Universidade de São Paulo
Faculdade de Medicina de Ribeirão Preto

**SÉRIES EPIDEMIOLÓGICAS NA PRESENÇA DE PONTOS DE MUDANÇA SOB
UM ENFOQUE BAYESIANO**

Charles Chen

Ribeirão Preto

2021

**SÉRIES EPIDEMIOLÓGICAS NA PRESENÇA DE PONTOS DE MUDANÇA SOB
UM ENFOQUE BAYESIANO**

Charles Chen

Área de Concentração: Saúde Pública

Orientador: Prof. Dr. Jorge Alberto Achcar

Dissertação de Mestrado apresentada ao Programa Saúde na Comunidade da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, para obtenção do título de mestre, como parte dos requisitos para obtenção do título de Mestre em Saúde Pública.

Ribeirão Preto

2021

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE
TRABALHO, POR QUALQUER CONVENCIONAL OU ELETRÔNICO, PARA FINS
DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE

FICHA CATALOGRÁFICA

Chen, Charles.

Séries epidemiológicas na presença de pontos de mudança sob um
enfoque Bayesiano, 2020.

Dissertação de Mestrado apresentada à Faculdade de Medicina de Ribeirão
Preto, Universidade de São Paulo no Programa Saúde na Comunidade

Orientador: Prof. Dr. Jorge Alberto Achcar

1. Pontos de mudanças. 2. Modelos de regressão. 3. Séries
epidemiológicas, métodos Bayesianos

FOLHA DE APROVAÇÃO

Chen, Charles. Séries epidemiológicas na presença de pontos de mudança sob um enfoque Bayesiano.

Dissertação de Mestrado apresentada ao Programa Saúde Pública da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, para obtenção do título de mestre.

Aprovado em: ____/____/____

BANCA EXAMINADORA

Prof. Dr. _____ INSTITUIÇÃO: _____

Julgamento _____ ASSINATURA: _____

Prof. Dr. _____ INSTITUIÇÃO: _____

Julgamento _____ ASSINATURA: _____

Prof. Dr. _____ INSTITUIÇÃO: _____

Julgamento _____ ASSINATURA: _____

A minha família e amigos

AGRADECIMENTOS

Aos Prof. Drº Jorge Alberto Achcar, Prof. Drº Edson Zangiacomi Martinez, Drº Ricardo Puziol de Oliveira e Drª Tatiana Reis Icuma por todo o apoio e ensinamentos.

Aos professores do curso de Pós Graduação em saúde da comunidade da FMRP/USP pelo empenho e dedicação prestados a ciência.

Aos funcionários da secretaria que me auxiliaram em todos os processos de cunho burocrático no decorrer do programa.

À Capes que apoiou financeiramente o projeto para que a dedicação em tempo integral fosse possível.

Aos meus pais que dedicaram a vida para que esse momento se tornasse realidade.

A todos que de certa forma me apoiaram a encarar esse desafio e estiveram presentes mesmo nos momentos mais nebulosos.

Um muito obrigado de coração.

RESUMO

Chen, Charles. **Séries epidemiológicas na presença de pontos de mudança sob um enfoque Bayesiano**. 2021. 84 páginas. Dissertação de mestrado – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2021.

Resumo: Neste estudo é introduzida uma abordagem Bayesiana para analisar dados de séries temporais de contagem na presença de um ou mais pontos de mudança. Esta situação é muito comum em muitas áreas de aplicação, especialmente considerando séries temporais de contagem de epidemiologia. Quando a contagem observada $n(t)$ os dados são diferentes de zero para cada tempo t , $t = 1, 2, \dots, N$ (número de vezes observadas), é proposto um modelo estatístico assumindo distribuições normais para o logaritmo dos dados transformados, ou seja, $Y(t) = \log[n(t)]$ na presença de um ou mais pontos de mudança. Em situações na presença de dados de contagem iguais a zero em momentos diferentes (ou seja, presença de contagem zero), o modelo estatístico baseado na transformação de logaritmo para os dados de contagem não é adequado na análise de dados. Para este caso, é proposto o uso de processos de Poisson não homogêneos (NHPP) assumindo uma modelagem PLP (power law process) para sua função intensidade na presença de pontos de mudança. Alguns exemplos com conjuntos de dados reais são apresentados para ilustrar a metodologia proposta.

Palavras-chave: séries temporais de contagem, modelo linear na presença de pontos de mudança, processo de Poisson não homogêneo na presença de pontos de mudança, processo de lei de potência, abordagem bayesiana.

ABSTRACT

Chen, Charles. Epidemiological series in the presence of points of change under a Bayesian approach. 2021. 84 pages. Master's Dissertation - Ribeirão Preto, Medical School, University of São Paulo, Ribeirão Preto, 2020

Abstract: In this study it is introduced a Bayesian approach to analyze count time series data in presence of one or more change-points. This situation is very common in many areas of application, especially considering epidemiology count time series. When the observed count $n(t)$ data are different of zero for each time t , $t = 1, 2, \dots, N$ (number of observed times), it is proposed a statistical model assuming normal distributions for the logarithm of the transformed data, that is, $Y(t) = \log[n(t)]$ in presence of one or more change-points. In situations in presence of count data equals to zero at different times (that is, presence of zero count) the statistical model based on the logarithm transformation for the count data is not suitable in the data analysis. For this case, it is proposed the use of non-homogeneous Poisson processes (NHPP) assuming a PLP (power law process) modeling for its intensity function in presence of change-points. Some examples with real data sets are presented to illustrate the proposed methodology.

Keywords: count time series, linear model in presence of change-points, non-homogeneous Poisson process in presence of change-points, power law process, Bayesian approach.

Índice

1. Introdução	11
1.1 O sarampo no Brasil	15
1.2 Objetivos	18
1.3 Introdução aos modelos de regressão lineares tradicionais	19
1.4 Introdução aos Processos de Poisson não Homogêneos	20
2. Introdução aos métodos Bayesianos	22
2.1 Probabilidade condicional e a fórmula de Bayes	24
2.2 Distribuições a Priori	25
2.3 Métodos de simulação para distribuições <i>a posteriori</i>	26
2.3.1 Amostrador de Gibbs	27
2.3.2 Algoritmo de Metropolis-Hastings	27
3. Modelagem estatística	29
3.1 Modelos de Regressão com pontos de mudança	29
3.2 Modelos de Regressão com efeitos defasados	29
3.3 Processos de Poisson Não Homogêneos na presença de k pontos de mudança	30
3.3.1 Função de Intensidade	30
3.3.2 Função de Verossimilhança	31
3.3.3 Inferência Bayesiana	32
4. Aplicações em dados reais	33
4.1 Um ponto de mudança	33
4.2 Dois pontos de mudança	36
4.3 Cinco pontos de mudança	41
4.4 NHPP na presença de vários pontos de mudança	44
Agradecimento	49
Referências Bibliográficas	50
Apêndice 1 – Medidas de Ajuste	54
• Mean average percent error (MAPE)	54
• Mean Absolute Deviation(MAD)	55
• Mean Square Deviation (MSD)	55
Apêndice 2 – Códigos de Programação OpenBugs	56

Capítulo 1

1. Introdução

A presença de um ou mais pontos de mudança são muito comuns em dados de séries temporais principalmente nas áreas de epidemiologia, economia, meio ambiente, monitoramento de condições clínicas, mudanças climáticas, análise de narrativa e imagens, entre diversas outras (Yamanashi e Takeuchi, 2002; Lavielle e Teyssiere, 2006; Chopin, 2007; Achcar et al. 2010, 2011; Liu et al., 2013; Jandhyala et al., 2013; Aminikhanghahi e Cook, 2017). Em diversos estudos podem ser observados a influência de alguns eventos que interferem abruptamente no comportamento dos dados, podendo desta forma, resultar em séries com um ou mais pontos de mudanças. Essas ocorrências podem ocorrer tanto com variáveis resposta tipo contínua, tipo contagem e outros tipos os quais podem assumir diferentes modelos paramétricos ou não paramétricos. Em se tratando de análises estatísticas, a identificação de pontos de mudança é de grande interesse para os pesquisadores considerando diferentes modelagens para predições de valores futuros de séries temporais, especialmente quando aplicados à dados de contagem frequentemente adotando o uso de uma distribuição Poisson padrão.

Nos casos específicos onde os dados de contagem são diferentes de zero para todos os tempos observados (dias, meses, anos) é possível assumir uma distribuição normal para os dados utilizando os dados transformados para uma escala logarítmica ou uma transformação mais geral (ver por exemplo, Box e Cox, 1964), isto é, $Y(t) = \log(n(t))$, onde $n(t)$ denota a contagem em um tempo t , e um modelo de regressão linear padrão assumindo erros com distribuição normal na presença de pontos de mudança pode ser adotado. Na prática, pode existir situações em que ocorre a presença de zeros em determinados tempos (presença de contagem igual a zero) impossibilitando a utilização da transformação logarítmica, sendo necessário a adoção de outras abordagens estatísticas para realização das análises das séries temporais contendo um ou mais pontos de mudança. Nesses casos, uma das possíveis alternativas seria a modelagem através dos Processos de Poisson Homogêneos (HPP) ou Processos de Poisson Não Homogêneos (NHPP) com formas paramétricas específicas para sua função de intensidade.

Considerando dados de contagem modelados por Processos de Poisson Homogêneos (HPP) e Processos de Poisson Não Homogêneos (NHPP) existem diversos estudos na literatura expondo inferências pontuais e/ou intervalares para pontos de mudanças, principalmente se utilizando de uma abordagem bayesiana (Kuo e Yang, 1996; Achcar et al., 1998; Cid e Achcar, 1999; Achcar, 2001; Pievatolo e Ruggeri, 2004). O emprego de métodos Bayesianos utilizando procedimentos de simulação MCMC (Markov Chain Monte Carlo) para simular amostras das distribuições a posteriori conjuntas de interesse (Gelfand e Smith, 1990; Smith e Roberts, 1993; Chib e Greenberg, 1995) de forma geral simplificam significativamente os processos de inferência sobre os parâmetros dos modelos dada a complexidade das equações de verossimilhança na presença de pontos de mudança quando comparado com outras abordagens, por exemplo, utilizando o método de estimação da máxima verossimilhança tradicional (MLE-maximum likelihood estimators ou estimadores de máxima verossimilhança). Modelos de contagem como os processos HPP e NHPP por diversas vezes já foram utilizados para na obtenção de inferências sobre parâmetros relacionados à pontos de mudança (Matthews e Farewell, 1982; Achcar e Bolfarine, 1989; Carlin et al., 1992; Dey e Purkayastha, 1997; Achcar e Loibel, 1998). Especificamente, Raftery e Akman (1986) apresentam uma abordagem Bayesiana para Processos de Poisson Homogêneo (HPP) na presença de um ponto de mudança. Ruggeri e Sivaganesan (2005) introduzem uma abordagem Bayesiana para pontos de mudanças em Processos de Poisson não Homogêneos (NHPP) adotando a lei de potência (PLP) como forma paramétrica para a função de intensidade.

Nesse estudo, será apresentado sob um enfoque Bayesiano a análise de dados de séries temporais transformados para uma escala logarítmica na presença de um ou mais pontos de mudança considerando uma distribuição Normal para os erros em modelos de regressão em situações com ausência de zeros nos registros de contagem, e a utilização de modelos baseados em processos NHPP como alternativa para séries temporais com presença de registros de contagem zero, uma situação muito comum em dados epidemiológicos uma vez que doenças consideradas erradicadas podem vir a apresentar novos registros ao longo do tempo.

No primeiro estudo, onde os dados não possuem registros de contagem zero, a variável de interesse (dado discreto) será transformada para uma escala logarítmica, desta forma possibilitando o ajuste de um modelo de regressão linear com erros ϵ independentes

e identicamente distribuídos com distribuição normal com média zero e variância σ^2 ($\varepsilon \sim N(0, \sigma^2)$) na presença de pontos de mudança considerando também efeitos defasados da variável resposta em conjunto com a covariável tempo. No segundo estudo, onde os dados possuem registros de contagem zero, adotaremos um modelo baseado em um Processo de Poisson não Homogêneo na presença de um ou mais pontos de mudança.

Como primeira motivação para o presente estudo, a Figura 1 introduz um conjunto de dados (série histórica mensal da contagem de visitantes na Nova Zelândia motivados por um estudo relacionado ao período de agosto de 1989 a maio de 2012). Nesta aplicação, todos os registros de contagem nos diferentes meses são diferentes de zero, desta forma é possível considerar um modelo estatístico com a transformação logarítmica dos dados. Na Figura 1 (dados na escala original e escala logarítmica), é possível verificar a presença de um ponto de mudança próximo ao 35º mês para um total de 154 meses.

Como uma segunda motivação para o presente estudo, na Figura 2 são apresentados dados relacionados a incidência (casos notificados) mensal de sarampo no Brasil no período de janeiro de 2012 a agosto de 2018 (ver também, Tabela 1). No caso de doenças que foram epidemicamente controladas e consideradas erradicadas como é o caso do sarampo, no decorrer dos anos é possível aparecer novos casos ou até novos surtos. Sendo assim, é comum as séries conterem registros de zeros em determinadas posições ou intervalos de tempo (Tabela 1). Neste caso em específico não é possível realizar uma transformação logarítmica da variável resposta de interesse sendo necessário adotar outra abordagem para os dados.

Figura 1 – Contagem mensal de visitantes na Nova Zelândia de Agosto de 1989 a Maio de 2012.

Figura 2 – Incidência de Sarampo no Brasil no período de janeiro 2012 a agosto 2018.

O objetivo principal para uma primeira aplicação com respostas transformadas para a escala logarítmica é referente à obtenção de inferências sobre os parâmetros do modelo de regressão linear com erros normais na presença de um ou mais pontos de mudança. Podemos destacar em especial a estimação do ponto de mudança, este considerado como um parâmetro desconhecido que será estimado a partir dos dados. O processo inferencial será realizado sob um enfoque Bayesiano e usando métodos de simulação MCMC (Monte Carlo em Cadeias de Markov) que serão executados para se obter inferências (médias, medianas, desvios-padrões, intervalos de credibilidade) que sumarizam os resultados das distribuições a posteriori de interesse. Para outras aplicações com presença de contagem igual a zero, são propostos modelos NHPP com função de intensidade PLP na presença de um ou mais pontos de mudança para análise dos dados onde também adotaremos como metodologia para a estimação dos parâmetros do modelo a inferência Bayesiana com o auxílio das simulações de MCMC.

1.1 O sarampo no Brasil

O ministério da saúde do Brasil em seu guia de bolso sobre doenças infecciosas e parasitárias descreve o sarampo como uma doença infecciosa aguda, de natureza viral, transmissível e altamente contagiosa. A transmissão se dá através do contato com secreções de indivíduos infectados pelo vírus, onde muitas vezes o contágio ocorre pelas vias aéreas não havendo necessidade de contato físico; o único modo de prevenção da doença é através da vacinação. Quando analisamos as complicações clínicas graves que o sarampo pode vir acarretar ao indivíduo e a sua rápida capacidade de proliferação prontamente se conclui que se trata de uma enfermidade que merece um acompanhamento contínuo.

Tabela 1 – Incidência mensal de sarampo no Brasil de Janeiro, 2012 a Agosto, 2018.

Data-base	N	Data-base	N	Data-base	N	Data-base	N
jan/2012	1	set/2013	8	mai/2015	26	jan/2017	0
fev/2012	0	out/2013	9	jun/2015	9	fev/2017	0
mar/2012	0	nov/2013	28	jul/2015	1	mar/2017	0
abr/2012	0	dez/2013	27	ago/2015	1	abr/2017	0
mai/2012	0	jan/2014	79	set/2015	0	mai/2017	0
jun/2012	0	fev/2014	89	out/2015	0	jun/2017	0
jul/2012	0	mar/2014	30	nov/2015	0	jul/2017	0
ago/2012	0	abr/2014	25	dez/2015	0	ago/2017	0
set/2012	0	mai/2014	15	jan/2016	0	set/2017	0
out/2012	0	jun/2014	34	fev/2016	0	out/2017	0
nov/2012	0	jul/2014	93	mar/2016	0	nov/2017	0
dez/2012	1	ago/2014	106	abr/2016	0	dez/2017	0
jan/2013	4	set/2014	69	mai/2016	0	jan/2018	0
fev/2013	0	out/2014	67	jun/2016	0	fev/2018	8
mar/2013	5	nov/2014	60	jul/2016	0	mar/2018	74
abr/2013	16	dez/2014	41	ago/2016	0	abr/2018	130
mai/2013	37	jan/2015	60	set/2016	0	mai/2018	282
jun/2013	25	fev/2015	41	out/2016	0	jun/2018	184
jul/2013	20	mar/2015	37	nov/2016	0	jul/2018	69
ago/2013	13	abr/2015	39	dez/2016	0	ago/2018	0

A Figura 3 mostra que nas décadas de 1970-1980, o Brasil sofria com altas taxas de incidência de sarampo, doença que se tornou de notificação compulsória a partir do ano 1968.

No ano de 1986, houve a pior epidemia de sarampo da história brasileira, onde houve uma incidência de 129.942 casos, representando uma incidência de 97,7 por 100.000 habitantes (Santos E. D. et al. ,1997). Nessa época houve diversas campanhas para controle da doença, mas foi em 1992 com o Plano de Eliminação do Sarampo que houve uma efetiva redução drástica no número de casos.

Figura 3 – Gráfico de incidência de sarampo no Brasil e Estratégias de Controle e Incidência do Sarampo (Ministério da Saúde).

Após diversas campanhas de imunização em caráter nacional foi obtido êxito no controle da doença, tanto que em 2016 o Brasil recebeu da Organização Pan Americana de Saúde (OPAS/OMS) o certificado de erradicação do sarampo. Porém em 2018, devido ao grande número de imigrantes oriundos da Venezuela, país que já enfrentava problemas com a doença, em conjunto com a deficiência da imunização no Brasil o vírus foi reintroduzido em circulação no país. Os casos de contágio no Brasil em sua grande maioria são importados, casos autóctones são muito raros, porém o ministério da saúde informa que possui suspeitas de alguns casos que estão em análise.

Devido a incidência prolongada por mais de 12 meses em território nacional, critério esse estabelecido pela OPAS/OMS, o Brasil teve seu certificado revogado. Em resposta, o país está tomando as devidas providências para reestabelecer o controle sobre a doença, através de campanhas nacionais de vacinação, aumentando a imunização e estabelecer novamente controle sobre a proliferação da doença.

