UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

ANÁLISE DE DADOS DE VIGILÂNCIA EPIDEMIOLÓGICA POR MEIO DE DIFERENTES TIPOS DE MODELOS DE SÉRIES TEMPORAIS

DISSERTAÇÃO DE MESTRADO

Caroline Pafiadache da Silva

Santa Maria, RS, Brasil

2014

ANÁLISE DE DADOS DE VIGILÂNCIA EPIDEMIOLÓGICA POR MEIO DE DIFERENTES TIPOS DE MODELOS DE SÉRIES TEMPORAIS

Caroline Pafiadache da Silva

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção, Área de Concentração em Gerência da Produção, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para a obtenção do grau de
Mestre em Engenharia de Produção

Orientador: Prof. Dr. Adriano Mendonça Souza

Santa Maria, RS, Brasil

2014

Universidade Federal de Santa Maria
Centro de Tecnologia
Programa de Pós-Graduação em Engenharia de Produção

A Comissão Examinadora, abaixo assinada,
aprova a Dissertação de Mestrado

ANÁLISE DE DADOS DE VIGILÂNCIA EPIDEMIOLÓGICA POR MEIO DE DIFERENTES TIPOS DE MODELOS DE SÉRIES TEMPORAIS

elaborada por
Caroline Pafiadache da Silva

como requisito parcial para a obtenção do grau de
Mestre em Engenharia de Produção

COMISSÃO EXAMINADORA:

Adriano Mendonça Souza, Dr.
(Presidente/Orientador)

Cleber Bisognin, Dr. (UFRGS)

Roselaine Ruviaro Zanini, Dra. (UFSM)

Santa Maria, 7 de Março de 2014

AGRADECIMENTOS

Aos meus pais e minha irmã pelo amor, apoio e incentivo proporcionando, como sempre, um ambiente adequado aos estudos.

Ao meu noivo pela ajuda incondicional em todos os momentos e “por acreditar no meu potencial”.

Ao Professor Dr. Adriano Mendonça Souza pela orientação, respeito, apoio e parceria durante toda a minha jornada acadêmica.

À Professora a qual eu jamais vou deixar de agradecer por tudo o que fez e faz por mim, Dra. Roselaine Ruviaro Zanini.

Ao Professor Dr. Cleber Bisognin por prontamente aceitar minha solicitação para participação na banca e se dispor a colaborar com o trabalho.

Aos professores do Programa de Pós-Graduação em Engenharia de Produção pelo comprometimento da partilha do conhecimento.

A todos os mestres que me auxiliaram de alguma forma no meu crescimento, em especial àqueles que acreditam e incentivam seus alunos, respeitam nossas limitações e enobrecem a profissão do professor.

À funcionária Márcia Regina Meneghini dos Santos pela atenção, gentileza e serviços competentemente prestados.

À Universidade Federal de Santa Maria pela excelência em ensino e pela oportunidade.

À Capes pelo apoio financeiro.

A todos os que emanam energia positiva e que de alguma forma contribuem com o meu crescimento intelectual.

RESUMO

Dissertação de Mestrado
Programa de Pós-Graduação em Engenharia de Produção
Universidade Federal de Santa Maria, RS, Brasil

ANÁLISE DE DADOS DE VIGILÂNCIA EPIDEMIOLÓGICA POR MEIO DE DIFERENTES TIPOS DE MODELOS DE SÉRIES TEMPORAIS

AUTORA: CAROLINE PAFIADACHE DA SILVA
ORIENTADOR: DR. ADRIANO MENDONÇA SOUZA
Data e Local da Defesa: Santa Maria, 7 de Março de 2014

A análise de séries históricas, obtidas nas bases de dados de saúde pública, desempenha um papel importante em processos de vigilância epidemiológica. No entanto, a implementação de metodologias de séries temporais ainda não se tornou uma rotina em meio aos profissionais dessa área. Neste trabalho são apresentados um *survey* da literatura sobre a análise de séries temporais empregada aos dados de vigilância epidemiológica e a aplicação prática de métodos estatísticos para a estimação de três modelos para doenças de notificação compulsória: modelagem de Box e Jenkins na presença e ausência da variável exógena ($ARIMA$ e $ARIMAX$) e modelo de vetor autorregressivo (VAR). Para isso, foi realizado um estudo transversal com dados secundários provenientes do SINAN (Sistema de Informação de Agravos de Notificação) constituído pelos casos de Hepatite A e Leptospirose, registrados no Rio Grande do Sul, no período de janeiro de 2008 a dezembro de 2012. Os modelos foram analisados e discutidos comparativamente por meio de medidas de desempenho. Os modelos $ARIMA$ apresentaram as melhores propriedades para a previsão de novos casos dos agravos estudados. A relação de causalidade unidirecional entre as doenças também foi estabelecida.

