões sobre a escala de profissionais e a alocação de recursos visando um melhor atendimento da população.

Referências

ANTUNES, J. L. F. et al. Using time series analysis in epidemiological studies. *Epidemiologia e Serviços de Saúde*, v. 24, n. 3, p. 565–576, set. 2015. Citado 3 vezes nas páginas 17, 33 e 38.

ARNDT, A. B. M.; TELLES, J. L.; KOWALSKI, S. C. Direct cost of femoral fracture from falls in elderly: analysis of the private section health in the city of Brasília, 2009. *Revista Brasileira de Geriatria e Gerontologia*, v. 14, n. 2, p. 221–231, jun. 2011. Citado na página 31.

BALEN, R. van et al. Hip fracture in elderly patients: outcomes for function, quality of life, and type of residence. *Clinical Orthopaedics and Related Research*, n. 390, p. 232–243, set. 2001. Citado na página 12.

BASTOW, M. D.; RAWLINGS, J.; ALLISON, S. P. Undernutrition, hypothermia, and injury in elderly women with fractured femur: an injury response to altered metabolism? *Lancet (London, England)*, v. 1, n. 8317, p. 143–146, jan. 1983. Citado na página 37.

BITTENCOURT, S. A.; CAMACHO, L. A. B.; LEAL, M. d. C. O sistema de informação hospitalar e sua aplicação na saúde coletiva. *Cadernos de Saúde Pública*, SciELO Public Health, v. 22, p. 19–30, 2006. Citado na página 39.

CASELLA, G.; GEORGE, E. Explaning the gibbs sampler. v. 46, p. 167–174, 08 1992. Citado na página 20.

CHIU, K. Y.; NG, T. P.; CHOW, S. P. Seasonal variation of fractures of the hip in elderly persons. *Injury*, v. 27, n. 5, p. 333–336, jun. 1996. Citado 6 vezes nas páginas 12, 32, 37, 41, 44 e 45.

CLAGUE, J. E. et al. Predictors of outcome following hip fracture. Admission time predicts length of stay and in-hospital mortality. *Injury*, v. 33, n. 1, p. 1–6, jan. 2002. Citado na página 12.

CLARK, S. Osteoporosis - the disease of the 21st century? *The Lancet*, Elsevier, v. 359, n. 9319, p. 1714, 2002. Citado 2 vezes nas páginas 40 e 41.

CONGDON, P. *Applied bayesian modelling*. [S.l.]: John Wiley & Sons, 2014. v. 595. Citado na página 18.

COOPER, C.; CAMPION, G.; MELTON, L. r. Hip fractures in the elderly: a world-wide projection. *Osteoporosis international*, Springer, v. 2, n. 6, p. 285–289, 1992. Citado 2 vezes nas páginas 40 e 41.

COOPER, C. et al. Secular trends in the incidence of hip and other osteoporotic fractures. *Osteoporosis International*, v. 22, n. 5, p. 1277, 2011. Citado na página 40.

CRAWFORD, J. R.; PARKER, M. J. Seasonal variation of proximal femoral fractures in the United Kingdom. *Injury*, v. 34, n. 3, p. 223–225, mar. 2003. Citado 6 vezes nas páginas 12, 32, 37, 41, 44 e 45.

CRAWLEY, M. *Statistics: An Introduction using R*. [S.l.]: Wiley, 2005. (Statistics: An Introduction Using R). ISBN 9780470022979. Citado na página 33.

CRYER, J.; CHAN, K. *Time Series Analysis: With Applications in R*. [S.l.]: Springer, 2008. (Springer Texts in Statistics). ISBN 9780387759586. Citado na página 21.

CUMMINGS, S. R.; MELTON, L. J. Epidemiology and outcomes of osteoporotic fractures. *The Lancet*, Elsevier, v. 359, n. 9319, p. 1761–1767, 2002. Citado na página 41.

CUNHA, P. T. S. d. et al. Hip fractures in the elderly: surgical treatment timing and its correlation with delirium and infection. *Acta Ortopédica Brasileira*, v. 16, n. 3, p. 173–176, 2008. Citado na página 12.

DICKEY, D. A.; FULLER, W. A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, v. 74, n. 366, p. 427–431, 1979. Citado na página 23.

DORE, N. et al. Improving care after hip fracture: the fracture? Think osteoporosis (FTOP) program. *BMC Geriatrics*, v. 13, p. 130, dez. 2013. Citado 2 vezes nas páginas 11 e 31.

DOUGLAS, S. et al. Seasonal variation of hip fracture at three latitudes. *Injury*, v. 31, n. 1, p. 11 – 19, 2000. Citado 2 vezes nas páginas 32 e 37.

EVANS, J. G.; SEAGROATT, V.; GOLDACRE, M. Secular trends in proximal femoral fracture, oxford record linkage study area and england 1968-86. *Journal of Epidemiology & Community Health*, BMJ Publishing Group Ltd, v. 51, n. 4, p. 424–429, 1997. Citado na página 41.