A volta da circulação do sarampo não é uma exclusividade brasileira, diversos países no mundo todo apresentaram nos últimos anos problemas relacionados à reincidência ou agravamento de transmissão da doença em suas fronteiras. Podemos citar alguns países que tiveram problemas similares aos do Brasil, como República Democrática do Congo em 2019 (Ducomble T, Gignoux E ,2020), Haiti em 2000 (Venczel et al., 2003), Holanda em 2013 (Van Dam ASG et al. 2020), China em 2014 (Ma C et al., 2019) e Ucrânia e Filipinas 2019 (OMS, 2020).

Os principais fatores para o retorno do sarampo em todos os países citados se dão quase que pelos mesmos motivos. Em alguns países o sistema de saúde pública se torna ineficiente, o acolhimento de imigrantes sobrecarrega a sua rede de atendimento, e em outros países, problemas internos como má administração na área de saúde e corrupção são fatores de retorno da epidemia.

Alguns grupos possuem restrições à vacinação devido a crenças religiosas ou desinformação. Dada a conectividade da nova era com muita informação virtual circulando quase que instantaneamente, sendo ela verídica ou infundada, tendenciosa ou imparcial como reportado por Arif et al. (2018) que mostra como palavras pesquisas em diversos sítios e locais podem gerar informações diferentes. Com a internet pode se disseminar rapidamente diversas informações falsas (fake news), como teorias infundadas vinculando vacinas ao desenvolvimento de doenças. Uma dessas fontes de desinformação bem conhecida do público, é um estudo publicado por Wakefield et al. (1998), que estabelecia suposta relação entre a vacina tríplice viral e a incidência de autismo. Esse estudo foi recolhido no ano de 2010 pela revista onde foi publicado o artigo por falta de comprovação científica.

Outro fato que colabora com a disseminação do vírus é o aumento do fluxo internacional de pessoas (Sarkar et al., 2019). Com o aumento do comércio e turismo a circulação internacional de pessoas tem aumentado consideravelmente nos últimos anos,

devido a meios de transporte mais rápidos e com maior capacidade que possibilitam rápidos deslocamentos de pessoas em grandes distancias a custos bem menores do que existia a décadas atrás; esse aumento de fluxo combinado com a alta capacidade de proliferação do vírus forma um ambiente propício para a disseminação da doença. Isso também é observado para outros vírus, como o caso recente do corona-vírus levando à Covid-19, que no ano de 2020 teve uma disseminação muito rápida levando a uma pandemia com altas incidências ainda ocorrendo em praticamente todos os países do mundo.

Os estudos indicam que o caminho mais efetivo é o da imunização, e levando em consideração os fatores expostos acima a doença só consegue se instalar novamente se houver baixa adesão da população à imunização. (Cherry e Zahn, 2018) indicam que indivíduos vacinados tendem a desenvolver sintomas menos graves quando comparados aos indivíduos não vacinados, e mesmo com a vacinação pode ser necessário mais de uma dose para manter os níveis de anticorpos. Os países que possuem problemas estruturais e administrativos possuem maiores dificuldades para implantar um projeto de vacinação efetivo. Vacinas são componentes dependentes de tecnologia, na ausência de polo tecnológico para fabricação no próprio país a aquisição deverá ser realizada por vias internacionais, elevando seu custo. Somente adquirir não é suficiente, deve-se providenciar local adequado para seu armazenamento, unidades equipadas para seu transporte e profissionais treinados para sua aplicação (Venczel et al., 2003; Kuroiwa et al., 2003).

1.2 Objetivos

- **Objetivo Geral**

O objetivo do estudo é propor a utilização de um modelo baseado em processos de Poisson não homogêneos com função intensidade PLP (Power Law Process ou processo de lei de potências) na presença de um ou mais pontos de mudança em séries epidemiológicas contendo registros de zero com particular destaque para a série epidemiológica de casos de sarampo no Brasil.

- **Objetivo específico**

Explorar a aderência do modelo proposto aplicado à série de dados epidemiológicos de sarampo do Brasil no período de janeiro 2012 a agosto 2018.

1.3 Introdução aos modelos de regressão lineares tradicionais

Modelos de regressão lineares são utilizados como uma das técnicas estatísticas mais usadas na análise de dados para modelar a relação entre uma variável resposta contínua Y (ou variável dependente) com uma ou mais covariáveis ou fatores também conhecidos como variáveis independentes ou variáveis explicativas, denotadas por $X_1, X_2, \dots, X_p$ que podem ser contínuas ou categóricas. No caso de apenas uma variável explicativa o modelo é usualmente chamado como um modelo de regressão linear simples. Caso o modelo possua mais de uma variável explicativa, ele é conhecido como regressão linear múltipla.

A importância dos modelos de regressão linear se dá principalmente pela sua capacidade de predição e na quantificação de relações entre variáveis. No caso da predição, utilizando-se de duas variáveis de dados observados, uma resposta denotada por Y e uma variável explicativa X , é possível ajustar um modelo capaz de prever valores de Y dado o valor da covariável X .

Na construção de modelos de regressão relacionando uma variável dependente Y com uma covariável X ou com um vetor de covariáveis $X_1, \dots, X_p$, é possível fazer inferências sobre a existência ou não de efeitos significativos de cada covariável específica na variável dependente Y , a partir de testes de hipóteses sobre os parâmetros de regressão do modelo de regressão ajustado. Isso comumente é feito assumindo testes de hipóteses com uma hipótese de nulidade H_0 dada pelo parâmetro de regressão correspondente ser igual à zero versus uma hipótese alternativa H_1 , considerando o parâmetro de regressão correspondente como diferente de zero.

Os modelos de regressão linear são frequentemente ajustados pelo método dos mínimos quadrados, porém para algumas situações específicas é possível utilizar o método da máxima verossimilhança (Draper e Smith, 1981). Na análise de regressão linear múltipla verifica-se o efeito conjunto das covariáveis $X_1, \dots, X_p$, na resposta Y .

Um modelo de regressão linear múltipla é dado por,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i, \quad (1)$$

onde, $\beta_0, \beta_1, \dots, \beta_p$, são parâmetros de regressão do modelo; $X_{i1}, \dots, X_{ip}$, são covariáveis com valores conhecidos; ε_i , são erros (variáveis aleatórias não observadas) assumidas com uma distribuição $N(0, \sigma^2)$; $i = 1, \dots, n$ onde n é o tamanho amostral.

Os estimadores dos parâmetros de regressão $\beta_0, \beta_1, \dots, \beta_p$, usualmente são estimados pelo método de mínimos quadrados (ver por exemplo, Draper e Smith, 1981).

1.4 Introdução aos Processos de Poisson não Homogêneos

Imagine a situação em um posto de saúde onde diversas pessoas chegam em busca de atendimento ao longo do dia, esta é uma clara situação de alocação contingencial no tempo. Sendo assim, podemos considerar um Processo de Poisson como uma boa abordagem para modelagem do problema.

Porém, quando analisamos mais profundamente a situação, pelo fato da procura por atendimento não ser uniformemente distribuído ao longo do expediente (não estacionaridade), em certos horários podem haver grande demanda (picos) e em outros horários nenhuma demanda. Essa peculiaridade deve ser considerada em nosso modelo, desta forma, o mais adequado seria propor uma abordagem através de um Processo de Poisson Não Homogêneo (NHPP – non-homogeneous Poisson process), visando dar versatilidade através da escolha de diferentes formas paramétricas para a função intensidade. Essa característica torna o modelo propício para descrever eventos “atípicos” onde a taxa de ocorrência se desenvolve ao longo do tempo (Insua et. al., 2012).

Utilizando o caso dos dados de sarampo no Brasil, seja $N = \{N_t : t \in \{0, T\}\}$, o processo que descreve o número de incidências acumuladas observadas no intervalo $[0, T]$ e sua função de intensidade dada por:

$$\lambda(t) = \frac{d}{dt} m(t) = \frac{d}{dt} E(N_t), \quad (2)$$

onde $m(t)$ é a função de valor médio.

Diversas formas paramétricas podem ser consideradas para a função intensidade $\lambda(t)$, que pode ser crescente, decrescente, forma de curva de banheira, unimodal, entre outros (Musa e Okamoto, 1984 e/ou Mudholkar et al., 1995).

Capítulo 2

2. Introdução aos métodos Bayesianos

Um dos principais objetivos da estatística é extrair informações sobre determinada população em estudo, tecnicamente falando, o objetivo principal é encontrar um estimador do valor de um parâmetro (usualmente conhecido como θ) relacionado com a distribuição dos dados, que forneça alguma noção sobre os dados observados. Esse parâmetro pode ter interpretações diretas como sendo média, mediana, variância ou desvio-padrão, mas também pode ser relacionado com a possível forma matemática de uma distribuição paramétrica ajustada aos dados sem uma interpretação direta. O desconhecimento sobre o valor real do parâmetro se torna presente nesses casos, podendo incidir sobre θ magnitudes de incertezas de graus variados. Sob o enfoque Bayesiano, estes diferentes graus de incerteza são representados por distribuições de probabilidades nos dados e nos parâmetros. Segundo (Paulino et al., 2003), o paradigma Bayesiano expõe que o que é desconhecido, no caso o parâmetro, é incerto e toda incerteza deve ser quantificada em termos probabilísticos. Sendo assim, é natural que os graus de conhecimento e incertezas a respeito de θ variem de pesquisador para pesquisador. Este fato faz com que além das quantidades observáveis, também os parâmetros de um modelo estatístico sejam considerados quantidades aleatórias.

Diferentemente da Inferência Clássica, a Inferência Bayesiana não se utiliza apenas da amostra para obter informação sobre o parâmetro. Apesar de parte dessa informação provir da amostra, representada pela função de verossimilhança, outra parte se deve ao conhecimento prévio do pesquisador, que se dá através de uma distribuição de probabilidade para o parâmetro, conhecida como distribuição *a priori*.

A estimação do parâmetro é realizada com a utilização conjunta dessas duas informações, através da fórmula de Bayes, e tem como resultado a distribuição *a posteriori* do parâmetro que representa todo o conhecimento existente sobre o problema investigado. A distribuição *a posteriori* é o principal elemento de toda análise bayesiana de dados, pois é ela que possui toda a informação necessária para a realização de inferências, permitindo a análise de qualquer aspecto do parâmetro em estudo, como

média, mediana, percentis, intervalos de credibilidade, entre outros (Carlin e Louis, 2003).

O surgimento e disponibilidade de vários algoritmos computacionais para simulação de amostras da distribuição a posteriori conjunta de interesse (simulação de Monte Carlo simples, simulação de Monte Carlo por importância, aproximações numéricas de Laplace, simulação MCMC) em geral implantados em variados *softwares* estatísticos livres existentes, como por exemplo, o software R (R Development Core Team, 2010) e o software OpenBugs (Spiegelhalter et al., 2003), facilitam o uso da metodologia Bayesiana na análise de dados em praticamente todas as áreas de aplicação, como tem sido observado nas últimas décadas.

Além da possibilidade de incorporação de informações prévias sobre parâmetros do modelo proposto fornecidas por experts da área de aplicação (distribuição *a priori*), a inferência Bayesiana possui a vantagem de ser bem mais flexível quanto ao ajuste de modelos muitas vezes bem complexos com a presença de grande número de parâmetros, sem a dependência de métodos assintóticos necessitando de grandes amostras dos dados (muitas vezes não disponíveis) usualmente considerados na inferência clássica como é usado com o método de máxima verossimilhança. Além disso, a metodologia Bayesiana proporciona maior facilidade de interpretação nas inferências obtidas usando intervalos de credibilidade para os parâmetros do modelo, obtidos com grande precisão, especialmente usando métodos de simulação como o método MCMC para obter amostras da distribuição a posteriori, sendo isso feito diretamente dos quantis da distribuição *a posteriori* (Paulino et al., 2003). Outro ponto importante: sob a metodologia Bayesiana usualmente não é necessário o uso de testes de hipóteses (valor-p) para as inferências de interesse o que é comum sob a abordagem clássica. Uma outra vantagem a ser mencionada sob o enfoque Bayesiano é relacionada à comparação de modelos onde é possível comparar o ajuste dos modelos propostos de forma simples e direta usando densidades preditivas, uma metodologia não existente sob o enfoque clássico.

2.1 Probabilidade condicional e a fórmula de Bayes

Considere uma partição do espaço amostral Ω , onde os eventos $A_1, A_2, \dots, A_k$ formam uma sequência de eventos mutuamente exclusivos, isto é, $\bigcup_{j=1}^k A_j = \Omega$ e $A_i \cap A_j = \emptyset$ (conjunto vazio) para $i \neq j$ tal que $P\left(\bigcup_{j=1}^k A_j\right) = \sum_{j=1}^k P(A_j) = 1$.

Sendo assim para qualquer outro evento B ($B \subset \Omega$), temos,

$$P(A_i|B) = \frac{P(A_i|B)P(A_i)}{\sum_{j=1}^k P(A_j|B)P(A_j)}, \text{ para todo } i \text{ variando de } 1 \text{ a } k. \quad (3)$$

Assumir agora um vetor de dados $y = (y_1, \dots, y_n)'$ e θ (quantidade desconhecida que pode ser um vetor de parâmetros) os parâmetros de uma distribuição de probabilidade associada com a variável aleatória Y_i com valores observados em $y_i, i = 1, \dots, n$.

Considere $y = (y_1, \dots, y_n)$ uma amostra aleatória onde os dados são independentes e identicamente distribuídos. Suponha que y é um vetor de observações de uma distribuição conjunta dada por $f(y|\theta)$ e seja $\pi(\theta)$ uma distribuição a priori para θ . Para θ assumindo valores discretos $\theta_1, \dots, \theta_k$, temos de (3) a distribuição a posteriori para θ_i dado y dada por:

$$\pi(\theta_i|y) = \frac{f(y|\theta_i)\pi(\theta_i)}{\sum_{j=1}^k f(y|\theta_j)\pi(\theta_j)} \quad (4)$$

onde o parâmetro θ também é conhecido como uma quantidade aleatória, sob o enfoque bayesiano.

Para θ assumindo valores contínuos num dado intervalo, podemos escrever (4) como:

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \quad (5)$$

onde a integral no denominador de (5) é definida no intervalo de variação de θ .

2.2 Distribuições a Priori

Na análise bayesiana, a distribuição *a priori* é utilizada a fim de representar o que já é conhecido sobre parâmetros desconhecidos, antes de se avaliar os dados. Deve-se ter muita cautela ao definir uma distribuição *a priori*, afinal se esta informação não for bem definida pode-se chegar a interpretações errôneas. Uma distribuição *a priori* para um parâmetro pode se dar de várias formas, sendo possível ocorrerem a partir de procedimentos subjetivos ou objetivos.

A distribuição *a priori* conjugada é uma *priori* informativa, onde a distribuição *a priori* e *a posteriori* pertencem à mesma classe de distribuições. A passagem de *priori* para *posteriori* envolve apenas uma simples mudança nos parâmetros, sem a necessidade de cálculos adicionais (Paulino et al., 2003).

A distribuição *a priori* também pode ser a incorporação do conhecimento de um pesquisador (expert), ou seja, o pesquisador se baseia na prática e no seu “*feeling*” para definir uma *priori*. Em muitas situações práticas esse conhecimento do especialista não existe ou, se existe, não é fidedigno. Nesse caso, caracteriza-se uma “ignorância *a priori*” (Paulino et al., 2003).

Nestas situações são utilizadas distribuições *a priori* não informativas, ou seja, *prioris* de referência “neutras” (Box e Tiao, 1973). A utilização deste tipo de distribuição *a priori* permite a comparação com os resultados obtidos pela inferência clássica, haja visto que através de uma *priori* não informativa, o modelo é baseado apenas na informação dos dados amostrais. Existem vários métodos para se definir *a priori* não informativa, podemos citar alguns exemplos como o Método de Bayes-Laplace, Método de Jeffreys, entre outros (Box e Tiao, 1973; Paulino et. al., 2003).

2.3 Métodos de simulação para distribuições *a posteriori*

Para a obtenção de sumários inferenciais de interesse das distribuições *a posteriori*, geralmente é necessário resolver integrais bayesianas que não apresentam solução analítica. Na prática, observa-se que os modelos utilizados nem sempre são simples para se obter os sumários *a posteriori*. Mesmo que se tenha uma *priori* e uma verossimilhança simples, a junção delas pode produzir uma distribuição *a posteriori* muito complicada (Paulino et. al., 2003).

Os métodos com base em simulação de amostras da distribuição *a posteriori* de interesse, como, por exemplo, o método de Monte Carlo com Cadeias de Markov (MCMC), passaram a ser muito utilizados com o avanço das técnicas computacionais. Esse método consiste na simulação de uma variável aleatória através de uma cadeia de Markov, no qual a sua distribuição assintoticamente se aproxima da distribuição *a posteriori* (Bernardo e Smith, 1994).

A cadeia de Markov é um processo estocástico no qual o próximo estado da cadeia depende somente do estado atual e dos dados. No entanto, existe uma relação com o estado inicial, que é descartado após um período de aquecimento, o chamado período “Burn in”.

As formas mais usuais dos métodos MCMC são os amostradores de Gibbs e o algoritmo de Metropolis-Hastings. Estas duas formas simulam amostras da distribuição *a posteriori* conjunta a partir das distribuições condicionais *a posteriori* para cada parâmetro (Gelfand e Smith, 1990; Chib e Greenberg, 1995).

O amostrador de Gibbs nos permite gerar amostras da distribuição *a posteriori* conjunta desde que as distribuições condicionais completas possuam formas fechadas ou conhecidas. Por outro lado, o algoritmo de Metropolis-Hastings permite gerar amostras de distribuições *a posteriori* conjunta com distribuições condicionais completas possuindo ou não uma forma conhecida ou fechada.

2.3.1 Amostrador de Gibbs

Suponha $\pi(\theta|y)$ uma distribuição *a posteriori* conjunta, sendo $\theta = (\theta_1, \dots, \theta_k)$, no qual desejamos obter inferências. Para isso, simulam-se quantidades aleatórias de distribuições condicionais completas $\pi(\theta_i|y, \theta_{(i)})$.

Considere os valores iniciais (arbitrários) para $\theta^{(0)} = \theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)}$. Desta forma, segue o seguinte algoritmo:

- Gerar $\theta_1^{(1)}$ de $\pi(\theta_1|y, \theta_2^{(0)}, \dots, \theta_k^{(0)})$;
- Gerar $\theta_2^{(1)}$ de $\pi(\theta_2|y, \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)})$;
- Gerar $\theta_3^{(1)}$ de $\pi(\theta_3|y, \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, \dots, \theta_k^{(0)})$; (6)

...