Palavras-chave: Vigilância epidemiológica, séries temporais, modelo $ARIMA$, modelo $ARMAX$, modelo VAR

ABSTRACT

Dissertação de Mestrado
Programa de Pós-Graduação em Engenharia de Produção
Universidade Federal de Santa Maria, RS, Brasil

DATA ANALYSIS OF SURVEILLANCE THROUGH DIFFERENT KINDS OF TIME SERIES MODELS

AUTHORESS: CAROLINE PAFIADACHE DA SILVA
ADVISER: DR. ADRIANO MENDONÇA SOUZA
Santa Maria, March 7th 2014

The analysis of time series obtained in the databases of public health plays an important role in processes of health surveillance. However, implementation of methodologies for time series has not yet become a routine in the midst of healthcare practitioners. The objective of this study is to present a theoretical review about time series analysis used for epidemiological surveillance data and practical application of statistical methods for the estimation of three models for notifiable diseases: the Box and Jenkins methodological in the presence and absence of exogenous variable ($ARIMAX$ and $ARIMA$) and vector autoregression (VAR) model. For this, we perfomed a cross-sectional study using secondary data from SINAN (Information System for Notifiable Diseases) consisting of cases of hepatitis A and leptospirosis recorded in Rio Grande do Sul, in the period January 2008 to December 2012. The models were analyzed and discussed through comparison of performance measures. The $ARIMA$ models presented the best properties for the prediction of new cases of the diseases studied. The one-way causality between the diseases was also established.

Key-words: epidemiological surveillance, time series, $ARIMA$ model, $ARIMAX$ model, VAR model.

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIC – Critério de Informação de Akaike

ADF – *Augumented Dickey-Fuller*

AR – Autorregressivo

$ARIMA$ – Modelo Autorregressivo Integrado de Média Móvel

$ARIMAX$ – Modelo Autorregressivo Integrado de Médias Móveis estendido

BIC – Critério de Informação Bayesiano

CDC – *Center for Disease Control*

Dr – Doutor

Dra – Doutora

FAC – Função de Autocorrelação

FACP – Função de Autocorrelação Parcial

H_0 – Hipótese Nula

H_1 – Hipótese Alternativa

KPSS - *Kwaiatkowski, Philips, Schmidt e Shin*

MA – Média Móvel

MAE - *Mean Absolute Error*

MAPE – *Mean Absolute Percentual Error*

MSE – *Mean Square Error*

$SARIMA$ – Modelo Sazonal Autorregressivo Integrado de Média Móvel

SIA/SUS – Sistema de Informações Ambulatoriais do SUS

SIH/SUS – Sistema de Informações Hospitalares do SUS

SIM – Sistema de Informações Sobre Mortalidade

SINAN – Sistema de Informação de Agravos e Notificação

SINASC – Sistema de Informações Sobre Nascidos Vivos

SUS – Sistema Único de Saúde

UTI – Unidade de Tratamento Intensivo

VAR – Vetor Autorregressivo

WHO – *World Health Organization*

LISTA DE ANEXOS

SUMÁRIO

1 INTRODUÇÃO

A Epidemiologia é a ciência que estuda os padrões da ocorrência de doenças em populações humanas e os fatores determinantes destes padrões. Enquanto a clínica aborda a doença em “nível individual”, a Epidemiologia aborda o processo saúde-doença em “grupos de pessoas” que podem variar de pequenos grupos até populações inteiras. O fato de a Epidemiologia, por muitas vezes, estudar morbidade, mortalidade ou agravos à saúde, deve-se, simplesmente, às limitações metodológicas da definição de saúde (JEKEL, 1999).