GIL-ALANA, L. A. Long memory, seasonality and time trends in the average monthly temperatures in Alaska. *Theoretical and Applied Climatology*, v. 108, n. 3-4, p. 385–396, maio 2012. Citado 2 vezes nas páginas 17 e 33.

HAENTJENS, P. et al. Meta-analysis: excess mortality after hip fracture among older women and men. *Annals of Internal Medicine*, v. 152, n. 6, p. 380–390, 2010. Citado 2 vezes nas páginas 31 e 41.

HANNAN, E. L. et al. Mortality and Locomotion 6 Months After Hospitalization for Hip Fracture: Risk Factors and Risk-Adjusted Hospital Outcomes. *JAMA*, v. 285, n. 21, p. 2736–2742, jun. 2001. Citado na página 12.

HYNDMAN, R. J. *fpp: Data for "Forecasting: principles and practice"*. [S.l.], 2013. R package version 0.5. Disponível em: <https://CRAN.R-project.org/package=fpp>. Citado na página 23.

JACOBSEN, S. J. et al. Seasonal variation in the incidence of hip fracture among white persons aged 65 years and older in the United States, 1984-1987. *American Journal of Epidemiology*, v. 133, n. 10, p. 996–1004, maio 1991. Citado 6 vezes nas páginas 12, 32, 37, 41, 44 e 45.

JACOBSEN, S. J. et al. Population-based study of the contribution of weather to hip fracture seasonality. *American Journal of Epidemiology*, v. 141, n. 1, p. 79–83, jan. 1995. Citado 6 vezes nas páginas 12, 32, 37, 41, 44 e 45.

JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112. Citado 6 vezes nas páginas 25, 26, 27, 28, 29 e 43.

JENA, D.; MUDDU, B. N.; RICHARDSON, J. B. Seasonal variation of proximal femoral fractures. *Orthopaedic Proceedings*, v. 85-B, n. SUPP_II, p. 117–117, fev. 2003. Citado 6 vezes nas páginas 12, 32, 37, 41, 44 e 45.

KHARAZMI, M. et al. Mortality After Atypical Femoral Fractures: A Cohort Study. *Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research*, v. 31, n. 3, p. 491–497, mar. 2016. Citado na página 31.

LARSEN, K. Gam: The predictive modeling silver bullet. *Multithreaded. Stitch Fix*, v. 30, 2015. Citado na página 30.

LAU, E. M. et al. The seasonality of hip fracture and its relationship with weather conditions in New South Wales. *Australian Journal of Public Health*, v. 19, n. 1, p. 76–80, fev. 1995. Citado 5 vezes nas páginas 32, 37, 41, 44 e 45.

LUNN, D. et al. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, v. 28, n. 25, p. 3049–3067, nov. 2009. Citado 2 vezes nas páginas 13 e 19.

MAZZUCCHELLI, R. et al. Short-term association between outdoor air pollution and osteoporotic hip fracture. *Osteoporosis International*, v. 29, n. 10, p. 2231–2241, Oct 2018. Citado na página 37.

MORETTIN, P.; TOLOI, C. de C. *Análise de séries temporais*. [S.l.]: Edgard Blucher, 2006. (ABE - Projeto Fisher). ISBN 9788521203896. Citado 4 vezes nas páginas 11, 14, 23 e 24.

PARKER, M. J.; MARTIN, S. Falls, hip fractures and the weather. *European Journal of Epidemiology*, v. 10, n. 4, p. 441–442, ago. 1994. Citado 6 vezes nas páginas 12, 32, 37, 41, 44 e 45.

PEDRAZZONI, M. et al. Seasonal variation in the incidence of hip fractures in Emilia-Romagna and Parma. *Bone*, v. 14 Suppl 1, p. S57–63, 1993. Citado 6 vezes nas páginas 12, 32, 37, 41, 44 e 45.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org/>. Citado na página 24.

RAJU, R. et al. A study to evaluate functional outcomes of distal femur fractures. *International Journal of Orthopaedics Sciences*, v. 3, n. 4i, p. 602–605, out. 2017. Citado na página 31.

RALIS, Z. A. Epidemic of fractures during period of snow and ice. *British Medical Journal (Clinical Research Ed.)*, v. 282, n. 6264, p. 603–605, fev. 1981. Citado 5 vezes nas páginas 32, 37, 41, 44 e 45.

SAKAKI, M. H. et al. Study of the proximal femoral fractures mortatlity in elderly patients. *Acta Ortopédica Brasileira*, v. 12, n. 4, p. 242–249, 2004. Citado 2 vezes nas páginas 12 e 31.

SERVéN, D. et al. *dswah/pyGAM: v0.8.0*. 2018. Disponível em: <https://doi.org/10.5281/zenodo.1476122>. Citado na página 30.

SOARES, D. S. et al. Femoral fractures in elderly Brazilians: a spatial and temporal analysis from 2008 to 2012. *Cadernos de Saúde Pública*, v. 30, n. 12, p. 2669–2678, dez. 2014. Citado na página 12.