- Gerar $\theta_k^{(1)}$ de $\pi(\theta_k|y, \theta_1^{(1)}, \dots, \theta_{k-1}^{(1)})$;

Substitua os valores iniciais por $\theta^{(1)} = \theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_k^{(1)}$, para uma nova realização. Os valores $\theta_1^{(z)}, \theta_2^{(z)}, \dots, \theta_k^{(z)}$, para z suficientemente grande, convergem para um valor da quantidade aleatória com distribuição $\pi(\theta|y)$ (Bernardo e Smith, 1994; Casela e George, 1992).

2.3.2 Algoritmo de Metropolis-Hastings

Suponha uma amostra de uma densidade não regular $\pi(\theta_i|\theta_{(i)})$, em que $\theta_{(i)} = \theta_{i-1}, \theta_{i+1}, \dots, \theta_k$. Seja $q(\theta, \beta)$ o núcleo de transição da distribuição $p(\theta)$ que representa $\pi(\theta_i|\theta_{(i)})$ e que transforma θ em β .

Desta forma o algoritmo é dado por:

- Inicie com $\theta^{(0)}$ e indicador de estado $j = 0$;

- Gerar um ponto β do núcleo de transição $q(\theta^{(j)}, \beta)$; (7)
- Atualizar θ^j por $\theta^{(j+1)} = \beta$, com probabilidade, $p = \min \left\{ \frac{p[\beta]q[\theta^{(j)}, \beta]}{p[\theta^{(j)}]q[\beta, \theta^j]} \right\}$,
ficar com $\theta^{(j)}$ com probabilidade $1 - p$.
- Repetir os dois últimos passos até obter uma distribuição estacionária.

Importante observação: se um valor candidato é rejeitado, então o valor atual é considerado na próxima etapa; no terceiro passo o valor de p não depende da constante normalizadora; o algoritmo de Metropolis-Hastings é especificado pela densidade candidata para geração $q(x, y)$ (Bernardo e Smith, 1994; Chib e Greenberg, 1995).

Na prática podemos usar alguns programas computacionais na simulação de amostras da distribuição *a posteriori* de interesse. Como um caso especial, com o uso do software OpenBugs, um software muito popular e de licença livre, as simulações antes trabalhosas ficam muito mais simples pois só é necessário especificar a função de verossimilhança e as distribuições a priori para cada parâmetro sem necessidade de especificar cada distribuição condicional a posteriori para cada parâmetro como será considerado neste estudo.

Capítulo 3

3. Modelagem estatística

3.1 Modelos de Regressão com pontos de mudança

O modelo de regressão com um ponto de mudança é dado pela seguinte formulação:

$$Y(t) = \beta_{0j} + \beta_{1j}T(t) + \epsilon(t) \quad , \text{ onde } t = 1, 2, \dots, N \text{ e } j = 1, 2 \quad (8)$$

Sendo assim, temos:

$$\begin{cases} Y(t) = \beta_{01} + \beta_{11}T(t) + \epsilon(t), & \text{para } T(t) \leq \zeta \\ Y(t) = \beta_{02} + \beta_{12}T(t) + \epsilon(t), & \text{para } T(t) > \zeta \end{cases} \quad (9)$$

onde $T(t)$ representa a covariável tempo; $\epsilon(t)$ são erros *iid's* com distribuição $N(0, \sigma_{\epsilon j}^2), j = 1, 2$, onde,

$$\begin{cases} \sigma_{\epsilon 1}^2, T(t) \leq \zeta \\ \sigma_{\epsilon 2}^2, T(t) > \zeta \end{cases} \quad (10)$$

Desta forma será necessário estimar os parâmetros $\beta_{01}, \beta_{11}, \beta_{02}, \beta_{12}, \sigma_{\epsilon 1}^2, \sigma_{\epsilon 2}^2$ e ζ . Em se tratando de abordagem Bayesiana podemos considerar uma distribuição a priori discreta para o ponto de mudança ζ assumindo valores no intervalo $1, 2, 3, \dots, N$.

3.2 Modelos de Regressão com efeitos defasados

Em algumas situações o modelo se mostra mais aderente quando efeitos defasados são considerados, isto é, existe uma estrutura de dependência entre as variáveis resposta.

A partir do modelo dado em (8) podemos generalizar a formulação considerando qualquer número n de atrasos como efeitos defasados. Usualmente se assumem atrasos de acordo com a demanda dos dados, por exemplo para capturar interferências referentes à sazonalidade. A formulação é dada abaixo,

$$\begin{aligned}
 Y(1) &= \beta_{0j} + \beta_{1j}T(1) + \epsilon(1) \\
 Y(2) &= \beta_{0j} + \beta_{1j}T(2) + \beta_{2j}Y(1) + \epsilon(1) \\
 &\vdots \\
 Y(t) &= \beta_{0j} + \beta_{1j}T(t) + \sum_{l=2}^n \beta_{lj}Y(t-l+1) + \epsilon(t), j = 1,2,
 \end{aligned} \tag{11}$$

onde temos os parâmetros,

$$\begin{cases} \beta_{01}, \beta_{11}, \dots, \beta_{n1}, \text{ para } T(t) \leq \zeta \\ \beta_{02}, \beta_{12}, \dots, \beta_{n2}, \text{ para } T(t) > \zeta \end{cases} \tag{12}$$

$\epsilon(t)$ são erros *iid's* com distribuição $N(0, \sigma_{\epsilon j}^2), j = 1,2$ para, $\begin{cases} \sigma_{\epsilon 1}^2, T(t) \leq \zeta \\ \sigma_{\epsilon 2}^2, T(t) > \zeta \end{cases}$

Para esse caso teremos um total de $2n + 5$ parâmetros para serem estimados e assumiremos que o ponto de mudança ζ assume a mesma priori discreta mencionada acima.

3.3 Processos de Poisson Não Homogêneos na presença de k pontos de mudança

3.3.1 Função de Intensidade

Dado as considerações das peculiaridades para a escolha do modelo introduzido na seção 1.4 consideramos como função de intensidade nesse estudo a Lei de Potência (ou PLP mencionado anteriormente) considerando a presença de k pontos de mudança.

$$\lambda(t|\theta) = \begin{cases} \lambda_1 = \frac{\beta_1}{\alpha_1} \left(\frac{t}{\alpha_1}\right)^{\beta_1-1}, & 0 < t < \zeta_1 \\ \lambda_2 = \frac{\beta_2}{\alpha_2} \left(\frac{t}{\alpha_2}\right)^{\beta_2-1}, & \zeta_1 \leq t < \zeta_2 \\ \lambda_3 = \frac{\beta_3}{\alpha_3} \left(\frac{t}{\alpha_3}\right)^{\beta_3-1}, & \zeta_2 \leq t < \zeta_3 \\ \lambda_4 = \frac{\beta_4}{\alpha_4} \left(\frac{t}{\alpha_4}\right)^{\beta_4-1}, & \zeta_3 \leq t < \zeta_4 \\ \vdots \\ \lambda_n = \frac{\beta_n}{\alpha_n} \left(\frac{t}{\alpha_n}\right)^{\beta_n-1}, & \zeta_{n-1} \leq t < \zeta_n \end{cases} \quad (13)$$

Utilizando a intensidade PLP obtemos a seguinte função de valor médio:

$$m(t|\theta) = \begin{cases} m_1(t), & 0 < t < \zeta_1 \\ m_2(t) - m_2(\zeta_1) + m_1(\zeta_1), & \zeta_1 < t < \zeta_2 \\ m_3(t) - m_3(\zeta_2) + m_2(\zeta_2) - m_2(\zeta_1) + m_1(\zeta_1), & \zeta_2 \leq t < \zeta_3 \\ \vdots \\ m_n(t) + \sum_{i=1}^n -m_i(\zeta_{i-1}) + m_{i-1}(\zeta_{i-1}), & \zeta_{n-1} \leq t < \zeta_n \end{cases} \quad (14)$$

3.3.2 Função de Verossimilhança

A função de verossimilhança para o modelo de Poisson não homogêneo é dada por (ver Cox e Lewis, 1966 e Insua et al., 2012),

$$L(T|\theta) = [\prod_{i=1}^n \lambda_j(t_i)] \exp(-m(t)), j = 1 \dots n \quad (15)$$

Isto é,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{N(\zeta_1)} \lambda_1(t_i) \exp\{-m_1(\zeta_1)\} \prod_{i=N(\zeta_1)+n-1}^{N(\zeta_n)} \lambda_{n-1}(t_i) \exp\{-m_{n-1}(\zeta_{n-1}) \pm m_{n-1}(\zeta_{n-2})\} \\ &\times \prod_{i=N(\zeta_1)+n}^{N(T)} \lambda_n(t_i) \exp\{-m_n(T) + m_n(\zeta_{n-1})\} \end{aligned}$$

3.3.3 Inferência Bayesiana

Aplicando o método Bayesiano no modelo com a presença de k pontos de mudanças assumindo a função de intensidade PLP citado na seção anterior, assumimos distribuições uniformes a priori para α_j e β_j , dadas por $\alpha_j \sim U(0, a_j)$ e $\beta_j \sim U(b_{1j}, b_{2j})$ para $j = 1 \dots n$, onde a_j , b_{1j} e b_{2j} são hiperparâmetros conhecidos. Para os pontos de mudanças $\zeta_k \sim U(c_{1k}, c_{2k})$, $k = 1, \dots, n$ também assumimos distribuições a priori uniformes onde c_{1k} e c_{2k} também são hiperparâmetros conhecidos e independentes. Métodos de Monte Carlos Cadeias de Markov (MCMC) são utilizados para simular amostras da distribuição conjunta a posteriori de interesse. Neste caso, o algoritmo de simulação passo a passo é dado por:

Passo 1: Adota-se um valor inicial $\zeta_k^{(0)}$, $k = 1, \dots, n$ e $\alpha_j^{(0)}$, $\beta_j^{(0)}$, $j = 1, \dots, n$, respectivamente. Indique os valores de ζ_k , $k = 1, \dots, n$ e α_j , β_j , $j = 1, \dots, n$ no i -ésimo passo por $\zeta_k^{(i)}$, $k = 1, \dots, n$ e $\alpha_j^{(i)}$, $\beta_j^{(i)}$, $j = 1, \dots, n$;

Passo 2: Gerar $\zeta_1^{(i)}$, $\zeta_k^{(i+1)}$, $k = 2, \dots, n$ e $\alpha_j^{(i+1)}$, $\beta_j^{(i+1)}$, $j = 1, \dots, n$ da distribuição a posteriori condicional, e repita o procedimento N vezes;

Passo 3: Calcule $\mu(\alpha, \beta, \zeta)$ da estimativa Bayesiana de Monte Carlo por $\sum_{i=B+1}^N \mu(\alpha^{(i)}, \beta^{(i)}, \zeta^{(i)}) / (N - B)$ onde B é considerado o período de *burn in*.

O procedimento abordado anteriormente pode ser generalizado para modelos de NHPP contendo quaisquer números de pontos de mudanças e qualquer função intensidade paramétrica.

Capítulo 4

4. Aplicações em dados reais

4.1 Um ponto de mudança

Nesta primeira aplicação utilizaremos um conjunto de dados oriundos do portal *Stats NZ*, considerado como uma fonte de dados oficiais da Nova Zelândia. Dele foi retirado a série histórica do período de agosto de 1989 a maio 2012 contendo a contagem mensal de indivíduos que visitaram o país com finalidade educacional. Ao todo foram registrados 154 meses em que nenhum dos meses apresentou registro zero de visitas, podendo desta forma a utilização da transformação para escala logarítmica na variável de interesse ($Y(t) = \log(\text{contagem de visitantes no tempo } t)$). Da Figura 1 podemos observar a presença de apenas um ponto de mudança ζ , desta forma, a partir dos dados em escala transformada podemos ajustar um modelo de regressão linear na presença de um ponto de mudança. Considerando os modelos regressão linear apresentados nas equações (8) e (11) os quais denotaremos nesta aplicação como modelo 1 e modelo 2, sendo a variável explicativa o tempo em meses na presença do ponto de mudança ζ assumindo as seguintes distribuições a priori independentes para os parâmetros dos modelos:

$$\zeta \sim Unif(1,154)$$

$$\psi_1 = \frac{1}{\sigma_{\epsilon_1}^2} \sim Gamma(1,1)$$

$$\psi_2 = \frac{1}{\sigma_{\epsilon_2}^2} \sim Gamma(1,1)$$

$$\beta_{0j} \sim N(5,100), j = 1,2$$

$$\beta_{ij} \sim N(5,100), i = 1,2, \dots, 6, j = 1,2$$

Na Tabela 2 são apresentadas as estimativas dos parâmetros das distribuições a posteriori geradas para ambos os modelos, em complemento, nas Figuras 4 e 5 são apresentados os gráficos dos dados observados sobrepostos pelos valores das médias estimadas e os histogramas dos valores simulados dos pontos de mudança utilizando

métodos MCMC e o software OpenBugs. A verificação da convergência do algoritmo de simulação MCMC foi observada a partir de gráficos das series temporais para amostras geradas para cada parâmetro.

Tabela 2 – Sumários a posteriori para os modelos 1 e 2 para os dados de visitantes na Nova Zelândia.

	Parâmetro	Média Post. (D.P.)	Int. Cred. (95%)
Modelo 1	β_{01}	7.3333 (0.2872)	(6.7100;7.8600)
	β_{02}	9.3855 (0.0508)	(9.2780;9.4820)
	β_{11}	0.1104 (0.0320)	(0.0629;0.1934)
	β_{12}	0.0009 (0.0005)	(-0.0001;0.0020)
	ζ	20.6770 (2.4136)	(14.0000;25.0000)
	ψ_1	3.8580 (1.2496)	(1.7960;6.6540)
	ψ_2	26.1801 (3.3462)	(19.9597;33.0800)
Modelo 2	β_{01}	6.9762 (0.3136)	(6.3500;7.5880)
	β_{02}	8.6834 (0.1117)	(8.4730;8.9120)
	β_{11}	0.1403 (0.4150)	(0.0605;0.2235)
	β_{12}	0.0002 (0.0003)	(-0.0005;0.0010)
	β_{21}	9.4E-05 (3.4E-05)	(2.5E-05;0.0002)
	β_{22}	9.0E-05 (1.2E-05)	(6.5E-05;0.0001)
	β_{31}	-6.3E-05 (3.6E-05)	(-0.0001;8.1E-06)
	β_{32}	-6.5E-05 (1.8E-05)	(-0.0001;-2.9E-05)
	β_{41}	4.6E-05 (3.7E-05)	(-2.8E-05;0.0001)
	β_{42}	4.8E-05 (1.9E-05)	(1.16E-05;8.8E-05)
	β_{51}	-2.2E-05 (3.6E-05)	(-9.4E-05;5.0E-05)
	β_{52}	-1.5E-05 (1.8E-05)	(-5.2E-05;1.9E-05)
	β_{61}	-1.9E-05 (2.7E-05)	(-7.2E-05;3.5E-05)
	β_{62}	5.4E-07 (1.2E-05)	(-2.2E-05;2.5E-05)
	ζ	15.6822 (1.7788)	(14.0000;19.0000)
	ψ_1	4.2664 (1.5892)	(1.7410;7.8910)
	ψ_2	41.6983 (5.2187)	(32.0797;52.5400)

Figura 4 – Verificação dos ajustes para os modelos 1 e 2, com valores observados versus médias ajustadas

Figura 5 – Histogramas dos valores simulados para os pontos de mudança dos modelos 1 e modelo 2

A partir da análise gráfica há indicação que o modelo 2 é mais aderente aos dados observados. Além disso, com propósito de reduzir as subjetividades da análise, no apêndice 1 são apresentadas algumas medidas de ajuste (MAPE, MAD e MSD) para estabelecer comparação quantitativa entre os modelos. Os resultados podem ser observados na Tabela 3 indicando que o modelo 2 realmente possui um ajuste melhor quando comparado ao modelo 1.

Tabela 3 – Medidas de ajuste para os modelos 1 e 2 dos dados de visitantes a Nova Zelândia.

Critério	Modelo 1	Modelo 2
MAPE	1.5561	0.9126
MAD	0.1399	0.0816
MSD	0.0393	0.0150

4.2 Dois pontos de mudança

Nesta segunda aplicação serão utilizados dados de séries epidemiológicas referente à incidência de tuberculose (TB) na cidade de Nova York entre os anos de 1970 a 2000. O gráfico dos dados sugere três movimentos de tendência nos dados (Figura 6): a primeira no período de (1970 a 1979), declínio possivelmente associado às boas políticas de controle da doença; o segundo período (1979 a 1992) houve uma incidência crescente possivelmente justificada pela epidemia de HIV que assolou a época; o terceiro período (1992 a 2000) houve um declínio novamente nos níveis de incidência (Achcar et al., 2008), muito provavelmente relacionado ao início de novas terapias voltadas ao tratamento da AIDS. Dos gráficos da Figura 6 (escala original e logarítmica), é possível observar claramente a presença de dois pontos de mudança próximos ao décimo ano e vigésimo segundo ano.

Figura 6 – Incidência de Tuberculose na cidade de Nova York no período de 1970 a 2000.

Para a análise bayesiana dos dados foi proposto um modelo de regressão linear na presença de dois pontos de mudança ζ_1 e ζ_2 , com modelo 1 se referindo ao modelo tradicional de regressão linear (8) e modelo 2 se referindo ao modelo com efeitos defasados (11), dado que não houve registros contendo zero na série epidemiológica estudada. Considerando apenas a covariável ano no modelo 1 assumimos as seguintes distribuições a priori independentes para os parâmetros:

$\zeta_1 \sim$ priori informativa discreta $[1,...,31]$ concentrado em torno do valor 5 (maior massa de probabilidade)

$\zeta_2 \sim$ priori informativa discreta $[1,...,31]$ concentrado em torno do valor 22 (maior massa de probabilidade)

$$\psi_1 = \frac{1}{\sigma_{\epsilon_1}^2} \sim Gamma(1,1)$$

$$\psi_2 = \frac{1}{\sigma_{\epsilon_2}^2} \sim Gamma(1,1)$$

$$\psi_3 = \frac{1}{\sigma_{\epsilon_3}^2} \sim Gamma(1,1)$$

$$\beta_{0j} \sim N(7,1), j = 1,2,3$$

$$\beta_{1j} \sim N(0,1), j = 1,2,3$$

Além destas, foram consideradas outras distribuições a priori informativas para os parâmetros dos modelos, especialmente para os modelos de dois pontos de mudança para otimizar a convergência das simulações através do amostrador de Gibbs/Metropolis-Hastings.