A Epidemiologia foi definida por Porta, Greenland e Last (2008) como “o estudo da ocorrência e distribuição de estados ou eventos relacionados com a saúde em populações específicas, incluindo o estudo dos determinantes que influenciam tais estados e a aplicação deste conhecimento para controlar os problemas de saúde”. A partir dessa definição pode-se perceber que os epidemiologistas estão preocupados não somente com a incapacidade, doença ou morte, mas também com a melhoria dos indicadores de saúde e com maneiras de promover saúde. Estes estudos epidemiológicos incluem vigilância, observação, testes de hipótese, pesquisas analíticas e experimentos.

A informação é instrumento essencial para a tomada de decisões em todos os ramos das atividades humanas. Dispor de informações de qualidade, que retratem de forma fidedigna a situação de saúde nas diferentes regiões, estados e municípios brasileiros, permite que o sistema planeje melhor suas ações de prevenção e controle de doenças, assim como de promoção da saúde. Para vigilância em saúde, essa informação se constitui no fator desencadeador do processo informação-decisão-ação.

As funções da vigilância epidemiológica podem ser resumidas em: coleta de dados; processamento dos dados coletados; análise e interpretação dos dados processados; recomendação das medidas de controle apropriadas; promoção das ações de controle indicadas; avaliação da eficácia e efetividade das medidas adotadas e divulgação de informações pertinentes.

Segundo a Secretaria de Vigilância em Saúde (BRASIL, 2005), a vigilância epidemiológica tem como propósito fornecer orientação técnica permanente para os

profissionais de saúde que, por sua vez, têm a responsabilidade de decidir sobre a execução de ações de controle de doenças e agravos, tornando disponíveis, para esse fim, informações atualizadas sobre a ocorrência dessas enfermidades, bem como dos fatores que as condicionam, numa área geográfica ou população definida. Subsidiariamente, a vigilância epidemiológica constitui-se em importante instrumento para o planejamento, organização e operacionalização dos serviços de saúde, bem como à normatização das atividades técnicas correlatas.

No Brasil, o Ministério da Saúde destaca cinco sistemas de informação em razão da maior relevância para a vigilância: Sistema de Informações de Agravos de Notificação (SINAN); Sistema de Informações Sobre Mortalidade (SIM); Sistema de Informações Sobre Nascidos Vivos (SINASC); Sistema de Informações Hospitalares (SIH/SUS) e Sistema de Informações Ambulatoriais do SUS (SIA/SUS). O SINAN apresenta maior importância para a vigilância em saúde por ser o instrumento de informação das doenças de notificação compulsória de abrangência nacional (BRASIL, 2002).

A partir destes sistemas de informação existem diferentes formas de utilização dos dados disponíveis para medir saúde e doença tais como: prevalência, incidência, mortalidade ou pode-se ainda usar simplesmente o número de casos existentes submetidos à alguma forma de monitoramento.

O uso de técnicas estatísticas para o controle e acompanhamento de doenças bem como para a detecção de epidemias de doenças de notificação compulsória auxiliam os profissionais responsáveis em tomadas de decisões efetivas de prevenção e controle com base nos resultados observados.

Dentre as diferentes formas de se analisar os dados epidemiológicos obtidos por meio dos sistemas de informação específicos, as técnicas de séries temporais constituem uma importante alternativa para avaliação do comportamento de uma doença em determinada população de estudo contribuindo para maior compreensão dos fenômenos epidemiológicos.

1.1 Justificativa e importância da pesquisa

A sensibilidade de um sistema de monitoramento de doenças é fundamental na avaliação das medidas de prevenção e controle implantadas, na identificação de mudanças no padrão epidemiológico de uma doença e na identificação de surtos e epidemias.

Um método estatístico adequado para o monitoramento em vigilância epidemiológica, de maneira geral, é aquele que identifica com maior rapidez a variabilidade na incidência de determinada doença. Quanto mais rápido o sistema sinalizar um comportamento anormal, mais rápida será a intervenção dos gestores de saúde no sentido de atuar no controle e prevenção para diminuição dessas ocorrências.

A análise de séries temporais é um procedimento de predição eficaz para novos casos de determinada doença e quando empregada de forma correta auxilia os membros da gestão em saúde na tomada de decisão em tempo real na medida em que o início de um surto ou epidemia é evidenciado.