SOUZA, M. M. et al. Seasonal variation of femoral fractures in the state of são paulo, southeast brazil. *Revista de Saúde Pública*, 2019. Citado 4 vezes nas páginas 45, 47, 48 e 51.

SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002. Citado na página 16.

WOOD, S. N. *Generalized additive models: an introduction with R*. [S.l.]: Chapman and Hall/CRC, 2017. Citado 2 vezes nas páginas 29 e 43.

Anexos

ANEXO A

Normas da Revista de Saúde Pública

Neste anexo apresentamos as normas para publicação da Revista de Saúde Pública, na qual foi publicado o Capítulo 3 na forma de artigo (intitulado "Seasonal variation offemoral fractures in the State ofSão Paulo, Southeast Brazil").

RSP Revista de Saúde Pública

ISSN 0034-8910 *versão impressa*
ISSN 1518-8787 *versão on-line*

INSTRUÇÕES AOS AUTORES

- Categorias de artigos
- Dados de identificação do manuscrito
- Conflito de Interesses
- Declaração de Documentos
- Preparo do Manuscrito
- Processo Editorial
- Taxa de Publicação
- Suplementos

Categorias de Artigos

São aceitos manuscritos nos idiomas: português, espanhol e inglês.

O texto de manuscrito de pesquisa original deve seguir a estrutura conhecida como IMRD: Introdução, Métodos, Resultados e Discussão (Estrutura do Texto). Manuscritos baseados em pesquisa qualitativa podem ter outros formatos, admitindo-se Resultados e Discussão em uma mesma seção e Considerações Finais/Conclusões. Outras categorias de manuscritos (revisões, comentários, etc.) seguem os formatos de texto a elas apropriados.

Os estudos devem ser apresentados de forma que qualquer pesquisador interessado possa reproduzir os resultados. Para isso estimulamos o uso das seguintes **recomendações**, de acordo com a categoria do manuscrito submetido:

·**CONSORT** checklist e fluxograma para ensaios controlados e randomizados
·**STARD** checklist e fluxograma para estudos de acurácia diagnóstica
·**MOOSE** checklist e fluxograma para metanálises e revisões sistemáticas de estudos observacionais
·**PRISMA** checklist e fluxograma para revisões sistemáticas e metanálises
·**STROBE** checklist para estudos observacionais em epidemiologia
·**RATS** checklist para estudos qualitativos

Pormenores sobre os itens exigidos para apresentação do manuscrito estão descritos de acordo com a categoria de artigos

Categorias de artigos

a) Artigos Originais

Incluem estudos observacionais, estudos experimentais ou quase-experimentais, avaliação de programas, análises de custo-efetividade, análises de decisão e estudos sobre avaliação de desempenho de testes diagnósticos para triagem populacional. Cada artigo deve conter objetivos e hipóteses claras, desenho e métodos utilizados, resultados, discussão e conclusões.
Incluem também ensaios teóricos (críticas e formulação de conhecimentos teóricos relevantes) e artigos dedicados à apresentação e discussão de aspectos metodológicos e técnicas utilizadas na pesquisa em saúde pública. Neste caso, o texto deve ser organizado em tópicos para guiar o leitor quanto aos elementos essenciais do argumento desenvolvido.

Instrumentos de aferição em pesquisas populacionais
Manuscritos abordando instrumentos de aferição podem incluir aspectos relativos ao desenvolvimento, a avaliação e à adaptação transcultural para uso em estudos populacionais, excluindo-se aqueles de aplicação clínica, que não se incluem no escopo da RSP.

Aos manuscritos de instrumentos de aferição, recomenda-se que seja apresentada uma apreciação detalhada do construto a ser avaliado, incluindo seu possível gradiente de intensidade e suas eventuais subdimensões. O desenvolvimento de novo instrumento deve estar amparado em revisão de literatura, que identifique explicitamente a insuficiência de propostas prévias e justifique a necessidade de novo instrumental.

Deve ser detalhada a proposição, a seleção e a confecção dos itens, bem como o emprego de estratégias para adequá-los às definições do construto, incluindo o uso de técnicas qualitativas de pesquisa (entrevistas em profundidade, grupos focais etc.), reuniões com painéis de especialistas, entre outras. O trajeto percorrido na definição da forma de mensuração dos itens e a realização de pré-testes com seus conjuntos preliminares necessitam ser descritos no texto. A avaliação das validades de face, conteúdo, critério, construto e/ou dimensional deve ser apresentada em detalhe.

Análises de confiabilidade do instrumento também devem ser apresentadas e discutidas, incluindo-se medidas de consistência interna, confiabilidade teste-reteste e/ou concordância inter-observador. Os autores devem expor o processo de seleção do instrumento final e situá-lo em perspectiva crítica e comparativa com outros instrumentos destinados a avaliar o mesmo construto ou construtos semelhantes.