Os sumários a posteriori de interesse da distribuição a posteriori para os parâmetros relacionados ao modelo 1 são apresentados na Tabela 4. Nela também constam os sumários a posteriori para os parâmetros do modelo 2, este considerando a presença de seis efeitos defasados e assumindo as mesmas distribuições a priori considerados no caso para uma covariável (ano). São elas:

$\zeta_1 \sim$ priori informativa discreta $[1,...,31]$ concentrado em torno do valor 10 (maior massa de probabilidade)

$\zeta_2 \sim$ priori informativa discreta $[1,...,31]$ concentrado em torno do valor 22 (maior massa de probabilidade)

$$\psi_1 = \frac{1}{\sigma_{\epsilon_1}^2} \sim Gamma(1,1)$$

$$\psi_2 = \frac{1}{\sigma_{\epsilon_2}^2} \sim Gamma(1,1)$$

$$\psi_3 = \frac{1}{\sigma_{\epsilon_3}^2} \sim \text{Gamma}(1,1)$$

$$\beta_{0j} \sim N(7,1), j = 1,2,3$$

$$\beta_{ij} \sim N(0,1) \quad i = 1, \dots, 6, j = 1,2,3$$

A verificação da convergência do algoritmo de simulação MCMC foi observada a partir de gráficos das series temporais para amostras geradas para cada parâmetro.

A Figura 7 apresenta os gráficos dos valores da série epidemiológica observados sobrepostos pelas médias estimadas dadas pelos modelos 1 e 2. Observa-se que o modelo 2 possui uma maior aderência aos dados; em complemento a Figura 8 apresenta o histograma dos pontos de mudança simulados pelo método de Gibbs/Metropolis-Hastings para o modelo 2. Além disso, pelos resultados mostrados na Tabela 4, as estimativas para os pontos de mudança são dadas por $\zeta_1 = 8.5923$ (próximo a 9) e $\zeta_2 = 26.2955$ (próximo a 25) assumindo o modelo 2 (isto é, o primeiro ponto de mudança foi estimado para o ano de 1978 que corresponde ao início da epidemia de HIV e o segundo ponto de mudança foi estimado para 1994, ano no qual ocorre o declínio dos casos de TB em Nova York, coincidindo com o início dos novos tratamentos para a epidemia de HIV). Os resultados estão em consonância com as conclusões obtidas por Achcar et al. (2008), estudo no qual foi empregado um modelo de Poisson não Homogêneo (NHPP) na presença de dois pontos de mudança se utilizando da abordagem Bayesiana através de métodos de MCMC. Importante destacar que a metodologia proposta apresenta uma modelagem mais simplificada para dados de séries epidemiológicas quando comparado à modelagem através do uso de NHPP.

Figura 7 – Médias estimadas para os modelos 1 e 2, valores observados versus médias estimadas

Figura 8– Histograma dos pontos de mudanças do modelo 2

Tabela 4 – Sumários a posteriori para os modelos 1 e 2

	Parâmetro	Média Post. (D.P.)	Int. Cred. (95%)
Modelo 1	β_{01}	7.7502 (0.4308)	(6.8350;8.5560)
	β_{02}	6.6156 (0.3797)	(5.8600;7.3520)
	β_{03}	8.4377 (0.9599)	(6.5020;10.2600)
	β_{11}	-0.0209 (0.0345)	(-0.1981;0.1903)
	β_{12}	0.0637 (0.0228)	(0.0192;0.1092)
	β_{13}	-0.0326 (0.0349)	(-0.0991;0.0375)
	ζ_1	8.7593 (1.0588)	(6.0000;11.0000)
	ζ_2	24.6699 (1.3325)	(22.0000;27.0000)
	ψ_1	3.8297 (1.9796)	(0.9616;8.5600)

	Parâmetro	Média Post. (D.P.)	Int. Cred. (95%)
Modelo 2	ψ_2	7.0740 (2.5454)	(2.9870;12.8400)
	ψ_3	3.4981 (1.7868)	(0.8930;7.7750)
	β_{01}	7.5355 (0.5933)	(6.2970;8.6550)
	β_{02}	6.6417 (0.7009)	(5.2500;8.0390)
	β_{03}	8.0085 (0.9972)	(6.0490;9.9690)
	β_{11}	-0.0227 (0.1724)	(-0.3868;0.3071)
	β_{12}	0.0540 (0.0911)	(-0.1311;0.2335)
	β_{13}	-0.0081 (0.1080)	(-0.2156;0.2010)
	β_{21}	0.0002 (0.0003)	(-0.0004;0.0008)
	β_{22}	0.0002 (0.0004)	(-0.0005;0.0010)
	β_{23}	0.0002 (0.0014)	(-0.0023;0.0028)
	β_{31}	2.8E-05 (0.0003)	(-0.0007;0.0007)
	β_{32}	-0.0001 (0.0004)	(-0.0010;0.0008)
	β_{33}	-0.0002 (0.0014)	(-0.0025;0.0022)
	β_{41}	-3.8E-05 (0.0003)	(-0.0007;0.0006)
	β_{42}	-0.0001 (0.0004)	(-0.0009;0.0009)
	β_{43}	-0.0001 (0.0014)	(-0.0025;0.0022)
	β_{51}	-1.4E-05 (0.0002)	(-0.0003;0.0003)
	β_{52}	-3.1E-05 (0.0002)	(-0.0005;0.0005)
	β_{53}	-0.0001 (0.0013)	(-0.0024;0.0021)
	β_{61}	0.0032 (0.9976)	(-1.9520;1.9560)
	β_{62}	-0.0017 (1.0003)	(-1.9690;1.9650)
	β_{63}	0.0001 (1.0037)	(-1.9720;1.9750)
	ζ_1	8.5923 (1.1624)	(6.0000;11.0000)
	ζ_2	26.2955 (1.1953)	(24.0000;27.0000)
	ψ_1	2.6233 (1.5539)	(0.5253;6.4230)
	ψ_2	5.9281 (2.3709)	(2.2260;11.4100)
	ψ_3	1.2920 (1.2342)	(0.0406;4.5760)

Com propósito de reduzir as subjetividades da análise utilizaremos novamente os valores de MAPE, MAD e MSD para estabelecer comparação quantitativa entre os

modelos. Os resultados podem ser observados na tabela 4 indicando que o modelo 2 realmente possui um ajuste melhor quando comparado ao modelo 1.

Tabela 5 Medidas de ajuste para os modelos 1 e 2 dos dados de TB para cidade de Nova York

Critério	Modelo 1	Modelo 2
MAPE	1.3687	0.7892
MAD	0.1052	0.0611
MSD	0.0153	0.0089

4.3 Cinco pontos de mudança

Com intuito de generalização da aplicabilidade do modelo utilizando a proposta do modelo linear citado anteriormente, é apresentado nesta seção uma aplicação assumindo a presença de cinco pontos de mudança. Para este estudo será adotado a série epidemiológica da incidência mensal de sarampo ocorrido na Itália no período de janeiro de 2011 a junho de 2018. Esse conjunto de dados foi obtido do site da Organização Mundial da Saúde (WHO) totalizando 90 registros (em meses). Da Figura 9 é possível verificar a presença de 5 pontos de mudança notáveis facilmente no quinto mês; trigésimo mês; trigésimo oitavo mês; septuagésimo quinto mês e octogésimo oitavo mês.

Figura 9 – Incidência de sarampo na Itália período de Janeiro 2011 a Junho 2018

Considerando a ausência de registros contendo zero na série podemos utilizar o modelo de regressão linear proposto anteriormente realizando a transformação em escala logarítmica da variável de interesse (casos de sarampo). Para o modelo iremos admitir como covariável a mês e efeitos defasados de cinco meses para a variável resposta, desta forma temos as seguintes distribuições independentes a priori para o modelo:

$\zeta_1 \sim$ priori informativa discreta $[1,...,90]$ concentrado em torno do valor 10 (maior massa de probabilidade)

$\zeta_2 \sim$ priori informativa discreta $[1,...,90]$ concentrado em torno do valor 30 (maior massa de probabilidade)

$\zeta_3 \sim$ priori informativa discreta $[1,...,90]$ concentrado em torno do valor 38 (maior massa de probabilidade)

$\zeta_4 \sim$ priori informativa discreta $[1,...,90]$ concentrado em torno do valor 75 (maior massa de probabilidade)

$\zeta_5 \sim$ priori informativa discreta $[1,...,90]$ concentrado em torno do valor 88 (maior massa de probabilidade)

$$\psi_j = \frac{1}{\sigma_{\epsilon_1}^2} \sim \text{Gamma}(1,1), j = 1, \dots, 6$$

$$\beta_{0j} \sim N(7,1), j = 1, \dots, 6$$

$$\beta_{1j} \sim N(0,1), j = 1, \dots, 6$$

Nesta aplicação foram utilizadas com maior frequência distribuições a priori informativas para facilitar a convergência do algoritmo na geração das amostras de Gibbs. A verificação da convergência do algoritmo de simulação MCMC foi observada a partir de gráficos das series temporais para amostras geradas para cada parâmetro. A Tabela 6 contém os sumários a posteriori; em complemento a Figura 10 apresenta os dados originais sobrepostos pelos valores das médias estimadas resultantes dos modelos.

Tabela 6 – Sumários a posteriori das distribuições a posteriori do modelo 1 e modelo 2 (efeitos defasados)

Parâmetro	Média Post (D.P.)	Int. Cred. (95%)
β_{01}	6.1118 (0.7077)	(4.8720;7.6860)
β_{02}	6.0774 (0.6513)	(4.7770;7.3160)
β_{03}	6.7984 (0.9947)	(4.8680;8.7460)
β_{04}	5.9838 (0.8068)	(4.4130;7.5750)
β_{05}	7.4359 (0.9808)	(5.5070;9.3730)
β_{06}	6.9828 (0.9932)	(5.0640;8.9380)
β_{11}	0.0779 (0.3241)	(-0.7746;0.5591)
β_{12}	-0.1104 (0.0379)	(-0.1838;-0.0355)
β_{13}	-0.0479 (0.0342)	(-0.1163;0.0172)
β_{14}	-0.0360 (0.0141)	(-0.0638;-0.0087)
β_{15}	-0.0213 (0.0127)	(-0.0462;0.0036)
β_{16}	-0.0125 (0.0127)	(-0.0372;0.0117)
ζ_1	5.6807 (1.2520)	(3.0000;7.0000)
ζ_2	28.3026 (0.6201)	(28.0000;30.0000)
ζ_3	38.5736 (1.3484)	(36.0000;40.0000)
ζ_4	73.1087 (0.3469)	(73.0000;74.0000)
ζ_5	86.9629 (1.0631)	(85.0000;89.0000)
ψ_1	2.1453 (1.4293)	(0.3015;5.7150)
ψ_2	0.7401 (0.2292)	(0.3685;1.2570)
ψ_3	0.9572 (0.4681)	(0.3171;2.1340)
ψ_4	1.1099 (0.3064)	(0.6147;1.8060)
ψ_5	1.4692 (0.5552)	(0.5908;2.7330)
ψ_6	2.2714 (1.4911)	(0.2953;5.9500)

Figura 10 –Valores observados versus valores das médias estimadas para os dados de Sarampo na Itália

4.4 NHPP na presença de vários pontos de mudança

Para este exemplo será adotado a série histórica das incidências de sarampo no Brasil no período de janeiro de 2012 a agosto de 2018, totalizando 80 meses registrados. Os dados podem ser visualizados na Tabela 1 e na Figura 2, ao analisá-los é possível observar que existem vários registros de zero em determinados intervalos, consequência da erradicação da doença através de imunização no período. O modelo de regressão linear tradicional utilizando transformação logarítmica já não se torna uma opção viável nesta situação. Sendo assim, é proposta a modelagem do problema através do Processo de Poisson Não Homogêneo citado anteriormente adotando a Lei de Potência (PLP) como sua função de intensidade. As distribuições independentes a priori para os parâmetros são dadas por:

$$\zeta_1 = U(300,600)$$

$$\zeta_2 = U(1000,1500)$$

$$\zeta_3 = U(2200,2350)$$

$$\alpha_1 = U(0,500)$$

$$\alpha_2 = U(0,100)$$

$$\alpha_3 = U(0,500)$$

$$\alpha_4 = U(0,500)$$

$$\beta_1 = U(1,3)$$

$$\beta_2 = U(0,1)$$

$$\beta_3 = U(1,3)$$

$$\beta_4 = U(0,1)$$

Assumir ζ_1, ζ_2 e ζ_3 denotando os pontos de mudança, $\alpha_1, \alpha_2, \alpha_3$ e α_4 denotando os parâmetros de escala e $\beta_1, \beta_2, \beta_3$ e β_4 denotando os parâmetros de forma. Note que os hiperparâmetros para os parâmetros de forma foram propositalmente definidos para terem o comportamento crescente, decrescente, crescente e decrescente, de acordo com a função intensidade (ver Figura 3). Foi utilizado para as simulações o amostrador de Gibbs/Metropolis-Hastings, considerando um *burn-in* de tamanho 1000 para eliminar as possíveis distorções nas simulações iniciais das iterações; após esse processo foram simuladas 1000 observações adicionais. As distribuições conjuntas a posteriori foram geradas através de métodos de MCMC com o uso do software OpenBugs. A convergência dos algoritmos foi verificada a partir de trace-plots das amostras simuladas; complementarmente a Tabela 7 apresenta as estimativas dos parâmetros das distribuições a posteriori. O resultado obtido das estimativas de Monte Carlos para a função de valor médio baseado nas 1000 amostras de Gibbs é dado por:

$$m(t|\theta) = \begin{cases} (t/139.5000)^{1.7620}, & \text{se } 0 < t < 454.1000 \\ (t/0.6705)^{0.9891}, & \text{se } 454.1 \leq t < 1256 \\ (t/312.8000)^{1.5850}, & \text{se } 1256 \leq t < 2252 \\ (t/0.1395)^{0.9368}, & \text{se } 2252 \leq t < T \end{cases}$$

Figura 9 – Valores observados versus médias estimadas para os modelos com 1, 2 e 3 pontos de mudanças.

Tabela 7 – Sumários a posteriori considerando os modelos com 1 ponto, 2 pontos e 3 pontos de mudanças.

	Parâmetro	Média Post (D.P.)	Int. Cred. (95%)
Modelo 1	α_1	70.2700(25.22)	(1.0540;70.6200)
	α_2	5.5070(12.22)	(0.8225;25.0200)
	β_1	1.1260(0.3221)	(0.1896;1.5060)
	β_2	0.9879(0.009)	(0.9690;1.6640)
	ζ_1	379.8000(159.1y)	(24.5000;454.5000)
Modelo 2	α_1	79.94 (8.6750)	(74.8900;99.9300)
	α_2	0.6337 (0.3244)	(0.0490;0.8800)
	α_3	122.3 (22.0600)	(109.500;173.2000)
	β_1	2.562 (0.1118)	(2.327;2.791)
	β_2	0.5611 (0.081)	(0.3681;0.6604)
	β_3	3.012 (0.030)	(3.000;3.094)
	ζ_1	1222 (7.652)	(1217;1240)
	ζ_2	2254 (0.9543)	(2252;2254)
Modelo 3	α_1	139.5000 (47.5400)	(58.2300;245.0000)
	α_2	0.6705 (0.0611)	(0.5145;0.7577)

Parâmetro	Média Post (D.P.)	Int. Cred. (95%)
α_3	312.8000 (119.0000)	(88.8800;490.6000)
α_4	0.1395 (0.0615)	(0.0204;0.2471)
β_1	1.7620 (0.4413)	(1.0740;2.7860)
β_2	0.9891 (0.0110)	(0.9599;0.9996)
β_3	1.5850 (0.2829)	(1.0600;2.1080)
β_4	0.9368 (0.0495)	(0.8048;0.9980)
ζ_1	454.1000 (2.5940)	(451.7000;456.4000)
ζ_2	1256.0000 (1.8660)	(1254.0000;1259.0000)
ζ_3	2252.0000 (0.2510)	(2252.0000;2252.0000)

A Figura 9 apresenta graficamente os resultados do modelo, sobreposição dos dados observados pelas médias estimadas pelos modelos.

Tabela 6 – Medidas de Ajustes para os modelos de NHPP

Critério	Modelo 1	Modelo 2	Modelo 3
MAPE	78,20%	235,67%	35,06%
MAD	386	52	45
MSD	243.634	9.143	9.210

Capítulo 5

Conclusões

Das aplicações para os modelos propostos em diferentes situações é possível concluir que os métodos introduzidos neste estudo são bastante consistentes para modelagem de series epidemiológicas na presença de um ou mais pontos de mudança. Nas séries históricas na ausência de zeros foi possível observar a boa aderência das estimativas em relação aos dados originais utilizando a transformação logarítmica das contagens assumindo modelos de regressão linear múltipla com erros normais na presença de pontos de mudança. Em situações em que ocorrem a presença de zeros na série histórica como no exemplo de dados de contagem de sarampo no Brasil, o Processo de Poisson Não Homogêneo (NHPP) mostrou ser uma boa alternativa para a modelagem dos dados. Uma observação cabível para ambos os modelos seria o fato da perda de precisão nas estimativas pontuais na presença de valores observados atípicos extremamente discrepantes (*Outliers*), porém as estimativas para os pontos de mudança calculadas pelo método de MCMC utilizando de abordagem Bayesiana apresentam boa precisão. Ao comparar as duas abordagens, o modelo de regressão contendo efeitos defasados se mostra mais sensível em captar as possíveis oscilações de volatilidade dos dados ao longo do tempo, situação que não acontece no NHPP.

É importante salientar que nas aplicações apresentadas nesse estudo foi observado que a escolha de distribuições a priori com alguma informação previa (uso de métodos Bayesianos empíricos) a partir dos dados é de grande importância na obtenção de inferências precisas, pois as inferências muitas vezes são sensíveis a escolha de distribuições a priori (Carlin and Louis, 2000). A escolha de distribuições a priori adequadas também pode ser obtida a partir da informação de especialistas na área de aplicação. Por exemplo, especialistas em saúde pública quando consideramos series epidemiológicas.

O uso de métodos clássicos de inferências para modelos de regressão de contagem na presença de pontos de mudança podem ser muito complexos ou as vezes difíceis de serem aplicados, o que torna o enfoque clássico muitas vezes inviável para aplicações em

áreas como saúde pública. Além disso, sob o enfoque Bayesiano há várias técnicas para discriminação de modelos baseadas nas distribuições preditivas de forma simples.