Frente ao perfil epidemiológico brasileiro e considerando o contexto dos procedimentos estatísticos aplicados à investigação de dados em vigilância em saúde pública o tema proposto também se justifica por ser subexplorado na literatura no que tange à análise e disseminação rotineira da informação coletada nos sistemas de vigilância em saúde.

1.2 Problema de pesquisa

Uma das principais preocupações em vigilância epidemiológica é o estudo da ocorrência e distribuição de eventos relacionados com a saúde em populações específicas. Com base na literatura existente é possível verificar que a análise de séries de dados, quando empregados para examinar uma única doença para uma determinada população, demonstra ser um procedimento quantitativo de destaque na análise de dados da saúde.

No entanto, é também de interesse no campo da Epidemiologia, a comparação de séries temporais com o propósito de se verificar a relação entre as mesmas. Dessa forma o problema de pesquisa consiste em investigar se a ocorrência de doenças com mesmas vias de transmissão são passíveis de serem

modeladas conjuntamente, verificando a existência de causalidade entre elas com consequente predição de novos casos a partir dos modelos obtidos.

1.3 Objetivos

1.3.1 Objetivo geral

Essa pesquisa tem por objetivo estimar e comparar diferentes modelos de séries temporais aplicados à dados de vigilância epidemiológica.

1.3.2 Objetivos específicos

• Revisar o referencial teórico sobre a análise de séries temporais empregada nos dados de vigilância epidemiológica;

• Abordar a técnicas de modelagem de Box e Jenkins ($ARIMA$) na ausência e presença de variável exógena e modelo de vetor autorregressivo (VAR);

• Comparar, por meio de medidas de desempenho, o ajuste dos modelos estimados;

• Apresentar uma aplicação prática desses procedimentos com a finalidade exemplificar o uso desses três modelos em dados de doenças notificadas compulsoriamente às Secretarias Estatuais de Saúde.

1.4 Estrutura do Trabalho

Esta dissertação está estruturada em seis capítulos. O primeiro consta da introdução, justificativa e importância, problema de pesquisa, objetivos geral e específico, e estrutura da pesquisa.

O segundo capítulo apresenta uma revisão de trabalhos sobre métodos de séries temporais aplicados aos dados de vigilância epidemiológica presentes na literatura.

O terceiro capítulo abrange o referencial teórico a cerca da metodologia utilizada no presente trabalho.

No quarto capítulo, de forma sistemática, está descrita a metodologia empregada no estudo.

O quinto capítulo compreende a discussão e análise dos resultados.

O último capítulo é composto pela conclusão deste trabalho e sugestões para futuras pesquisas.

2 REVISÃO DA LITERATURA

O método de análise de séries temporais denominado $ARIMA$ (Autoregressive integrated moving Averange) foi proposto por Box e Jenkins em 1976 e tem sido utilizado com sucesso em diversas áreas principalmente na economia e previsão de demandas .

Na área da saúde, Choi e Thacker (1981) sugeriram a utilização do modelo $ARIMA$ para prever o número esperado de mortes por influenza e pneumonia em uma cidade dos Estados Unidos com base em dados históricos de mortalidade. O método utilizado na época pelo CDC (Center for Disease Control) era baseado na análise de regressão linear. A análise comparativa entre os métodos (proposto e utilizado pelo CDC) demonstrou que modelos $ARIMA$ forneceram uma previsão mais exata para as doenças. A partir disso, surgiram outros estudos com a aplicação de métodos de séries temporais na área da saúde com ênfase em vigilância epidemiológica.

A vigilância do vírus influenza é de fundamental importância não só devido às epidemias anuais de gripe, mas também pelo risco de novas pandemias. Estudos a cerca desse agravo têm sido realizado em vários países com o auxílio de técnicas de séries temporais.

O monitoramento da mortalidade e morbidade como indicadores da atividade da influenza, na província de Barcelona (Espanha), foi estudado por Dominguez *et al.* (1996). Modelos autorregressivos foram propostos para cada indicador, no entanto as duas séries foram também analisadas em conjunto para facilitar a detecção de possíveis implicações entre elas. O estudo conjunto das séries mostrou que o índice de mortalidade pode ser modelada separadamente da morbidade, mas a morbidade é altamente influenciada pelo número de casos de mortalidade registrados. Assim, os autores consideram que o modelo baseado na mortalidade geral pode ser usado para detectar atividades epidemiológicas de influenza.