Para os manuscritos sobre **adaptação transcultural** de instrumentos de aferição, além de atender, de forma geral, às recomendações supracitadas, faz-se necessário explicitar o modelo teórico norteador do processo. Os autores devem, igualmente, justificar a escolha de determinado instrumento para adaptação a um contexto sociocultural específico, com base em minuciosa revisão de literatura. Finalmente, devem indicar explicitamente quais e como foram seguidas as etapas do modelo teórico de adaptação no trabalho submetido para publicação.
Obs: O instrumento de aferição deve ser incluído como anexo dos artigos submetidos.

No preparo do manuscrito, além das recomendações citadas, verifique as instruções de formatação a seguir.

Formatação:

- Devem conter até 3500 palavras (excluindo resumos, tabelas, figuras e referências).
- Número de tabelas/figuras: até 5 no total.
- Número de referências: até 30 no total.
- Resumos no formato estruturado com até 300 palavras.

b) Comunicações breves – São relatos curtos de achados que apresentam interesse para a saúde pública, mas que não comportam uma análise mais abrangente e uma discussão de maior fôlego.

Formatação:
Sua apresentação deve acompanhar as mesmas normas exigidas para artigos originais.

- Devem conter até 1500 palavras (excluindo resumos tabelas, figuras e referências)
- Número de tabelas/figuras: uma tabela ou figura.
- Número de referências: até 5 no total.
- Resumos no formato narrativo com até 100 palavras.

c) Artigos de revisão

Revisão sistemática e meta-análise - Por meio da síntese de resultados de estudos originais, quantitativos ou qualitativos, objetiva responder à pergunta específica e de relevância para a saúde pública. Descreve com pormenores o processo de busca dos estudos originais, os critérios utilizados para seleção daqueles que foram incluídos na revisão e os procedimentos empregados na síntese dos resultados obtidos pelos estudos revisados. Consulte:

MOOSE checklist e fluxograma para metanálises e revisões sistemáticas de estudos observacionais
PRISMA checklist e fluxograma para revisões sistemáticas e metanálises

Revisão narrativa/crítica - A revisão narrativa ou revisão crítica apresenta caráter descritivo-discursivo, dedicando-se à apresentação compreensiva e à discussão de temas de interesse científico no campo da Saúde Pública. Deve apresentar formulação clara de um objeto científico de interesse, argumentação lógica, crítica teórico-metodológica dos trabalhos consultados e síntese

conclusiva. Deve ser elaborada por pesquisadores com experiência no campo em questão ou por especialistas de reconhecido saber.

Formatação:

- Devem conter até 4000 palavras (excluindo resumos, tabelas, figuras e referências).
- Número de tabelas/figuras: até 5 no total.
- Número de referências: sem limites.
- Resumos no formato estruturado com até 300 palavras, ou narrativo com até 150 palavras.

d) Comentários

Visam a estimular a discussão, introduzir o debate e "oxigenar" controvérsias sobre aspectos relevantes da saúde pública. O texto deve ser organizado em tópicos ou subitens destacando na Introdução o assunto e sua importância. As referências citadas devem dar sustentação aos principais aspectos abordados no artigo.

Formatação:

- Devem conter até 2000 palavras (excluindo resumos, tabelas, figuras e referências).
- Número de referências: até 30 no total.
- Número de tabelas/figuras: até 5 no total.
- Resumos no formato narrativo com até 150 palavras.

Publicam-se também Cartas Ao Editor com até 600 palavras e até 5 referências.

Dados de Identificação do Manuscrito

Autoria

O conceito de autoria está baseado na contribuição substancial de cada uma das pessoas listadas como autores, no que se refere sobretudo à concepção do projeto de pesquisa, análise e interpretação dos dados, redação e revisão crítica. A contribuição de cada um dos autores deve ser explicitada em declaração para esta finalidade. Não se justifica a inclusão de nome de autores cuja contribuição não se enquadre nos critérios acima.

Dados de identificação dos autores (cadastro)

Nome e sobrenome: O autor deve seguir o formato pelo qual já é indexado nas bases de dados.

Correspondência: Deve constar o nome e endereço do autor responsável para troca de correspondência.

Instituição: Podem ser incluídas até três hierarquias institucionais de afiliação (por exemplo: departamento, faculdade, universidade).

Coautores: Identificar os coautores do manuscrito pelo nome, sobrenome e instituição, conforme a ordem de autoria.

Financiamento da pesquisa: Se a pesquisa foi subvencionada, indicar o tipo de auxílio, o nome da agência financiadora e o respectivo número do processo.

Apresentação prévia: Tendo sido apresentado em reunião científica, indicar o nome do evento, local e ano da realização.

Conflito de Interesses

Quando baseado em tese ou dissertação, indicar o nome do autor, título, ano, nome do programa de pós-graduação e instituição onde foi apresentada.

A confiabilidade pública no processo de revisão por pares e a credibilidade de artigos publicados dependem em parte de como os conflitos de interesses são administrados durante a redação, revisão por pares e tomada de decisões pelos editores.

Conflitos de interesses podem surgir quando autores, revisores ou editores possuem interesses que, aparentes ou não, podem influenciar a elaboração ou avaliação de manuscritos. O conflito de interesses pode ser de natureza pessoal, comercial, política, acadêmica ou financeira.