No estudo foi observado que o uso da metodologia Bayesiana especialmente usando o software Openbugs o custo computacional foi pequeno e sem necessidade de grande esforço e treinamento para a programação computacional necessária na geração de amostras da distribuição a posteriori conjunta de interesse, o que pode ser de grande utilidade para a área de saúde pública na análise de series epidemiológicas na presença de pontos de mudança.

Outro ponto importante: na modelagem dos dados de contagem usando processos de Poisson não-homogêneos foi utilizado o processo PLP, mas nas aplicações muitas outras formas paramétricas poderiam ser assumidas para a função intensidade (ver por exemplo, Kuo e Yang (1996)).

Podemos generalizar a aplicação do modelo para outras séries epidemiológicas, como por exemplo, nas series epidemiológicas da epidemia recente de Covid-19 onde há vários pontos de mudança nas series relacionadas a estratégias de isolamento social, ondas diferentes da epidemia ou possíveis mutações do vírus que torna a infecção mais transmissível. Essa modelagem leva a melhor entendimento da pandemia, verificação das medidas de isolamento e suporte para as autoridades da área de saúde pública na tomada de decisões.

Uma vez que os modelos se mostraram aderentes quando utilizamos o tempo como covariável, o próximo passo também seria avaliar a viabilidade de correlacionar o modelo a covariáveis da área de planejamento da saúde pública, desta forma é possível utilizar o modelo para estimação de possíveis casos/surtos de sarampo de acordo com a cobertura vacinal.

Agradecimento

Este estudo foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Referências Bibliográficas

Achcar, J., Martinez, E., Ruffino-Netto, A., Paulino, C., Soares, P. (2008). A statistical model investigating the prevalence of tuberculosis in New York City using counting processes with two change-points. *Epidemiology & Infection*, 136(12):1599- 1605.

Achcar, J. A. (2001). Bayesian analysis for software reliability data. In *Advances in Reliability*, volume 20 of *Handbook of Statistics*, 733 - 748. Elsevier.

Achcar, J. A., Bolfarine, H. (1989). Constant hazard against a change-point alternative: a Bayesian approach with censored data. *Communications in Statistics-Theory and Methods*, 18(10):3801-3819.

Achcar, J.A., Dey, D.K., Niverthi, M. (1998). A Bayesian approach using nonhomogeneous Poisson processes for software reliability models. In *Frontiers in Reliability*, 1-18. World Scientific.

Achcar, J. A., Loibel, S. (1998). Constant hazard function models with a change point: A Bayesian analysis using Markov chain monte Carlo methods. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 40(5):543-555.

Achcar, J. A., Rodrigues, E. R., Paulino, C. D., Soares, P. (2010). Non-homogeneous Poisson models with a change-point: an application to ozone peaks in Mexico City. *Environmental and Ecological Statistics*, 17(4):521-541.

Achcar, J. A., Rodrigues, E. R., Tzintzun, G. (2011). Using non-homogeneous Poisson models with multiple change-points to estimate the number of ozone exceedances in mexico city. *Environmetrics*, 22(1):1-12.

Aminikhanghahi, S., Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339-367.

Arif N, Al-Jefri M, Bizzi IH, Perano GB, Goldman M, Haq I, Chua KL, Mengozzi M, Neunez M, Smith H, Ghezzi P. (2018) *Fake News or Weak Science? Visibility and Characterization of Antivaccine Webpages Returned by Google in Different Languages and Countries*. Front Immunol. 2018 Jun 5;9:1215. doi: 10.3389/fimmu.2018.01215. Erratum in: Front Immunol. 2019 Oct 01;10:2252. PMID: 29922286; PMCID: PMC5996113.

Carlin, B. P., Gelfand, A. E., Smith, A. F. (1992). Hierarchical Bayesian analysis of change-point problems. *Applied Statistics*, 41(2): 389-405.

Carlin, B. , Louis, T. A. (2000), Bayes and Empirical Bayes Methods for Data Analysis, 2nd ed. London: Chapman & Hall.

Cherry JD, Zahn M. (2018) *Clinical Characteristics of Measles in Previously Vaccinated and Unvaccinated Patients in California*. Clinical Infectious Diseases, 67(9):1315-1319.

Chib, S., Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327-335.

Chopin, N. (2007). Dynamic detection of change points in long time series. *Annals of the Institute of Statistical Mathematics*, 59(2):349-366.

Cid, J. E. R., Achcar, J. A. (1999). Bayesian inference for nonhomogeneous Poisson processes in software reliability models assuming non-monotonic intensity functions. *Computational Statistics & Data Analysis*, 32(2):147-159.

Dey, D. K., Purkayastha, S. (1997). Bayesian approach to change point problems. *Communications in Statistics-Theory and Methods*, 26(8):2035-2047.

Domingues, C. M. et al. (1997) A evolução do sarampo no Brasil e a situação atual. *Inf. Epidemiol. Sus* [online]. 6(1):7-19.

Draper, N.R.; Smith, H.S. (1981) *Applied regression analysis*. Canada: Wiley Online Publisher.

Ducomble T, Gignoux E. (2020) *Learning from a massive epidemic: measles in DRC*. Lancet Infect Dis, 20(5):542. doi: 10.1016/S1473-3099(20)30265-6. Epub 2020 Mar 31. PMID: 32243817.

Gelfand, A. E., Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398-409.

Insua, D. R., Ruggeri, F., Wiper, M. P. (2012) *Bayesian analysis of stochastic process models* (ebook), Wiley Series in Probability and Statistics.

Jandhyala, V., Fotopoulos, S., MacNeill, I., Liu, P. (2013). Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, 34(4):423-446.

Kuroiwa C, Xayyavong P, Vongphrachanh P, Khampapongpane B, Yamanaka M, Nakamura S. (2003) *Difficulties in measles elimination: prevalence of measles antibodies before and after mass vaccination campaign in Laos*. Vaccine. 21(5-6):479-84. doi: 10.1016/s0264-410x(02)00480-2. PMID: 12531647.

Kuo, L., Yang, T. Y. (1996). Bayesian computation for nonhomogeneous Poisson processes in software reliability. *Journal of the American Statistical Association*, 91(434):763-773.

Lavielle, M. , Teyssiere, G. (2006). Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287-306.

Liu, S., Yamada, M., Collier, N., Sugiyama, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72-83.

Ma C, Rodewald L, Hao L, et al. (2019) *Progress Toward Measles Elimination — China, January 2013–June 2019*. MMWR Morb Mortal Wkly, 68:1112–1116.

Matthews, D. E., Farewell, V. T. (1982). On testing for a constant hazard against a change-point alternative. *Biometrics*, 38: 463-468.

Mudholkar, G. S., Srivastava, D. K., Freimer, M. (1995). The exponentiated Weibull family: A reanalysis of the bus-motor-failure data. *Technometrics*, 37(4):436-445.

Musa, J. D. , Okumoto, K. (1984). A logarithmic Poisson execution time model for software reliability measurement. In *Proceedings of the 7th international conference on Software engineering*, 230-238. IEEE Press.

Paulino, C.D., Amaral, M. A., Murteira, B. (2003). *Estatística Bayesiana*. Lisboa: Fundação Calouste Gulbenkian, 446 p. ISBN 972-31-1043-1

Pievatolo, A. , Ruggeri, F. (2004). Bayesian reliability analysis of complex repairable systems. *Applied Stochastic Models in Business and Industry*, 20(3):253-264.

Raftery, A. E. , Akman, V. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika*, 73(1): 85-89.

Ruggeri, F., Sivaganesan, S. (2005). On modeling change points in non-homogeneous Poisson processes. *Statistical Inference for Stochastic Processes*, 8(3):311-329.

Sarkar S, Zlojutro A, Khan K, Gardner L. (2019) *Measles resurgence in the USA: how international travel compounds vaccine resistance*. The Lancet Infectious Diseases. 19(7):684-686. doi: 10.1016/S1473-3099(19)30231-2.

Smith, A. F., Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):3-23.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D. (2003). WinBUGS user manual.

Venczel L, Dobbins J, André J, Laender F, Izurieta H, Delorme P, Voltaire HC. (2003) *Measles eradication in the Americas: experience in Haiti*. The Lancet Infectious Diseases, 2:103-108.

Van Dam ASG, Woudenberg T, de Melker HE, Wallinga J, Hahné SJM. (2020) *Effect of vaccination on severity and infectiousness of measles during an outbreak in the Netherlands, 2013-2014*. Epidemiology & Infection, 148: e81, 2020 03 23.

Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, et al.(1998) *Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children.*, 28(351):637–41. doi:10.1016/S0140-6736(97)11096-0

World Health Organization (2019) *Immunisation, vaccine and biologicals. New measles surveillance data from WHO*. Geneva: WHO. <https://www.who.int/immunization/newsroom/new-measles-data-august-2019/en/>

Yamanishi, K., Takeuchi, J.I. (2002). A unifying framework for detecting outliers and change points from non-stationary time series data. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 676-681.

Apêndice 1 – Medidas de Ajuste

- **Mean average percent error (MAPE)**

O MAPE expressa a precisão do modelo em percentual, dado pela equação:

$$\left[\frac{\sum_{t=1}^n |(y_t - \hat{y}_t) / y_t|}{n} \right] \times 100$$

onde, y_t representa as observações, $\hat{y}_t$ são as estimações resultantes do modelo e n número de observações do conjunto de dados. Essa medida expressa percentualmente a diferença entre valores observados e estimados, por se tratar de um percentual sua interpretação se torna muito mais simples quando comparado a outras estatísticas. Por exemplo, caso o valor resultante de MAPE seja 5, entendemos que nosso modelo possui uma divergência total de 5% entre os valores observado e os valores estimados.

- **Mean Absolute Deviation(MAD)**

O MAD expressa a precisão do modelo em valores absolutos, dado pela equação:

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

onde, y_t representa as observações, $\hat{y}_t$ são as estimações resultantes do modelo e n número de observações do conjunto de dados. Essa medida expressa a diferença absoluta entre os valores observados e estimados, não sendo tão sensível na presença de Outliers (Valores atípicos) quando comparado ao MSD.

- **Mean Square Deviation (MSD)**

O MSD expressa a precisão do modelo em valores quadráticos, dado pela equação:

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

onde, y_t representa as observações, $\hat{y}_t$ são as estimações resultantes do modelo e n número de observações do conjunto de dados. Essa medida expressa a diferença quadrática entre os valores observados e estimados, sendo mais sensível na presença de Outliers (Valores atípicos) quando comparado ao MAD. Ao realizar análise conjunta de MSD e MAD, caso o desvio quadrático esteja muito maior em relação ao desvio absoluto podemos concluir que existem Outliers nos valores estimados do nosso modelo.


```

+ (beta[J[i]]-1)*(log(t[i])-log(alpha[J[i]]))
L[i] <- lambda[i]*m
J[i] <- 1+step(t[i]-tau1-.5) + step(t[i]-tau2-.5)+step(t[i]-tau3-.5)
}
m <- exp(-(pow((tau1/alpha[1]),beta[1])+ pow((tau2/alpha[2]),beta[2]) +
pow((tau3/alpha[3]),beta[3]) + pow((T/alpha[4]),beta[4])- pow((tau1/alpha[2]),beta[2])
- pow((tau2/alpha[3]),beta[3]) - pow((tau3/alpha[4]),beta[4]))/N)
tau1 ~dunif(300,600)
tau2 ~dunif(1000,1500)
tau3 ~dunif(2200,2350)
alpha[1] ~dunif(0,500)
alpha[2] ~dunif(0,100)
alpha[3] ~dunif(0,500)
alpha[4] ~dunif(0,500)
beta[1] ~dunif(1,3)
beta[2] ~dunif(0,1)
beta[3] ~dunif(1,3)
beta[4] ~dunif(0,1)
}

```

```

list(N=1863,T=2435,t=c(3,345,383,390,386,367,429,432,455,455,448,469,486,
479,484,469,457,467,484,471,483,460,474,476,478,458,
457,503,515,503,501,499,513,512,516,503,507,516,508,
499,497,514,507,495,490,489,493,514,500,488,491,515,
507,501,508,500,487,516,497,496,488,495,504,499,533,
536,536,535,531,523,541,537,522,535,545,535,537,521,
525,525,519,533,535,547,530,532,530,545,520,577,578,
563,577,578,553,571,560,574,556,570,569,576,549,569,
576,567,575,559,551,601,593,589,606,604,595,603,587,
586,595,597,580,587,628,633,613,634,626,612,619,623,
670,661,648,654,656,641,655,655,652,697,693,678,691,
692,695,700,690,699,693,671,697,686,671,696,676,692,
673,690,682,672,674,690,694,677,688,676,695,726,730,
716,705,714,718,715,716,723,713,704,708,724,721,717,
704,717,730,707,717,703,721,711,716,708,723,715,734,
760,742,741,757,750,747,736,749,738,744,740,741,762,
739,748,759,760,756,740,748,739,761,745,756,745,741,
751,733,735,761,750,750,752,741,739,732,750,759,734,
735,754,735,746,762,748,746,738,759,736,761,757,753,
758,750,750,762,760,746,755,757,747,733,733,760,734,
759,746,747,761,759,740,736,762,754,743,745,751,753,
766,773,776,784,765,781,787,781,786,780,777,776,766,
775,788,785,779,783,778,770,778,781,779,770,768,788,

```

783,790,787,781,769,768,767,772,786,770,776,777,781,
 781,786,778,777,778,765,774,776,781,783,773,773,790,
 768,768,789,771,790,771,786,766,788,764,768,772,769,
 776,766,767,763,777,767,783,764,764,765,766,771,766,
 767,785,776,766,767,790,782,767,787,783,785,795,798,
 804,792,811,816,808,807,819,799,817,812,798,807,811,
 793,799,795,810,801,799,797,815,803,800,811,809,820,
 815,798,851,827,846,827,825,824,823,828,846,828,844,
 839,844,839,846,850,843,827,831,850,822,838,847,823,
 848,875,860,867,866,876,854,868,881,863,853,862,861,
 878,875,872,906,910,907,910,897,901,908,905,900,884,
 895,895,900,883,911,909,899,895,905,908,910,905,906,
 907,905,908,885,884,901,907,905,886,898,891,913,925,
 932,922,931,925,918,940,940,941,915,941,934,926,922,
 937,935,922,925,920,937,940,935,943,921,942,914,914,
 916,929,943,941,913,928,915,921,923,928,925,925,919,
 933,941,942,942,937,932,918,930,932,922,917,922,928,
 942,919,916,932,924,919,919,933,938,940,922,943,919,
 913,934,926,916,936,933,927,943,915,922,931,918,929,
 925,918,937,943,933,920,933,921,942,913,930,938,913,
 952,950,957,963,965,969,968,971,958,966,958,953,951,
 954,973,955,962,945,954,946,956,965,964,971,962,958,
 960,953,950,944,956,960,969,974,956,945,969,963,951,
 967,953,962,949,951,953,944,969,964,947,963,959,953,
 970,965,974,950,971,964,967,956,958,958,965,949,973,
 950,945,951,968,971,959,948,965,955,947,973,974,947,
 963,948,951,962,968,959,945,944,973,972,967,955,967,
 963,944,959,956,952,959,955,950,944,955,960,973,948,
 965,947,989,975,1004,980,987,979,991,986,987,987,991,
 996,986,990,1001,1001,975,996,993,991,999,982,975,979,
 986,984,998,1003,991,978,978,989,985,1004,983,990,977,
 978,1000,981,987,977,990,1002,989,990,984,976,988,997,
 997,987,1003,985,1003,1000,993,978,993,1001,990,983,990,
 998,998,985,987,979,991,1028,1029,1012,1023,1026,1006,1007,
 1026,1032,1034,1012,1024,1014,1018,1007,1021,1013,1034,1022,1017,
 1030,1013,1030,1029,1014,1023,1033,1018,1017,1029,1022,1035,1034,
 1013,1022,1034,1005,1010,1021,1031,1025,1005,1007,1027,1008,1015,
 1026,1016,1014,1023,1027,1031,1018,1023,1034,1008,1010,1024,1024,
 1011,1009,1013,1005,1014,1013,1013,1014,1064,1041,1062,1053,1056,
 1061,1038,1049,1056,1036,1062,1049,1059,1056,1041,1058,1050,1064,
 1057,1048,1058,1045,1040,1049,1057,1047,1058,1052,1040,1055,1046,
 1040,1054,1051,1056,1040,1039,1037,1048,1052,1061,1058,1057,1058,
 1048,1058,1044,1049,1039,1049,1044,1058,1045,1038,1061,1044,1065,
 1036,1059,1054,1070,1077,1086,1088,1072,1078,1086,1080,1069,1074,