Upshur *et al.* (1999) avaliaram a existência significativa na relação entre o vírus da gripe e os casos de internação em pessoas com mais de 65 anos com pneumonia, doença pulmonar crônica ou insuficiência cárdica, na cidade Ontário (Canadá) no período de 1988 a 1993. Por meio de comparação de modelos $ARIMA$

ajustados, os pesquisadores verificaram a existência de correlação do vírus com as duas primeiras doenças, salientaram ainda o poder e a utilidade do uso de métodos de séries temporais no estudo epidemiológico de doenças transmissíveis.

A incidência de mortalidade por influenza (H1N1) ocorrida em 1918 foi analisada por Goldstein *et al.* (2009) nas cidades Filadélfia e Nova Iorque, EUA (Estados Unidos). Esta doença conhecida na época como gripe espanhola foi responsável pela morte de cerca de 600.000 pessoas somente neste país. Uma curva de incidência para a taxa de mortalidade diária causada por esta pandemia foi ajustada por um modelo de séries temporais. Os autores verificaram que a redução da taxa de mortalidade foi ocasionada principalmente pela conscientização e mudança de comportamento da população.

A transmissão da gripe é frequentemente associada a fatores climáticos. Soebiyanto *et al.* (2010) analisaram o papel desses fatores sobre a Epidemiologia da transmissão da gripe em duas regiões caracterizadas por clima quente: Hong Kong (China) e Maricopa County (Arizona, EUA). Estas duas regiões apresentam temperaturas semelhantes, porém os índices pluviométricos são distintos. O estudo mostrou que a inclusão das variáveis climáticas resulta em modelos com melhor desempenho do que os univariados, onde os casos de gripe dependem apenas de seus valores passados e sinal de erro. Para a cidade de Hong Kong, o melhor modelo foi obtido quando a temperatura da superfície do terreno, precipitação e umidade relativa foram incluídas como covariáveis. Já para Maricopa County foram incluídas também a temperatura atmosférica máxima e pressão do ar. Os autores consideram que para países sem sistemas avançados de vigilância da gripe, as variáveis ambientais podem ser utilizadas para estimar a transmissão dessa doença no presente e no futuro próximo.

Anderson e Grenfell (1984) propuseram a utilização de técnicas de análise de séries temporais (autocorrelação e análise espectral) para examinar tendências seculares oscilatórias em dados históricos de doenças infecciosas e o impacto dos programas de vacinação em massa sobre esses fenômenos. As doenças: coqueluche e caxumba foram estudadas em dados da Inglaterra e País de Gales e o sarampo em dados da Inglaterra, País de Gales, Escócia, América do Norte e França. Os autores identificaram sazonalidade das doenças no período pré-vacinação, redução de flutuações epidêmicas de sarampo e supressão de ciclos anuais de coqueluche na Inglaterra e País de Gales após a vacinação.

Anos mais tarde, em 2005, Girard *et al.* analisaram as consequências de estratégias de vacinação contra a coqueluche na Inglaterra e País de Gales. Modelos $ARIMA$ foram ajustados em termos de efeitos sobre a saúde e custos financeiros dos programas de vacinação. Os custos diretos de programas de vacinação, com a taxa de cobertura de 90%, permaneceram sistematicamente menores do que o custo esperado com os gastos com a doença coqueluche.

Programas de vacinação também foram estudados no Brasil, Antunes *et al.* (2007) avaliaram o programa de vacinação em idosos contra o vírus da gripe em dados da cidade de São Paulo por meio de comparação das taxas de mortalidade antes e após as campanhas anuais de vacinação em massa. As taxas foram monitoradas e comparadas utilizando modelos de séries temporais ajustados para cada período. Desta forma, os autores verificaram que o índice de mortalidade geral por pneumonia e influenza caiu 26,3% após a vacinação.