Quando os autores submetem um manuscrito, eles são responsáveis por reconhecer e revelar conflitos financeiros ou de outra natureza que possam ter influenciado seu trabalho. Os autores devem reconhecer no manuscrito todo o apoio financeiro para o trabalho e outras conexões financeiras ou pessoais com relação à pesquisa. O relator deve revelar aos editores quaisquer conflitos de interesse que poderiam influir em sua opinião sobre o manuscrito, e, quando couber, deve declarar-se não qualificado para revisá-lo.

Se os autores não tiverem certos do que pode constituir um potencial conflito de interesses, devem contatar a secretaria editorial da Revista.

Declaração de Documentos

Em conformidade com as diretrizes do **International Committee of Medical Journal Editors**, são solicitados alguns documentos e declarações do (s) autor (es) para a avaliação de seu manuscrito. Observe a relação dos documentos abaixo e, nos casos em que se aplique, anexe o documento ao processo. O momento em que tais documentos serão solicitados é variável:

Documento/declaração	Quem assina	Quando anexar
a. Carta de Apresentação	Todos os autores	Na submissão
b. Declaração de responsabilidade	Todos os autores	Na submissão
c. Responsabilidade pelos Agradecimentos	Autor responsável	Após a aprovação
d. Transferência de Direitos Autorais	Todos os autores	Após a aprovação

a) CARTA DE APRESENTAÇÃO

A carta deve ser assinada por todos os autores e deve conter:

- Informações sobre os achados e conclusões mais importantes do manuscrito, esclarecendo seu significado para a saúde pública.
- Se os autores têm artigos publicados na linha de pesquisa do manuscrito, mencionar até três.
- Declaração de responsabilidade de cada autor: ter contribuído substancialmente para a concepção e planejamento, ou análise e interpretação dos dados; ter contribuído significativamente na elaboração do rascunho ou na revisão crítica do conteúdo; e ter participado da aprovação da versão final do manuscrito. Para maiores informações sobre critérios de autoria, consulte o site da RSP.
- Declaração de potenciais conflitos de interesses dos autores.
- Atestar a exclusividade da submissão do manuscrito à RSP.

Responder- Qual a novidade do seu estudo? Por que deve ser publicado nesta revista?

b. DECLARAÇÃO DE RESPONSABILIDADE
Segundo o critério de autoria do *International Committee of Medical Journal Editors*, autores devem contemplar todas as seguintes condições: (1) Contribuí substancialmente para a concepção e planejamento, ou análise e interpretação dos dados; (2) Contribuí significativamente na elaboração do rascunho ou na revisão crítica do conteúdo; e (3) Participei da aprovação da versão final do manuscrito.

No caso de grupo grande ou multicêntrico ter desenvolvido o trabalho, o grupo deve identificar os indivíduos que aceitam a responsabilidade direta pelo manuscrito. Esses indivíduos devem contemplar totalmente os critérios para autoria definidos acima e os editores solicitarão a eles as declarações exigidas na submissão de manuscritos. O autor correspondente deve indicar claramente a forma de citação preferida para o nome do grupo e identificar seus membros. Normalmente serão listados no final do texto do artigo.

Aquisição de financiamento, coleta de dados, ou supervisão geral de grupos de pesquisa, somente, não justificam autoria.

Todas as pessoas relacionadas como autores devem assinar declaração de responsabilidade.

Aquisição de financiamento, coleta de dados, ou supervisão geral de grupos de pesquisa, somente, não justificam autoria.

Todas as pessoas relacionadas como autores devem assinar declaração de responsabilidade.

c. AGRADECIMENTOS

Devem ser mencionados os nomes de pessoas que, embora não preencham os requisitos de autoria, prestaram colaboração ao trabalho. Será preciso explicitar o motivo do agradecimento, por exemplo, consultoria científica, revisão crítica do manuscrito, coleta de dados, etc. Deve haver permissão expressa dos nomeados e o autor responsável deve anexar a Declaração de Responsabilidade pelos Agradecimentos. Também pode constar desta parte apoio logístico de instituições.

d. TRANSFERÊNCIA DE DIREITOS AUTORAIS

Todos os autores devem ler, assinar e enviar documento transferindo os direitos autorais. O artigo só será liberado para publicação quando esse documento estiver de posse da RSP.

Preparo do Manuscrito

Título no idioma original do manuscrito e em inglês
O título deve ser conciso e completo, contendo informações relevantes que possibilitem recuperação do artigo nas bases de dados. O limite é de 90 caracteres, incluindo espaços. Se o manuscrito for submetido em inglês, fornecer também o título em português.

Título resumido
Deve conter até 45 caracteres.

Descritores
Devem ser indicados entre 3 a 10, extraídos do vocabulário **"Descritores em Ciências da Saúde" (DeCS)**, nos idiomas português, espanhol e inglês, com base no **Medical Subject Headings (MeSH)**. Se não forem encontrados descritores adequados para a temática do manuscrito, poderão ser indicados termos livres (ou *key words*) mesmo não existentes nos vocabulários citados.