1091,1077,1069,1069,1074,1068,1067,1074,1094,1094,1078,1076,1080,
 1071,1083,1088,1081,1066,1077,1078,1069,1093,1086,1086,1087,1076,
 1073,1070,1088,1070,1072,1098,1118,1126,1112,1097,1109,1105,1119,
 1114,1118,1106,1111,1125,1112,1122,1116,1117,1113,1126,1101,1118,
 1101,1117,1127,1119,1126,1119,1103,1114,1097,1118,1099,1107,1120,
 1109,1126,1110,1123,1100,1114,1124,1105,1107,1120,1110,1111,1098,
 1097,1108,1119,1123,1115,1107,1107,1122,1110,1117,1115,1098,1118,
 1143,1139,1132,1136,1128,1139,1151,1136,1149,1142,1132,1131,1146,
 1146,1129,1136,1149,1152,1133,1154,1151,1148,1136,1137,1138,1131,
 1139,1134,1146,1151,1151,1136,1150,1142,1141,1136,1139,1141,1145,
 1130,1133,1158,1182,1158,1165,1162,1181,1178,1186,1167,1157,1173,
 1175,1182,1166,1168,1177,1171,1175,1178,1174,1185,1181,1164,1181,
 1173,1177,1186,1182,1173,1178,1180,1168,1158,1173,1175,1158,1177,
 1205,1214,1206,1192,1202,1191,1204,1211,1216,1195,1206,1195,1188,
 1193,1209,1204,1212,1187,1210,1210,1208,1187,1215,1214,1199,1211,
 1201,1214,1206,1189,1190,1207,1216,1209,1193,1190,1212,1205,1187,
 1222,1245,1226,1224,1234,1240,1227,1246,1235,1239,1222,1237,1247,
 1237,1236,1240,1228,1247,1234,1237,1242,1219,1224,1234,1247,1223,
 1266,1276,1254,1274,1254,1252,1272,1252,1257,1306,1319,2238,2236,
 2238,2230,2240,2243,2248,2225,2256,2259,2277,2260,2260,2259,2275,
 2267,2274,2270,2258,2269,2259,2277,2258,2268,2257,2268,2255,2261,
 2256,2262,2270,2263,2268,2261,2262,2255,2261,2273,2274,2263,2271,
 2264,2272,2257,2255,2265,2271,2270,2256,2253,2256,2268,2253,2275,
 2258,2282,2270,2267,2256,2269,2255,2253,2258,2275,2266,2273,2270,
 2267,2267,2263,2269,2272,2263,2277,2266,2266,2263,2256,2277,2260,
 2275,2256,2283,2290,2298,2285,2312,2311,2299,2306,2290,2299,2301,
 2309,2308,2307,2289,2287,2293,2304,2312,2305,2299,2307,2287,2306,
 2300,2305,2284,2312,2284,2307,2288,2311,2296,2302,2298,2302,2290,
 2305,2302,2303,2305,2289,2292,2299,2287,2287,2310,2297,2306,2304,
 2302,2285,2306,2283,2284,2291,2308,2310,2293,2304,2297,2310,2288,
 2298,2293,2295,2291,2304,2304,2297,2308,2303,2311,2296,2303,2292,
 2293,2291,2291,2285,2293,2312,2294,2307,2285,2306,2307,2309,2283,
 2291,2292,2293,2310,2304,2312,2292,2311,2289,2303,2305,2307,2291,
 2312,2296,2308,2299,2291,2297,2297,2293,2302,2286,2305,2303,2284,
 2304,2293,2310,2288,2287,2292,2291,2299,2310,2295,2289,2286,2291,
 2306,2297,2329,2343,2330,2336,2335,2333,2318,2319,2326,2315,2332,
 2315,2334,2314,2320,2335,2342,2320,2322,2342,2320,2343,2322,2313,
 2323,2320,2322,2322,2313,2315,2326,2319,2328,2342,2335,2334,2316,
 2324,2332,2340,2313,2313,2333,2317,2339,2327,2319,2313,2343,2315,
 2322,2334,2334,2337,2313,2342,2338,2336,2315,2322,2319,2329,2339,
 2332,2336,2318,2337,2335,2331,2323,2326,2341,2318,2321,2324,2329,
 2328,2331,2327,2343,2332,2343,2330,2328,2317,2337,2327,2339,2341,
 2329,2325,2339,2331,2314,2342,2333,2332,2321,2326,2338,2340,2330,
 2314,2338,2316,2340,2333,2327,2313,2325,2317,2343,2330,2329,2336,

2343,2343,2313,2315,2343,2319,2339,2330,2321,2317,2341,2341,2336,
 2332,2341,2320,2326,2318,2342,2317,2323,2322,2323,2316,2324,2337,
 2343,2321,2323,2315,2314,2330,2337,2333,2331,2340,2322,2338,2324,
 2342,2334,2339,2339,2318,2333,2341,2329,2343,2322,2343,2315,2337,
 2341,2338,2338,2332,2342,2341,2342,2324,2331,2326,2337,2314,2318,
 2336,2332,2319,2326,2318,2318,2313,2335,2338,2340,2339,2335,2319,
 2313,2314,2340,2338,2319,2330,2327,2317,2343,2333,2343,2318,2338,
 2335,2331,2342,2329,2322,2322,2329,2340,2332,2327,2313,2315,2341,
 2332,2314,2325,2329,2329,2329,2330,2324,2319,2329,2333,2336,2333,
 2339,2313,2338,2333,2327,2315,2331,2338,2323,2323,2313,2319,2333,
 2324,2315,2324,2328,2317,2324,2334,2322,2325,2318,2337,2340,2339,
 2333,2317,2342,2318,2315,2324,2334,2337,2337,2321,2334,2342,2323,
 2334,2315,2328,2343,2321,2335,2314,2337,2337,2341,2328,2371,2369,
 2350,2359,2354,2372,2362,2372,2349,2354,2373,2353,2347,2362,2347,
 2354,2369,2350,2352,2370,2349,2357,2373,2368,2356,2373,2348,2372,
 2353,2357,2351,2348,2349,2353,2373,2360,2366,2368,2368,2357,2362,
 2371,2363,2367,2368,2364,2344,2368,2366,2349,2360,2362,2360,2351,
 2372,2355,2364,2352,2364,2353,2362,2348,2369,2355,2359,2357,2347,
 2372,2348,2346,2354,2358,2363,2361,2348,2352,2360,2348,2356,2353,
 2372,2370,2371,2362,2369,2351,2356,2350,2350,2353,2366,2347,2370,
 2362,2372,2354,2363,2372,2359,2356,2371,2363,2357,2356,2350,2370,
 2368,2365,2347,2348,2346,2371,2364,2353,2347,2373,2350,2361,2357,
 2352,2370,2355,2357,2368,2353,2367,2344,2371,2373,2344,2364,2366,
 2352,2348,2345,2356,2366,2370,2351,2355,2364,2351,2355,2350,2355,
 2369,2367,2353,2354,2372,2360,2360,2363,2346,2369,2346,2361,2372,
 2347,2371,2356,2355,2366,2360,2354,2373,2362,2369,2369,2352,2345,
 2349,2346,2367,2348,2371,2358,2368,2350,2362,2347,2345,2353,2364,
 2402,2390,2389,2401,2382,2388,2393,2384,2388,2399,2387,2382,2401,
 2398,2390,2378,2404,2392,2403,2397,2392,2376,2375,2385,2376,2381,
 2393,2392,2384,2386,2402,2401,2404,2388,2387,2378,2387,2403,2395,
 2374,2374,2391,2376,2386,2386,2391,2375,2392,2374,2396,2386,2403,
 2388,2393,2398,2386,2395,2402,2399,2382,2386,2384,2377,2374,2387,
 2375,2395,2377,2398))

Dados de sarampo na Italia

model

```
{
for (i in 1:N) {
  J[i] <- 1+step(i-k1)+step(i-k2) +step(i-k3) +step(i-k4) +step(i-k5)
  y[i] <- log(measles.Italy[i])
  y[i] ~ dnorm(mean[i],tau[J[i]])
}
for (i in 1:N) {
```


00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,
 0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.0
 0,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00),p3=c(0.00,0.00,0.00,0.00,0.00,
 0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.0
 0,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.10,0.15,0.50,0.15,0.10,0
 .00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00
 ,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.
 00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00),p4=c(0.00,0.0
 0,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0
 .00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00
 ,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.
 00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,
 0.10,0.15,0.50,0.15,0.10,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.0
 0),p5=c(0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,
 0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.0
 0,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.0
 .00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00
 ,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.10,0.15,0.
 50,0.15,0.10,0.00))

Artigo submetido para Statistical Methods and Applications

Noname manuscript No.
(will be inserted by the editor)

Bayesian estimation for one or more change-points for count times series

Ricardo Puziol de Oliveira · Jorge
Alberto Achcar · Charles Chen

Received: date / Accepted: date

Abstract In this study it is introduced a Bayesian approach to analyze count time series data in presence of one or more change-points. This situation is very common in many areas of application, especially considering epidemiology count time series. When the observed count $n(t)$ data are different of zero for each time t , $t = 1, 2, \dots, N$ (number of observed times), it is proposed a statistical model assuming normal distributions for the logarithm of the transformed data, that is, $Y(t) = \log[n(t)]$ in presence of one or more change-points. In situations in presence of count data equals to zero at different times (that is, presence of zero count) the statistical model based on the logarithm transformation for the count data is not suitable in the data analysis. For this case, it is proposed the use of non-homogeneous Poisson processes (NHPP) assuming a PLP (power law process) modeling for its intensity function in presence of change-points. Some examples with real data sets are presented to illustrate the proposed methodology.

Keywords count time series · linear model in presence of change-points · non homogeneous Poisson process in presence of change-points · power law process · Bayesian approach.

Ricardo Puziol de Oliveira
Medical School, University of São Paulo, Ribeirão Preto, SP, Brazil
E-mail: rpuziol.oliveira@gmail.com

Jorge Alberto Achcar
Medical School, University of São Paulo, Ribeirão Preto, SP, Brazil

Charles Chen
Medical School, University of São Paulo, Ribeirão Preto, SP, Brazil

1 Introduction

The presence of one or more change points is very common for time series data especially in epidemiology, economy, environment, medical condition monitoring, climate change detection, speech and image analysis, human activity analysis or other applications (see for example, Yamanishi and Takeuchi, 2002; Lavielle and Teyssiere, 2006; Chopin, 2007; Achcar et al., 2010, 2011; Liu et al., 2013; Jandhyala et al., 2013; Aminikhanghahi and Cook, 2017). Usually there are many time series where the occurrence of some event related to the data could abruptly change the behavior of the time series, that is, there is the presence of one or more change-points. This situation can occur with continuous responses, count responses or other kind of data assuming different parametric or non-parametric models. In the statistical analysis, detection of change points is useful in modeling and prediction of a time series, especially for count data where it is common the use of standard discrete Poisson distributions.

As a great simplification in the data analysis, all discrete observed count could be different of zero in all observed times (days, months or years) where it is possible to assume normal distributions for the logarithm transformed data, $Y(t) = \log(n(t))$, and a standard linear regression model with normal errors in presence of change-points. In other situations when there are the presence of count data equals to zero at specified times (presence of zero counts) the logarithm transformation for the count data is not possible and it is needed to use other statistical approaches in the analysis of time series in presence of one or more change-points. In this case, as a possible modeling approach for the count data it is proposed the use of homogeneous Poisson processes (HPP) or non-homogeneous Poisson processes (NHPP) with specified parametric forms for its intensity functions.

Considering count data modeled by homogeneous Poisson processes (HPP) or non-homogeneous Poisson processes (NHPP) there are many studies introduced in the literature presenting point and interval inferences for the change-points especially under a Bayesian approach (see for example Kuo and Yang, 1996; Achcar et al., 1998; Cid and Achcar, 1999; Achcar, 2001; Pievatolo and Ruggeri, 2004). The use of Bayesian methods using standard MCMC (Markov Chain Monte Carlo) simulation methods (see for example Gelfand and Smith, 1990; Smith and Roberts, 1993; Chib and Greenberg, 1995) usually gives great simplifications to get the inferences of interest given the complexity of the likelihood functions in presence of change-points when compared with other inference approaches as for example, using standard maximum likelihood estimation (MLE) methods. count processes like HPP or NHPP processes have been extensively used by different authors to get inference for change-points (see for example Matthews and Farewell, 1982; Achcar and Bolfarine, 1989; Carlin et al., 1992; Dey and Purkayastha, 1997; Achcar and Loibel, 1998). In special, Raftery and Akman (1986) presented a Bayesian analysis for homogeneous Poisson processes in presence of a change-point. Ruggeri and Sivaganesan (2005) introduced a Bayesian analysis for change-points in NHPPs

considering PLP (power law process) processes assuming a random number of change-points.

In this study, it is assumed a Bayesian modeling approach for the logarithms of the count time series in presence of one or more change-points considering normal errors in situations without the presence of zero count and a NHPP modeling approach in situations in presence of zero count, a common situation for epidemiological time series since a particular disease could appear or disappear after several years without incidence (zero counts during a long time). In the first case (no presence of zero count) the responses of interest (discrete data) are transformed to a logarithmic scale to be modeled by a usual linear regression model in presence of change-points assuming normal errors also including lagged effects of the responses in addition to the time covariates. In the second case (presence of zero count) it is assumed a NHPP modeling approach for situations in presence of one or more change-points.

As a first motivation for the present study, Figure 1 introduces a data set (time series related to monthly counts of visitors to New Zealand from August 1989 to May 2012 where the visit purpose is education). In this application, all monthly count are different of zero, that is, it is possible to model the logarithms of the visitors count in each assumed month. From Figure 1 (original and logarithm scales), it is possible to see the presence of a change-point close to the 35th month (data set with $N = 154$ observations).

Fig. 1 Monthly count time series of visitors to New Zealand from August 1989 to May 2012 where the visit purpose is education (Source: Statistics New Zeland).

As a second motivation for this study, Figure 2 presents the incidences (notification cases) of measles disease in Brazil from January, 2012 to August, 2018 (follow-up of 80 months). The dataset is presented in Table 1. Since measles is an epidemic almost eradicated but that in some times there is the notification of new cases, it is common to have time series with zero count values,

as it is observed in this application (see also Table 1) where it is not possible to model the logarithms of the count, that is, it is needed other statistical modeling approach in the data analysis.

Fig. 2 Measles disease count time series in Brazil from January, 2012 to August, 2018 with a follow-up of 80 months (Source: World Health Organization).

Table 1 Measles disease count (N) in Brazil from January, 2012 to August, 2018 with a follow-up of 80 months (Source: World Health Organization).

Year	N	Year	N	Year	N	Year	N
2012	1	2013	8	2015	26	2017	0
2012	0	2013	9	2015	9	2017	0
2012	0	2013	28	2015	1	2017	0
2012	0	2012	27	2015	1	2017	0
2012	0	2014	79	2015	0	2017	0
2012	0	2014	89	2015	0	2017	0
2012	0	2014	30	2015	0	2017	0
2012	0	2014	25	2015	0	2017	0
2012	0	2014	15	2016	0	2017	0
2012	0	2014	34	2016	0	2017	0
2012	0	2014	93	2016	0	2017	0
2012	1	2014	106	2016	0	2017	0
2013	4	2014	69	2016	0	2018	0
2013	0	2014	67	2016	0	2018	8
2013	5	2014	60	2016	0	2018	74
2013	16	2014	41	2016	0	2018	130
2013	37	2015	60	2016	0	2018	282
2013	25	2015	41	2016	0	2018	184
2013	20	2015	37	2016	0	2018	69
2013	13	2015	39	2016	0	2018	0

The main goal in the first application is related to inferences for the parameters (especially the change-points considered as unknown parameters that should be estimated from the data set) of linear regression models with normal errors in presence of one or more change-points. The inferences of interest are obtained under a Bayesian approach and standard MCMC simulation methods to get the posterior summaries of interest. For the second application, it is proposed the use of NHPP models with a specified parametric PLP form for its intensity function in presence of change-points in the data analysis. The paper is organized as follows: Section 2 introduces the linear regression model in presence of k change points; Section 3 presents the NHPP model in presence of change points. Section 4 introduces four applications. Finally Section 5 presents some discussion on the obtained results.

2 Linear Regression Models in Presence of k Change-points

For the setting of the models, let $N \geq 1$ be a fixed integer number that records the amount of observed data (in the time series introduced in Figure 1, it represents the visit count to New Zealand related to education from August 1989 to May 2012). In our approach it is assumed $Y(t), t = 1, 2, \dots, N$ indicating the times series in the logarithm scale. As a first modeling approach, let us assume the presence of only one unknown change-point ζ where the proposed model is given by a simple linear regression model for the time series with normal errors and a constant variance, that is,

$$Y(t) = \beta_{0j} + \beta_{1j} T(t) + \epsilon(t), \quad (1)$$

for $t = 1, 2, \dots, N$ and $j = 1, 2$ where $Y(t) = \beta_{01} + \beta_{11} T(t) + \epsilon(t)$ for $T(t) \leq \zeta$ (ζ is a change-point that should be estimated from the data) and $Y(t) = \beta_{02} + \beta_{12} T(t) + \epsilon(t)$ for $T(t) > \zeta$; T denotes the covariate *time* and $\epsilon(t)$ is a noise assumed to be independent and identically distributed with a normal distribution $N(0, \sigma_{\epsilon_j}^2), j = 1, 2$ where $\sigma_{\epsilon_1}^2$ for $T(t) \leq \zeta$ and $\sigma_{\epsilon_2}^2$ for $T(t) > \zeta$. In this way there are seven parameters to be estimated: $\beta_{01}, \beta_{11}, \beta_{02}, \beta_{12}, \sigma_{\epsilon_1}^2, \sigma_{\epsilon_2}^2$ and ζ . The change-point ζ is assumed under a Bayesian approach to be a discrete random variable assuming discrete values $1, 2, 3, \dots, N$.

For a Bayesian analysis, it could be assumed a non-informative discrete uniform prior on the integers $1, 2, 3, \dots, N$ or some informative discrete prior distribution considering prior information from experts or obtained from a preliminary data analysis. For example, from the plot of the time series introduced in Figure 1, it is possible to give greater prior mass probabilities for the points close to the month value 35 since this value indicates a possible change-point ζ for the time series. In this way, it is assumed the following independent prior distributions for the parameters of the models: $\zeta \sim$ discrete uniform distribution for the integers $1, 2, 3, \dots, N$; $\Psi_1 = 1/\sigma_{\epsilon_1}^2 \sim \text{Gamma}(a_1, b_1)$; $\Psi_2 = 1/\sigma_{\epsilon_2}^2 \sim \text{Gamma}(a_2, b_2)$; $\beta_{01} \sim N(c_{01}, d_{01})$; $\beta_{02} \sim N(c_{02}, d_{02})$; $\beta_{11} \sim N(c_{11}, d_{11})$; $\beta_{12} \sim N(c_{12}, d_{12})$ where the hyperparameters are assumed known; $\text{Gamma}(a, b)$ denotes a gamma distribution with mean equals to a/b and variance equals to

a/b^2 ; $N(c, d^2)$ denotes a normal distribution with mean equals to c and variance equals to d^2 .

In some situations better modeling fit are obtained considering the presence of lagged effects of the previous responses which implies in a dependence structure among the responses. In this case, the model (1) could be generalized considering a linear regression model in presence of the covariate time T and also including any number n of lagged effects (for example, seasonal effects of 6 or 12 months), that is, the model is given by,

$$\begin{aligned} Y(1) &= \beta_{0j} + \beta_{1j} T(1) + \epsilon(1) \\ Y(2) &= \beta_{0j} + \beta_{1j} T(2) + \beta_{2j} Y(1) + \epsilon(2) \\ &\vdots \\ Y(t) &= \beta_{0j} + \beta_{1j} T(t) + \sum_{l=2}^n \beta_{lj} Y(t-l+1) + \epsilon(t) \end{aligned} \quad (2)$$

where $j = 1, 2$ and $\beta_{01}, \beta_{11}, \dots, \beta_{n1}$ for $T(t) \leq \zeta$; $\beta_{02}, \beta_{12}, \dots, \beta_{n2}$ for $T(t) > \zeta$ and $\epsilon(t)$ is a noise assumed to be independent and identically distributed with a normal distribution $N(0, \sigma_{\epsilon_j}^2)$, $j = 1, 2$ with $\sigma_{\epsilon_1}^2$ for $T(t) \leq \zeta$ and $\sigma_{\epsilon_2}^2$ for $T(t) > \zeta$. In this case, there are $2n + 5$ parameters to be estimated and it is also assumed the same discrete prior distribution for the change-point ζ and the same prior distributions for the other parameters assumed for the model (1).