Figuras e Tabelas
Todos os elementos gráficos ou tabulares apresentados serão identificados como figura ou tabela, e numerados sequencialmente a partir de um, e não como quadros, gráficos,etc.

Resumo
São publicados resumos em português, espanhol e inglês. Para fins de cadastro do manuscrito, deve-se apresentar dois resumos, um na língua original do manuscrito e outro em inglês (ou em português, em caso de manuscrito apresentado em inglês). As especificações quanto ao tipo de resumo estão descritas em cada uma das **categorias de artigos**. Como regra geral, o resumo deve incluir: objetivo do estudo, principais procedimentos metodológicos (população em estudo, local e ano de realização, métodos observacionais e analíticos), principais resultados e conclusões.

Estrutura do texto
Introdução – Deve ser curta, relatando o contexto e a justificativa do estudo, apoiados em referências pertinentes ao objetivo do manuscrito, que deve estar explícito no final desta parte. Não devem ser mencionados resultados ou conclusões do estudo que está sendo apresentado.

Métodos – Os procedimentos adotados devem ser descritos claramente; bem como as variáveis analisadas, com a respectiva definição quando necessária e a hipótese a ser testada. Devem ser descritas a população e a amostra, instrumentos de medida, com a apresentação, se possível, de medidas de validade; e conter informações sobre a coleta e processamento de dados. Deve ser incluída a devida referência para os métodos e técnicas empregados, inclusive os métodos estatísticos; métodos novos ou substancialmente modificados devem ser descritos, justificando as razões para seu uso e mencionando suas limitações. Os critérios éticos da pesquisa devem ser respeitados. Os autores devem explicitar que a pesquisa foi conduzida dentro dos padrões éticos e aprovada por comitê de ética.

Resultados – Devem ser apresentados em uma sequência lógica, iniciando-se com a descrição dos dados mais importantes. Tabelas e figuras devem ser restritas àquelas necessárias para argumentação e a descrição dos dados no texto deve ser restrita aos mais importantes. Os gráficos devem ser utilizados para destacar os resultados mais relevantes e resumir relações complexas. Dados em gráficos e tabelas não devem ser duplicados, nem repetidos no texto. Os resultados numéricos devem especificar os métodos estatísticos utilizados na análise. Material extra ou suplementar e detalhes técnicos podem ser divulgados na versão eletrônica do artigo.

Discussão – A partir dos dados obtidos e resultados alcançados, os novos e importantes aspectos observados devem ser interpretados à luz da literatura científica e das teorias existentes no campo. Argumentos e provas baseadas em comunicação de caráter pessoal ou divulgadas em documentos restritos não podem servir de apoio às argumentações do autor. Tanto as limitações do trabalho quanto suas implicações para futuras pesquisas devem ser esclarecidas. Incluir somente hipóteses e generalizações baseadas nos dados do trabalho. As conclusões devem finalizar esta parte, retomando o objetivo do trabalho.

Referências
Listagem: As referências devem ser normalizadas de acordo com o **estilo *Vancouver* - Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication**, ordenadas por ordem de citação. Os títulos de periódicos devem ser referidos de forma abreviada, de acordo com o PubMed e grafados no formato itálico. No caso de publicações com até seis autores, citam-se todos; acima de seis, citam-se os seis primeiros, seguidos da expressão latina “et al”. Referências de um mesmo autor devem ser organizadas em ordem cronológica crescente. Sempre que possível incluir o DOI do documentado citado, de acordo com os exemplos a seguir.

Exemplos:
Artigos de periódicos
Narvai PC. Cárie dentária e flúor: uma relação do século XX. *Cienc Saude Coletiva*. 2000;5(2):381-92. DOI:10.1590/S1413-81232000000200011
Zinn-Souza LC, Nagai R, Teixeira LR, Latorre MRDO, Roberts R, Cooper SP, et al. Fatores associados a sintomas depressivos em estudantes do ensino médio de São Paulo, Brasil. *Rev Saude Publica*. 2008;42(1):34-40. DOI:10.1590/S0034-89102008000100005

Livros

Wunsch Filho V, Koifman S. Tumores malignos relacionados com o trabalho. In: Mendes R, coordenador. Patologia do trabalho. 2. ed. São Paulo: Atheneu; 2003. v.2, p. 990-1040.

Foley KM, Gelband H, editors. Improving palliative care for cancer Washington: National Academy Press; 2001[citado 2003 jul 13] Disponível em: http://www.nap.edu/catalog.php?record_id=10149

Para outros exemplos recomendamos consultar as normas (“Citing Medicine”) da National Library of Medicine, disponível em
http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=citmed.
Referências a documentos não indexados na literatura científica mundial, em geral de divulgação circunscrita a uma instituição ou a um evento (teses, relatórios de pesquisa, comunicações em eventos, dentre outros) e informações extraídas de documentos eletrônicos, não mantidas permanentemente em sites, se relevantes, devem figurar no rodapé das páginas do texto onde foram citadas.

Citação no texto
A referência deve ser indicada pelo seu número na listagem, na forma de **expoente** antes da pontuação no texto, sem uso de parênteses, colchetes ou similares. Nos casos em que a citação do nome do autor e ano for relevante, o número da referência deve ser colocado a seguir do nome do autor. Trabalhos com dois autores devem fazer referência aos dois autores ligados por “e”. Nos outros casos apresentar apenas o primeiro autor (seguido de ‘et al.’ em caso de autoria múltipla).

Exemplos:
A promoção da saúde da população tem como referência o artigo de Evans e

Stoddart9, que considera a distribuição de renda, desenvolvimento social e reação individual na determinação dos processos de saúde-doença.

Segundo Lima et al.9 (2006), a prevalência de transtornos mentais em estudantes de medicina é maior do que na população em geral.

Tabelas
Devem ser apresentadas no final do texto, após as referências bibliográficas, numeradas consecutivamente com algarismos arábicos, na ordem em que foram citadas no texto. A cada uma deve-se atribuir um título breve, não se utilizando traços internos horizontais ou verticais. As notas explicativas devem ser colocadas no rodapé das tabelas e não no cabeçalho ou título. Se houver tabela extraída de outro trabalho, previamente publicado, os autores devem solicitar formalmente autorização da revista que a publicou, para sua reprodução.
Para composição de uma tabela legível, o número máximo é de 10 colunas, dependendo da quantidade do conteúdo de cada casela. Notas em tabelas devem ser indicadas por letras e em sobrescrito.

Figuras
As ilustrações (fotografias, desenhos, gráficos, etc.) devem ser citadas como Figuras e numeradas consecutivamente com algarismos arábicos, na ordem em que foram citadas no texto e apresentadas após as tabelas. Devem conter título e legenda apresentados na parte inferior da figura. Só serão admitidas para publicação figuras suficientemente claras e com qualidade digital, preferentemente no formato vetorial. No formato JPEG, a resolução mínima deve ser de 300 dpi. Não se aceitam gráficos apresentados com as linhas de grade, e os elementos (barras, círculos) não podem apresentar volume (3-D). Se houver figura extraída de outro trabalho, previamente publicado, os autores devem solicitar autorização, por escrito, para sua reprodução.

***Checklist* para submissão**
1. Nome e instituição de afiliação de cada autor, incluindo e-mail e telefone.
2. Título do manuscrito, em português e inglês, com até 90 caracteres, incluindo os espaços entre as palavras.
3. Título resumido com 45 caracteres.
4. Texto apresentado em letras arial, corpo 12, em formato Word ou similar (doc, docx e rtf).
5. Resumos estruturados para trabalhos originais de pesquisa em dois idiomas, um deles obrigatoriamente em inglês.
6. Resumos narrativos para manuscritos que não são de pesquisa em dois idiomas, um deles obrigatoriamente em inglês.
7. Carta de Apresentação, constando a responsabilidade de autoria e conflito de interesses, assinada por todos os autores.
8. Nome da agência financiadora e número (s) do (s) processo (s).
9. Referências normalizadas segundo estilo Vancouver, ordenadas por ordem de citação, verificando se todas estão citadas no texto.
10. Tabelas numeradas sequencialmente, com título e notas, com no máximo 10 colunas.
11. Figura no formato vetorial ou em pdf, ou tif, ou jpeg ou bmp, com resolução mínima 300 dpi; em se tratando de gráficos, devem estar sem linhas de grade e sem volume.
12. Tabelas e figuras não devem exceder a cinco, no conjunto.

Processo Editorial

a) **Revisão da redação científica**

Para ser publicado, o manuscrito aprovado é submetido à revisão da redação científica, gramatical e de estilo. A RSP se reserva o direito de fazer alterações visando a uma perfeita comunicação aos leitores. O autor responsável terá acesso a todas as modificações sugeridas até a última prova enviada. Inclusive a versão em inglês do artigo terá esta etapa de revisão.

b) **Provas**

Após sua aprovação pelos editores, o manuscrito será revisado por uma equipe que fará a revisão da redação científica (clareza, brevidade, objetividade e solidez), gramatical e de estilo.

O autor responsável pela correspondência receberá uma prova, em arquivo de texto (doc, docx ou rtf), com as observações/alterações feitas pela equipe de leitura técnica. O prazo para a revisão da prova é de dois dias.

Caso ainda haja dúvidas nessa prova, a equipe editorial entrará em contato para revisão, até que se chegue a uma versão final do texto. Em seguida, o texto final passará por uma revisão gramatical. Após essa revisão o autor receberá nova prova, no formato final para publicação. Nessa última revisão podem ser feitas apenas correções de erros, pois não serão admitidos mais ajustes de forma. O prazo para a revisão da prova final é de um dia.