Now, in presence of k change points, let us assume model (1) for $t = 1, 2, \dots, N$ and $j = 1, 2, \dots, k + 1$ where β_{01} and β_{11} are the regression parameters for the model related to the dataset corresponding to $T(t) \leq \zeta_1$ (ζ_1 is the first change-point); β_{02} and β_{12} are the regression parameters for the model related to the dataset corresponding to $\zeta_1 < T(t) \leq \zeta_2$ (ζ_2 is the second change-point); β_{03} and β_{13} are the regression parameters for the model related to the dataset corresponding to $\zeta_2 < T(t) \leq \zeta_3$ and so on. In addition, $\epsilon_j(t)$ is a noise assumed to be independent and identically distributed with a normal distribution $N(0, \sigma_{\epsilon_j}^2)$, $j = 1, 2, \dots, k + 1$ where $\sigma_{\epsilon_1}^2$ for $T(t) \leq \zeta_1$; $\sigma_{\epsilon_2}^2$ for $\zeta_1 < T(t) \leq \zeta_2$; $\sigma_{\epsilon_3}^2$ for $\zeta_2 < T(t) \leq \zeta_3$, $\dots$, $\sigma_{\epsilon_{k+1}}^2$ for $\zeta_k < T(t) \leq T$ (follow-up time). In this case, there are $(k + 1)(n + 2) + k$ parameters to be estimated.

3 Use of Non-homogeneous Poisson Processes (NHPP) in Presence of Change-points

In some applications, especially in epidemiological times series it is possible to have many times with zero occurrence of a specified disease, that is, many zero values. In the epidemiology field, this fact may be due to the eradication of the disease by application of vaccines to the susceptible population or to improvements in health conditions and a possible recidivism could be due to lack of vaccines or financial resources for health. In this way it is not possible to use the logarithm transformation of the disease count in the statistical

data analysis and it is proposed the use of models based on non-homogeneous Poisson Processes (NHPP) in presence of change points. As a motivation for this approach Figure 2 presents the incidence (notification cases) of measles disease in Brazil from January, 2012 to August, 2018 (follow-up of 80 months). The dataset is also presented in Table 1.

3.1 The Likelihood Function

In the construction of the likelihood function, let us assume without loss of generality, the incidence data of measles in Brazil introduced in Table 1 and Figure 2. The incidence (notification cases) of measles disease in Brazil between January, 2012 to August, 2018 (80 months) presents three significant change-points close to the the months 17, 23 and 29 (see Figure 1 and Table 1). To model the number of measles in Brazil in this period, it is considered the use of a point process to the monthly cumulative count of measles starting in January, 2012. Since it is not observed the exact days of occurrence (only it is observed the count of measles in each month) it is used an uniform distribution to simulate the times (in days) for the measles occurrence in each month from January, 1 of 2012 to December, 30 of 2018, using only the month count information, where it is observed a total time of observation equals to $T = 2435$ days (see Figure 3).

From the cumulative plot of Figure 3, it is possible to see three change-points close to 500th day, 1200th day and 2200th day in the follow-up period of $T = 2435$ days. For the data analysis, it is proposed the use of a non-homogeneous Poisson process (NHPP) with a PLP intensity function in the presence of three change-points under a Bayesian approach using Markov chain Monte Carlo (MCMC) methods (see for example, Gelfand and Smith, 1990).

In the construction of the likelihood function, let $N(t)$ be the cumulative number of measles count that are observed during the interval $(0, t)$ and assume that $N(t)$ is modeled by a NHPP (Non-Homogeneous Poisson Process) with intensity function $\lambda(t) = dm(t)/dt = dE[N(t)]/dt$, where $m(t)$ is the mean value function. Different parametric forms could be assumed for the intensity function $\lambda(t)$ (increasing, decreasing, bathtub shape, unimodal, among many others, see for example, Musa and Okumoto, 1984 or Mudholkar et al., 1995). In this study, it is assumed power-law processes (PLP) in the presence of three change-points with intensity function for the overall process given by,

$$\lambda(t | \theta) = \begin{cases} \lambda_1 = \frac{\beta_1}{\alpha_1} \left(\frac{t}{\alpha_1} \right)^{\beta_1-1} & \text{if } 0 < t < \zeta_1 \\ \lambda_2 = \frac{\beta_2}{\alpha_2} \left(\frac{t}{\alpha_2} \right)^{\beta_2-1} & \text{if } \zeta_1 \leq t < \zeta_2 \\ \lambda_3 = \frac{\beta_3}{\alpha_3} \left(\frac{t}{\alpha_3} \right)^{\beta_3-1} & \text{if } \zeta_2 \leq t < \zeta_3 \\ \lambda_4 = \frac{\beta_4}{\alpha_4} \left(\frac{t}{\alpha_4} \right)^{\beta_4-1} & \text{if } t \geq \zeta_3 \end{cases} \quad (3)$$

Fig. 3 Histogram and cumulative plot for the daily count of measles disease (simulated from information of Table 1) in Brazil from January, 2012 to August, 2018 (Source: World Health Organization).

where $\theta = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \beta_4, \zeta_1, \zeta_2, \zeta_3)$ is the vector of parameters. Equivalently, letting $m_j(t) = m(t|\theta_j), j = 1, 2, 3, 4$, the corresponding mean value function is given by,

$$m(t|\theta) = \begin{cases} m_1(t) & \text{if } 0 < t < \zeta_1 \\ m_2(t) - m_2(\zeta_1) + m_1(\zeta_1) & \text{if } \zeta_1 \leq t < \zeta_2 \\ m_3(t) - m_3(\zeta_2) + m_2(\zeta_2) - m_2(\zeta_1) + m_1(\zeta_1) & \text{if } \zeta_2 \leq t < \zeta_3 \\ m_4(t) - m_4(\zeta_3) + m_3(\zeta_3) - m_3(\zeta_2) + m_2(\zeta_2) - m_2(\zeta_1) + m_1(\zeta_1) & \text{if } t \geq \zeta_3 \end{cases} \quad (4)$$

where $m_l = (t/\alpha_l)\beta_l, l = 1, 2, 3, 4$. Observe that the intensity function $\lambda_j(t)$ in equation (3) is constant for $\beta_j = 1$ decreases for $\beta_j < 1$ and increases for $\beta_j > 1, j = 1, 2, 3, 4$. This process is related to the Weibull probability model with parameters $(\alpha_j, \beta_j), j = 1, 2, 3, 4$.

Assuming that the data are observed up to a total time T , where the epochs of occurrence of cases are denoted by $t_i, i = 1, \dots, n; 0 < t_1 < t_2 < \dots < t_{N(\zeta_1)} < t_{N(\zeta_1)+1} < \dots < t_{N(\zeta_2)} < t_{N(\zeta_2)+1} < \dots < t_{N(\zeta_3)} < t_{N(\zeta_3)+1} < \dots < t_n < T$, the likelihood function for θ in presence of three change-points ζ_1, ζ_2 and ζ_3 is given by,

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{N(\zeta_1)} \lambda_1(t_i) \exp\{-m_1(\zeta_1)\} \prod_{i=N(\zeta_1)+1}^{N(\zeta_2)} \lambda_2(t_i) \exp\{-m_2(\zeta_2) + m_2(\zeta_1)\} \\
&\times \prod_{i=N(\zeta_2)+1}^{N(\zeta_3)} \lambda_3(t_i) \exp\{-m_3(\zeta_3) + m_3(\zeta_2)\} \\
&\times \prod_{i=N(\zeta_3)+1}^{N(T)} \lambda_4(t_i) \exp\{-m_4(T) + m_4(\zeta_3)\} \quad (5)
\end{aligned}$$

where $\lambda_j(t)$ is given in equation (3) and $m_j(t)$ is given in equation (4).

Remark 1. In the likelihood function (5), observe that $N(s+t) - N(s)$ given θ has a Poisson distribution $P(m(s+t|\theta) - m(s|\theta))$ for $t > 0$ and independent increments. Thus, the sampling distribution for the between occurrence times, say U_i , has density,

$$f_{U_{1|\theta}}(t) = \lambda(t|\theta) \exp\{-m(t|\theta)\} \quad (6)$$

and,

$$f_{U_{2|U_1}}(t) = \lambda(s+t|\theta) \exp\{-m(s+t|\theta) + m(s|\theta)\} \quad (7)$$

and so on. In this way, it is obtained the likelihood for the data $D_T = \{n; t_1, \dots, t_{N(\zeta_1)}, t_{N(\zeta_1)+1}, \dots, t_{N(\zeta_2)}, t_{N(\zeta_2)+1}, \dots, t_{N(\zeta_3)}, t_{N(\zeta_3)+1}, \dots, t_n, T\}$ in the presence of three change-points.

3.2 A Bayesian Analysis

For a Bayesian analysis of the NHPP process model with PLP intensity function given in equation (3) in presence of three change-points ζ_1, ζ_2 and ζ_3 , it is assumed uniform prior distributions for α_j and β_j , given by $\alpha_j \sim U(0, a_j)$ and $\beta_j \sim U(b_{1j}, b_{2j})$ for $j = 1, 2, 3, 4$ where a_j, b_{1j} and b_{2j} are known hyperparameters. Let us also assume informative uniform prior distributions $\zeta_k \sim U(c_{1k}, c_{2k}), k = 1, 2, 3$ for the change-points where c_{1k} and c_{2k} are also known hyperparameters. It is further considered prior independence among the parameters. Markov Chain Monte Carlo (MCMC) methods such as the Gibbs sampling algorithm or the Metropolis-Hastings algorithm are assumed to simulate samples for the joint posterior distribution of interest. In this case, the steps of the Gibbs sampling algorithm are given by,

- Step 1: Choose initial values $\zeta_k^{(0)}, k = 1, 2, 3$ and $\alpha_j^{(0)}, \beta_j^{(0)}, j = 1, 2, 3, 4$, respectively. Denote the values of $\zeta_k, k = 1, 2, 3$ and $\alpha_j, \beta_j, j = 1, 2, 3, 4$ at the i th step by $\zeta_k^{(i)}, k = 1, 2, 3$ and $\alpha_j^{(i)}, \beta_j^{(i)}, j = 1, 2, 3, 4$;
- Step 2: Generate $\zeta_1^{(i)}, \zeta_k^{(i+1)}, k = 2, 3$ and $\alpha_j^{(i+1)}, \beta_j^{(i+1)}, j = 1, 2, 3, 4$ from the conditional posterior distributions and repeat this step N times;
- Step 3: Calculate the Monte Carlo Bayesian estimate of $\mu(\alpha, \beta, \zeta)$ by $\sum_{i=B+1}^N \mu(\alpha^{(i)}, \beta^{(i)}, \zeta^{(i)}) / (N - B)$ where B is the burn-in period.

This modeling approach can be generalized to other NHPP in presence of any number of change-points and any other parametric form for the intensity function.

4 Applications

For all Gibbs simulation approaches considered in the first three applications, it is considered a burn-in sample of size 11,000 to eliminate the effects of the initial values assumed in the iterative approach and 1,000 additional Gibbs samples (every 100th sample from 100,000 generated Gibbs samples). The generation of samples for the joint posterior distribution is obtained using MCMC methods (Gibbs sampling or Metropolis-Hastings algorithms) and the OpenBugs software (Spiegelhalter et al., 2003). Convergence of the algorithm is verified from trace plots of the simulated samples.

4.1 New Zealand Education Visits in the Period from August 1989 to May 2012

Let us assume as a first application the data set (time series related to monthly number of visitors to New Zealand from August 1989 to May 2012 where the visit purpose is education) given in Figure 1. It is assumed the data in logarithm scale, that is, $Y(t) = \log(\text{number of visitors in the month } t)$. From Figure 1, it is observed the presence of only one change-point ζ . Since the count time series presents values different of zero for all times (months) it is possible to assume a linear regression model for the logarithms of the data set in presence of one change-point (see Section 2). In this way, assuming the linear regression models (1) and (2) denoted here as models 1 and 2 with time given by months in presence of a change-point ζ is assumed the following independent prior distributions for the parameters of the models: $\zeta \sim \text{uniform discrete uniform distribution for the integers } 1, 2, 3, \dots, 154$; $\Psi_1 = 1/\sigma_{\epsilon_1}^2 \sim \text{Gamma}(1, 1)$; $\Psi_2 = 1/\sigma_{\epsilon_2}^2 \sim \text{Gamma}(1, 1)$; $\beta_{0j} \sim N(5, 100)$, $j = 1, 2$; and $\beta_{ij} \sim N(0, 1)$, $i = 1, 2, \dots, 6$, $j = 1, 2$. Table 2 presents the posterior summaries of interest. In addition, Figure 4 and 5 present the plots of the observed time series, the fitted means and the histogram of the Gibbs simulated sample of change-points assuming models 1 and 2. From these plots, it is observed that model 2 is better fitted by the data set.

In order to decide by the best model, it is used some measures (MAPE, MAD and MSD introduced in Appendix 1) for the evaluation of the fit of the time series models. The results of selected criteria are presented in Table 3 where it could be concluded that the model 2 is again better fitted by the data set.

Fig. 4 Fitted data assuming models 1 (left panel) and 2 (right panel) for the monthly count time series of visitors to New Zealand from August 1989 to May 2012.

Fig. 5 Histogram for the simulated change-points for model 1 (left panel) and model 2 (right panel)

4.2 Incidence of Tuberculosis in New York City in the Period 1970-2000

Let us assume as a second application the time series related to reported year incidences of tuberculosis (TB) in New York City between the years of 1970 and 2000. This data set presents three trends (see Figure 6): a first period (1970-1979) where the trend of declining incidence was probably associated with good control programs; a second period (1979-1992) where there was an increase in incidence rates possibly associated by the HIV epidemic; a third period (1992-2000) where again there is a decline in incidence rates (see Achcar et al., 2008) possible related to the beginning of better therapies for AIDS.

Table 2 Posterior summaries for model 1 and model 2 for the monthly number of visitors to New Zealand from August 1989 to May 2012.

	Param.	Post. Mean (S.D.)	95% Cred. Int.
Model 1	β_{01}	7.3333 (0.2872)	(6.7100, 7.8600)
	β_{02}	9.3855 (0.0508)	(9.2780, 9.4820)
	β_{11}	0.1104 (0.0320)	(0.0629, 0.1934)
	β_{12}	0.0009 (0.0005)	(-0.0001, 0.0020)
	ζ	20.6770 (2.4136)	(14.0000, 25.0000)
	Ψ_1	3.8580 (1.2496)	(1.7960, 6.6540)
	Ψ_2	26.1801 (3.3462)	(19.9597, 33.0800)
Model 2	β_{01}	6.9762 (0.3136)	(6.3500, 7.5880)
	β_{02}	8.6834 (0.1117)	(8.4730, 8.9120)
	β_{11}	0.1403 (0.4150)	(0.0605, 0.2235)
	β_{12}	0.0002 (0.0003)	(-0.0005, 0.0010)
	β_{21}	9.4E-05 (3.4E-05)	(2.5E-05, 0.0002)
	β_{22}	9.0E-05 (1.2E-05)	(6.5E-05, 0.0001)
	β_{31}	-6.3E-05 (3.6E-05)	(-0.0001, 8.1E-06)
	β_{32}	-6.5E-05 (1.8E-05)	(-0.0001, -2.9E-05)
	β_{41}	4.6E-05 (3.7E-05)	(-2.8E-05, 0.0001)
	β_{42}	4.8E-05 (1.9E-05)	(1.16E-05, 8.8E-05)
	β_{51}	-2.2E-05 (3.6E-05)	(-9.4E-05, 5.0E-05)
	β_{52}	-1.5E-05 (1.8E-05)	(-5.2E-05, 1.9E-05)
	β_{61}	-1.9E-05 (2.7E-05)	(-7.2E-05, 3.5E-05)
	β_{62}	5.4E-07 (1.2E-05)	(-2.2E-05, 2.5E-05)
	ζ	15.6822 (1.7788)	(14.0000, 19.0000)
	Ψ_1	4.2664 (1.5892)	(1.7410, 7.8910)
	Ψ_2	41.6983 (5.2187)	(32.0797, 52.5400)

Table 3 MAPE, MAD and MSD values for each proposed model for the New Zealand data.

Criteria	Model 1	Model 2
MAPE	1.5561	0.9126
MAD	0.1399	0.0816
MSD	0.0393	0.0150

From Figure 6 (original and logarithm scales), it is possible to see the presence of two change-points close to the 10th year and 22th year.

For the Bayesian analysis of this data set it is considered the linear regression models in presence of two change-points ζ_1 and ζ_2 in presence or not of the six lagged effects of previous incidences since all yearly measles count are different of zero. Considering only the covariate year for the model 1, it is assumed the following independent prior distributions for the parameters of the model: $\zeta_1 \sim$ discrete informative distribution for the integers $1, 2, 3, \dots, 31$ concentrated in the value 10 (greater probability mass); $\zeta_2 \sim$ discrete informative

Fig. 6 Incidence of tuberculosis in NYC in the period from 1970 to 2000.

distribution for the integers $1, 2, 3, \dots, 31$ concentrated in the value 22 (more probability mass); $\Psi_1 = 1/\sigma_{\epsilon_1}^2 \sim \text{Gamma}(1, 1)$; $\Psi_2 = 1/\sigma_{\epsilon_2}^2 \sim \text{Gamma}(1, 1)$; $\Psi_3 = 1/\sigma_{\epsilon_3}^2 \sim \text{Gamma}(1, 1)$; $\beta_{0j} \sim N(7, 1)$, $j = 1, 2, 3$; and $\beta_{1j} \sim N(0, 1)$, $j = 1, 2, 3$. Moreover, it was considered more informative prior distributions for the parameters of the model, especially for the two change-points to get convergence of the Gibbs sampling algorithm.