.tigos submetidos em português ou espanhol serão vertidos para o inglês. Aproximadamente uma semana após o autor ter finalizado a prova do artigo, a RSP enviará a versão em inglês do artigo para apreciação do autor. Nesta revisão, o autor deverá atentar para possíveis erros de interpretação, vocabulário da área e principalmente, equivalência de conteúdo com a versão "original aprovada". O prazo de revisão da versão em inglês é de dois dias.

A Revista adota o sistema de publicação continuada (*rolling pass*). Desta forma, a publicação do artigo se torna mais rápida: não depende de outros artigos para fechamento de um fascículo, mas do processo individual de cada artigo. Por isso, solicitamos o cumprimento dos prazos estipulados.

Taxa de Publicação

Embora as revistas recebam subvenções de instituições públicas, estas não são suficientes para sua manutenção. Assim, a cobrança de taxa de publicação passou a ser alternativa para garantir os recursos necessários para produção da RSP.

A USP garante os recursos básicos, mas não são suficientes. Assim, temos que contar com recursos complementares, além das agências de fomento.

A RSP em 2016 completa 50 anos de publicação e somente em 2012 iniciou a cobrança de taxa de artigos, fato este imperioso para garantir sua continuidade, sobretudo permitindo-lhe evoluir com tecnologias mais avançadas, mas que exigem também maior qualidade e recursos tecnológicos.

O valor cobrado é avaliado regularmente. Assim, para os artigos submetidos a partir de **janeiro de 2017**, o valor da taxa será de 2.200,00 para artigo original, revisão e comentário, e de 1.500,00 para comunicação breve.

A RSP fornecerá aos autores os documentos necessários para comprovar o pagamento da taxa, perante instituições empregadoras, programas de pós-graduação ou órgãos de fomento à pesquisa.

Suplementos

a) CARTA DE APRESENTAÇÃO

Cidade, _[dia]__ de Mês de Ano.

Prezado Sr. Editor, *Revista de Saúde Pública*

Submetemos à sua apreciação o trabalho "_____[título]_____", o qual se encaixa nas áreas de interesse da RSP. A revista foi escolhida [colocar justificativa da escolha da revista para a publicação do manuscrito].

O autor 1 participou da concepção, planejamento, análise, interpretação e redação do trabalho; e, o autor 2 participou na interpretação e redação do trabalho. Ambos os autores aprovaram a versão final encaminhada.

O trabalho está sendo submetido exclusivamente à RSP. Os autores não possuem conflitos de interesse ao presente trabalho. (Se houver conflito, especificar).

_____
nome completo do autor 1 + assinatura

_____
nome completo do autor 2 + assinatura

b) DECLARAÇÃO DE RESPONSABILIDADE

Eu, (nome por extenso), certifico que participei da autoria do manuscrito intitulado (título) nos seguintes termos:

"Certifico que participei suficientemente do trabalho para tornar pública minha responsabilidade pelo seu conteúdo. "

"Certifico que o manuscrito representa um trabalho original e que nem este manuscrito, em parte ou na íntegra, nem outro trabalho com conteúdo substancialmente similar, de minha autoria, foi publicado ou está sendo considerado para publicação em outra revista, quer seja no formato impresso ou no eletrônico, exceto o descrito em anexo. "

"Atesto que, se solicitado, fornecerei ou cooperarei totalmente na obtenção e fornecimento de dados sobre os quais o manuscrito está baseado, para exame dos editores. "

Contribuição:
_____

_____ _____
Local, data Assinatura

c) DECLARAÇÃO DE RESPONSABILIDADE PELOS AGRADECIMENTOS

Eu, (nome por extenso do autor responsável pela submissão), autor do manuscrito intitulado (título completo do artigo):

- Certifico que todas as pessoas que tenham contribuído substancialmente à realização deste manuscrito, mas não preencheram os critérios de autoria, estão nomeados com suas contribuições específicas em Agradecimentos no manuscrito.
- Certifico que todas as pessoas mencionadas nos Agradecimentos forneceram a respectiva permissão por escrito.

_____/_____/_____
_____
DATA NOME COMPLETO E ASSINATURA

d) DECLARAÇÃO DE TRANSFERÊNCIA DE DIREITOS AUTORAIS

Concordo que os direitos autorais referentes ao manuscrito [TÍTULO], aprovado para publicação na Revista de Saúde Pública, serão propriedade exclusiva da Faculdade de Saúde Pública, sendo possível sua reprodução, total ou parcial, em qualquer outro meio de divulgação, impresso ou eletrônico, desde que citada a fonte, conferindo os devidos créditos à Revista de Saúde Pública.

Autores:

_____	_____
Local, data	NOME COMPLETO + Assinatura
_____	_____
Local, data	NOME COMPLETO + Assinatura

[Home] [Sobre a revista] [Corpo editorial] [Assinaturas]

Todo o conteúdo do periódico, exceto onde está identificado, está licenciado sob uma Licença Creative Commons

Avenida Dr. Arnaldo, 715
01246-904 São Paulo SP Brasil
Tel./Fax: +55 11 3061-7985.

e-Mail
revsp@usp.br