The posterior summaries of interest considering model 1 are presented in Table 5. Table 5 also presents the posterior summaries considering a second model given by 2 assuming the presence of six lagged effects and the same prior distributions as considered in the case of only one covariate (year) that is, $\zeta_1 \sim$ discrete informative distribution for the integers $1, 2, 3, \dots, 31$ concentrated in the value 10 (greater probability mass); $\zeta_2 \sim$ discrete informative distribution for the integers $1, 2, 3, \dots, 31$ concentrated in the value 22 (more probability mass); $\Psi_1 = 1/\sigma_{\epsilon_1}^2 \sim \text{Gamma}(1, 1)$; $\Psi_2 = 1/\sigma_{\epsilon_2}^2 \sim \text{Gamma}(1, 1)$; $\Psi_3 = 1/\sigma_{\epsilon_3}^2 \sim \text{Gamma}(1, 1)$; $\beta_{0j} \sim N(7, 1)$, $j = 1, 2, 3$; and $\beta_{ij} \sim N(0, 1)$, $i = 1, \dots, 6$, $j = 1, 2, 3$.

Figure 7 presents the plots of the observed time series and the fitted means assuming model 1 and model 2. From these plots, it is observed that model 2 is better fitted by the data set. Figure 8 also presents the histogram of the Gibbs simulated samples of change-point assuming model 2. In addition, from the results of Table 5, the change points are estimated by $\zeta_1 = 8.5923$ (close to 9) and $\zeta_2 = 26.2955$ (close to 25) assuming model 2 (that is, the change points are estimated in the year 1978 which corresponds to the beginning of the HIV epidemic and other change-point estimated in the year 1994 where there is the beginning of declining count of TB in NYC - beginning of new drugs for HIV epidemic). These results are in agreement with the results obtained by Achcar et al. (2008) assuming non-homogeneous Poisson processes (NHPP) in the presence of the two-change points presented in Section 3 under a Bayesian

approach using MCMC methods. It is important to point out that the proposed methodology gives a great simplicity of modeling epidemiological time series when compared with the use of a NHPP modeling approach.

Fig. 7 Fitted data assuming models 1 (left panel) and 2 (right panel) for the incidence of tuberculosis in NYC in the period from 1970 to 2000.

Fig. 8 Histogram for the simulated change-point ζ_1 (left panel) and change-point ζ_2 (right panel) assuming model 2.

In order to decide by the best model, it is used MAPE, MAD and MSD measures for the evaluation of the fit of the time series models. The results of selected criteria are presented in Table 5, where it could be concluded that the model 2 is better fitted by the data set.

Table 4 MAPE, MAD and MSD values for each proposed model for the incidence of tuberculosis in NYC data.

Criteria	Model 1	Model 2
MAPE	1.3687	0.7892
MAD	0.1052	0.0611
MSD	0.0153	0.0089

4.3 Incidence of Measles in Italy in the Period from January, 2011 to June, 2018

To illustrate the application of the proposed methodology, let us also assume a third application in presence of five change-points (models introduced in Section 2). In this way, let us assume the time series related to month report of incidence of measles in Italy for the period ranging from January, 2011 to June, 2018 (data obtained from the World Health Organization). From Figure 9 it is possible to see the presence of five change-points close to the 5h month; 30th month; 38th month; 75th month and 88th month (data set with $N = 90$ observations).

Fig. 9 Measles disease count time series in Italy from January, 2011 to June, 2018 with a follow-up of 90 months (Source: World Health Organization).

Since in this case, the dataset does not present zero count, it is possible to assume the linear models introduced in section 2 considering the logarithms of the measles count reported for each month in Italy for the period ranging from January, 2011 to June, 2018 in presence of five change-points. Considering the covariate month and five lagged effects of the responses, it is assumed the following independent prior distributions for the parameters of the model: $\zeta_1 \sim$

Table 5 Posterior summaries for model 1 (only covariate year) and model 2 (presence of lagged effects of previous incidences).

	Param.	Post. Mean (S.D.)	95% Cred. Int.
Model 1	β_{01}	7.7502 (0.4308)	(6.8350, 8.5560)
	β_{02}	6.6156 (0.3797)	(5.8600, 7.3520)
	β_{03}	8.4377 (0.9599)	(6.5020, 10.2600)
	β_{11}	-0.0209 (0.0345)	(-0.1981, 0.1903)
	β_{12}	0.0637 (0.0228)	(0.0192, 0.1092)
	β_{13}	-0.0326 (0.0349)	(-0.0991, 0.0375)
	ζ_1	8.7593 (1.0588)	(6.0000, 11.0000)
	ζ_2	24.6699 (1.3325)	(22.0000, 27.0000)
	Ψ_1	3.8297 (1.9796)	(0.9616, 8.5600)
	Ψ_2	7.0740 (2.5454)	(2.9870, 12.8400)
	Ψ_3	3.4981 (1.7868)	(0.8930, 7.7750)
Model 2	β_{01}	7.5355 (0.5933)	(6.2970, 8.6550)
	β_{02}	6.6417 (0.7009)	(5.2500, 8.0390)
	β_{03}	8.0085 (0.9972)	(6.0490, 9.9690)
	β_{11}	-0.0227 (0.1724)	(-0.3868, 0.3071)
	β_{12}	0.0540 (0.0911)	(-0.1311, 0.2335)
	β_{13}	-0.0081 (0.1080)	(-0.2156, 0.2010)
	β_{21}	0.0002 (0.0003)	(-0.0004, 0.0008)
	β_{22}	0.0002 (0.0004)	(-0.0005, 0.0010)
	β_{23}	0.0002 (0.0014)	(-0.0023, 0.0028)
	β_{31}	2.8E-05 (0.0003)	(-0.0007, 0.0007)
	β_{32}	-0.0001 (0.0004)	(-0.0010, 0.0008)
	β_{33}	-0.0002 (0.0014)	(-0.0025, 0.0022)
	β_{41}	-3.8E-05 (0.0003)	(-0.0007, 0.0006)
	β_{42}	-0.0001 (0.0004)	(-0.0009, 0.0009)
	β_{43}	-0.0001 (0.0014)	(-0.0025, 0.0022)
	β_{51}	-1.4E-05 (0.0002)	(-0.0003, 0.0003)
	β_{52}	-3.1E-05 (0.0002)	(-0.0005, 0.0005)
	β_{53}	-0.0001 (0.0013)	(-0.0024, 0.0021)
	β_{61}	0.0032 (0.9976)	(-1.9520, 1.9560)
	β_{62}	-0.0017 (1.0003)	(-1.9690, 1.9650)
	β_{63}	0.0001 (1.0037)	(-1.9720, 1.9750)
	ζ_1	8.5923 (1.1624)	(6.0000, 11.0000)
	ζ_2	26.2955 (1.1953)	(24.0000, 27.0000)
	Ψ_1	2.6233 (1.5539)	(0.5253, 6.4230)
	Ψ_2	5.9281 (2.3709)	(2.2260, 11.4100)
	Ψ_3	1.2920 (1.2342)	(0.0406, 4.5760)

discrete informative distribution for the integers $1, 2, 3, \dots, 90$ concentrated in the value 5 (more probability mass); $\zeta_2 \sim$ discrete informative distribution for the integers $1, 2, 3, \dots, 90$ concentrated in the value 30 (more probability mass); $\zeta_3 \sim$ discrete informative distribution for the integers $1, 2, 3, \dots, 90$

concentrated in the value 38 (more probability mass); $\zeta_4 \sim$ discrete informative distribution for the integers $1, 2, 3, \dots, 90$ concentrated in the value 75 (more probability mass); $\zeta_5 \sim$ discrete informative distribution for the integers $1, 2, 3, \dots, 90$ concentrated in the value 88 (more probability mass); $\Psi_j = 1/\sigma_{\epsilon_j}^2 \sim \text{Gamma}(1, 1)$; $\beta_{0j} \sim N(7, 1)$, $\beta_{1j} \sim N(0, 1)$, $j = 1, 2, 3, 4, 5, 6$. In this application, it was used more informative prior distributions for the parameters of the model, especially for the two change-points to get convergence of the Gibbs sampling algorithm. Table 6 presents the posterior summaries of interest. In addition, Figure 10 presents the plots of the observed time series and the fitted means assuming the proposed model.

Fig. 10 Observed data and fitted means for the measles disease count in Italy from January, 2011 to June, 2018 with a follow-up of 90 months.

4.4 Incidence of Measles in Brazil in the Period from January, 2011 to June, 2018

In this application it is considered the incidence (notification cases) of measles disease in Brazil from January, 2012 to August, 2018 (follow-up of 80 months). The dataset is presented in Table 1 and Figure 2. From Table 1, it is observed the presence of a great number of zero count in the reported incidences of measles in Brazil. In this way, it is assumed the use of NHPP modeling approach introduced in Section 3. Assuming the NHPP with intensity function given by (3) it is assumed the following independent prior distributions for the parameter of the models:

- Change-points: $\zeta_1 \sim U(300, 600)$, $\zeta_2 \sim U(1000, 1500)$ and $\zeta_3 \sim U(2200, 2350)$
- Scale parameters: $\alpha_1 \sim U(0, 500)$, $\alpha_2 \sim U(0, 100)$, $\alpha_3 \sim U(0, 500)$ and $\alpha_4 \sim U(0, 500)$. Shape parameters: $\beta_1 \sim U(1, 3)$, $\beta_2 \sim U(0, 1)$, $\beta_3 \sim U(1, 3)$, and $\beta_4 \sim U(0, 1)$.

Table 6 Posterior summaries for model 1(only covariate month).

Param.	Post. Mean (S.D.)	95% Cred. Int.
β_{01}	6.1118 (0.7077)	(4.8720, 7.6860)
β_{02}	6.0774 (0.6513)	(4.7770, 7.3160)
β_{03}	6.7984 (0.9947)	(4.8680, 8.7460)
β_{04}	5.9838 (0.8068)	(4.4130, 7.5750)
β_{05}	7.4359 (0.9808)	(5.5070, 9.3730)
β_{06}	6.9828 (0.9932)	(5.0640, 8.9380)
β_{11}	0.0779 (0.3241)	(-0.7746, 0.5591)
β_{12}	-0.1104 (0.0379)	(-0.1838, -0.0355)
β_{13}	-0.0479 (0.0342)	(-0.1163, 0.0172)
β_{14}	-0.0360 (0.0141)	(-0.0638, -0.0087)
β_{15}	-0.0213 (0.0127)	(-0.0462, 0.0036)
β_{16}	-0.0125 (0.0127)	(-0.0372, 0.0117)
ζ_1	5.6807 (1.2520)	(3.0000, 7.0000)
ζ_2	28.3026 (0.6201)	(28.0000, 30.0000)
ζ_3	38.5736 (1.3484)	(36.0000, 40.0000)
ζ_4	73.1087 (0.3469)	(73.0000, 74.0000)
ζ_5	86.9629 (1.0631)	(85.0000, 89.0000)
ψ_1	2.1453 (1.4293)	(0.3015, 5.7150)
ψ_2	0.7401 (0.2292)	(0.3685, 1.2570)
ψ_3	0.9572 (0.4681)	(0.3171, 2.1340)
ψ_4	1.1099 (0.3064)	(0.6147, 1.8060)
ψ_5	1.4692 (0.5552)	(0.5908, 2.7330)
ψ_6	2.2714 (1.4911)	(0.2953, 5.9500)

Observe that the hyperparameters for the uniform prior distributions for the shape parameters were chosen to have respective increasing, decreasing, increasing and decreasing forms for the intensity function (see Figure 3). For the Gibbs simulation approach considered in this application, it was considered a burn-in sample of size 1,000 to eliminate the effects of the initial values assumed in the iterative approach and 1,000 additional Gibbs samples. The generation of samples for the joint posterior distribution was obtained using MCMC methods and the OpenBugs software. Convergence of the algorithm is verified from trace plots of the simulated samples. Table 7 presents the posterior summaries of interest. The Monte Carlo estimate of mean value function based on the 1000 generated Gibbs samples is given by,

$$m(t | \theta) = \begin{cases} (t/139.5000)^{1.7620} & \text{if } 0 < t < 454.1000 \\ (t/0.6705)^{0.9891} - 622.8072 & \text{if } 454.1000 \leq t < 1256.0000 \\ (t/312.8000)^{1.5850} + 1093.6650 & \text{if } 1256.0000 \leq t < 2252.0000 \\ (t/0.1395)^{0.9368} - 7805.5530 & \text{if } 2252.0000 \leq t < T \end{cases} \quad (8)$$

Table 7 Posterior summaries for the NHPP model(three change-points).

Param.	Post. Mean (S.D.)	95% Cred. Int.
α_1	139.5000 (47.5400)	(58.2300, 245.0000)
α_2	0.6705 (0.0611)	(0.5145, 0.7577)
α_3	312.8000 (119.0000)	(88.8800, 490.6000)
α_4	0.1395 (0.0615)	(0.0204, 0.2471)
β_1	1.7620 (0.4413)	(1.0740, 2.7860)
β_2	0.9891 (0.0110)	(0.9599, 0.9996)
β_3	1.5850 (0.2829)	(1.0600, 2.1080)
β_4	0.9368 (0.0495)	(0.8048, 0.9980)
ζ_1	454.1000 (2.5940)	(451.7000, 456.4000)
ζ_2	1256.0000 (1.8660)	(1254.0000, 1259.0000)
ζ_3	2252.0000 (0.2510)	(2252.0000, 2252.0000)

Figure 11, presents the accumulated numbers of reported measles in in each time (day) occurrence from the observed data and using the estimated mean value function.

Fig. 11 Observed accumulated numbers of measles in in each time (day) occurrence and the estimated mean value function (measles count in Brazil).

5 Discussion of the Results

The results of this study could be very useful, especially to model epidemiological time series in presence of one or more change-points. It is important to point out that the times series modeling approach proposed in this study could be given by a logarithm transformation of the time series (count) when it is not observed the presence of zero count (Section 2). In this way, to apply

this simple proposed methodology it is needed time series with values different of zero.

On other hand, in some cases, especially with epidemic time series with rare occurrences, it is possible to have many zero count (days, months or years) where there is no occurrence of cases of the disease. For this case, it is proposed to assume NHPP modeling approach in presence of one or more change-points introduced in Section 3, where it is needed the exact time (calendar time) of each individual disease occurrence to model the accumulated number of occurrences in each time. It is important to point out that in both proposed modeling approaches the estimation of change-points could be obtained with good accuracy using MCMC methods under a Bayesian approach as it was observed in the four applications presented in Section 4.

References

- Achcar J, Martinez E, Ruffino-Netto A, Paulino C, Soares P (2008) A statistical model investigating the prevalence of tuberculosis in New York City using counting processes with two change-points. *Epidemiology & Infection* 136(12):1599–1605
- Achcar JA (2001) Bayesian analysis for software reliability data. In: *Advances in Reliability, Handbook of Statistics*, vol 20, Elsevier, pp 733 – 748, DOI [https://doi.org/10.1016/S0169-7161\(01\)20031-5](https://doi.org/10.1016/S0169-7161(01)20031-5), URL <http://www.sciencedirect.com/science/article/pii/S0169716101200315>
- Achcar JA, Bolfarine H (1989) Constant hazard against a change-point alternative: a Bayesian approach with censored data. *Communications in Statistics-Theory and Methods* 18(10):3801–3819
- Achcar JA, Loibel S (1998) Constant hazard function models with a change point: A Bayesian analysis using markov chain monte carlo methods. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 40(5):543–555
- Achcar JA, Dey DK, Niverthi M (1998) A Bayesian approach using nonhomogeneous Poisson processes for software reliability models. In: *Frontiers in reliability*, World Scientific, pp 1–18
- Achcar JA, Rodrigues ER, Paulino CD, Soares P (2010) Non-homogeneous Poisson models with a change-point: an application to ozone peaks in mexico city. *Environmental and Ecological Statistics* 17(4):521–541
- Achcar JA, Rodrigues ER, Tzintzun G (2011) Using non-homogeneous Poisson models with multiple change-points to estimate the number of ozone exceedances in mexico city. *Environmetrics* 22(1):1–12
- Aminikhanghahi S, Cook DJ (2017) A survey of methods for time series change point detection. *Knowledge and information systems* 51(2):339–367
- Carlin BP, Gelfand AE, Smith AF (1992) Hierarchical Bayesian analysis of changepoint problems. *Applied statistics* pp 389–405
- Chib S, Greenberg E (1995) Understanding the Metropolis-Hastings algorithm. *The american statistician* 49(4):327–335

- Chopin N (2007) Dynamic detection of change points in long time series. *Annals of the Institute of Statistical Mathematics* 59(2):349–366
- Cid JER, Achcar JA (1999) Bayesian inference for nonhomogeneous Poisson processes in software reliability models assuming nonmonotonic intensity functions. *Computational Statistics & Data Analysis* 32(2):147–159
- Dey DK, Purkayastha S (1997) Bayesian approach to change point problems. *Communications in Statistics-Theory and Methods* 26(8):2035–2047
- Gelfand AE, Smith AF (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association* 85(410):398–409
- Jandhyala V, Fotopoulos S, MacNeill I, Liu P (2013) Inference for single and multiple change-points in time series. *Journal of Time Series Analysis* 34(4):423–446
- Kuo L, Yang TY (1996) Bayesian computation for nonhomogeneous Poisson processes in software reliability. *Journal of the American Statistical Association* 91(434):763–773
- Lavielle M, Teyssiere G (2006) Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal* 46(3):287–306
- Liu S, Yamada M, Collier N, Sugiyama M (2013) Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks* 43:72–83
- Matthews DE, Farewell VT (1982) On testing for a constant hazard against a change-point alternative. *Biometrics* pp 463–468
- Mudholkar GS, Srivastava DK, Freimer M (1995) The exponentiated Weibull family: A reanalysis of the bus-motor-failure data. *Technometrics* 37(4):436–445
- Musa JD, Okumoto K (1984) A logarithmic Poisson execution time model for software reliability measurement. In: *Proceedings of the 7th international conference on Software engineering*, IEEE Press, pp 230–238
- Pievatolo A, Ruggeri F (2004) Bayesian reliability analysis of complex repairable systems. *Applied Stochastic Models in Business and Industry* 20(3):253–264
- Raftery AE, Akman V (1986) Bayesian analysis of a Poisson process with a change-point. *Biometrika* pp 85–89
- Ruggeri F, Sivaganesan S (2005) On modeling change points in non-homogeneous Poisson processes. *Statistical Inference for Stochastic Processes* 8(3):311–329
- Smith AF, Roberts GO (1993) Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B (Methodological)* pp 3–23
- Spiegelhalter D, Thomas A, Best N, Lunn D (2003) Winbugs user manual
- Yamanishi K, Takeuchi Ji (2002) A unifying framework for detecting outliers and change points from non-stationary time series data. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp 676–681