Fernández-Péres *et al.* (1998) descreveram o uso de análise de séries temporais na avaliação da incidência de infecção hospitalar no Hospital Geral de Guadalajara (Espanha). A principal hipótese era a de que a ocorrência mensal de infecção poderia estar associada a fatores relacionados ao trabalho dos profissionais da saúde tais como: controle e treinamento dos funcionários, greves apoiadas por médicos e movimentação de pessoal. Após o ajuste de um modelo $ARIMA$, o controle e treinamento de pessoal pelo sistema de vigilância imposta foi associado a uma diminuição de 3,63% na incidência do acumulado mensal, os períodos de greve médica indicaram um acréscimo de 4,34% nos casos de infecção, um aumento de 0,18% foi vinculado a cada novo contrato de enfermagem e evidências foram obtidas para a possível relação entre a incidência de infecções e dias de feriados. Dessa forma, os autores sugerem a necessidade de um melhor controle do treinamento dos funcionários do hospital e a adoção de determinadas medidas preventivas durante períodos de recesso a fim de se evitar novos casos.

Para Williamson e Hudson (1999), um sistema de monitoramento de dados de vigilância em saúde pública deve associar a análise do comportamento de determinada doença ao rastreamento de sinais de desequilíbrio da série. Os autores analisaram dezessete tipos de doenças de notificação compulsória nos Estados Unidos em um sistema de dois estágios: ajuste de modelos de Box-Jenkins univariados seguido de inspeção dos resíduos por meio de gráficos de controle estatístico de processo. Para sete diferentes doenças foi possível desenvolver

modelos adequadamente ajustados e aptos ao monitoramento pelos gráficos de controle. As outras doenças, segundo os autores, apresentaram problemas como séries curtas, dados faltantes ou ausência de padronização em sua coleta.

No Brasil, a evolução temporal dos casos de AIDS notificados, segundo o grau de escolaridade, foi descrita por Fonseca *et al.*, (2000). As taxas de incidência para ambos os sexos foram calculadas segundo dois graus de escolaridade: grau 1 (casos com até oito anos de estudo) e grau 2 (com mais de oito anos de estudo), por região e ano de diagnóstico. Os autores concluíram que a epidemia de AIDS no Brasil se iniciou nos estratos sociais de maior escolaridade, com progressiva disseminação para os estratos sociais de menor escolaridade em todas as regiões para ambos os sexos.

Otero *et al.* (2001), descreveram a evolução da mortalidade por desnutrição em indivíduos com mais de 60 anos nas Regiões Metropolitanas dos Estados do Rio de Janeiro e São Paulo, entre 1980 e 1996. Um modelo $SARIMA$ foi proposto para prever a ocorrência de casos em cada uma das regiões. Os resultados apontaram maior concentração de óbitos nos meses de junho e julho na região de São Paulo e em janeiro para o Rio de Janeiro. Os autores julgaram que conhecendo melhor essa população e suas características é possível gerar subsídios para o planejamento de ações preventivas e de investimento em qualidade de vida para esta faixa etária.

Estudos sobre a poluição do ar e saúde humana têm progredido a partir de estudos descritivos dos primeiros fenômenos e a associação do aumento de efeitos adversos à saúde com episódios extremos de poluição. Em um artigo de revisão, Bell *et al.* (2004) consideraram que métodos estatísticos avançados são necessários para estudar a evolução da mortalidade e morbidade associadas às diferentes partículas tóxicas de poluentes, e constataram ainda que a evidência epidemiológica a partir de estudos de séries temporais tem desempenhado um papel crucial no estabelecimento de padrões na atenção à saúde básica.

Muñoz-Tudurí *et al.* (2006), utilizaram o método de séries temporais para analisar as causas de mortalidade em uma cidade da ilha de Minorca (Espanha) entre os anos de 1634 e 1997. O ajuste de um único modelo $ARIMA$ para uma série longa é difícil porque tendências a longo prazo podem induzir a uma mudança gradual da série a qual não é captada pelo modelo. Esse problema pode ser contornado por meio de uma divisão do conjunto de dados originais e a estimação de modelos para cada um desses subconjuntos. A série de mortalidade foi então

dividida em períodos de cinquenta anos e os pesquisadores puderam observar as componentes tendência e sazonalidade em cada intervalo e seus efeitos nos fatores determinantes nos níveis de mortalidade daquela cidade.

A malária (*Plasmodium falciparum*) é um dos principais problemas de saúde pública nas regiões tropicais com potencial para aumentar significativamente em resposta às mudanças climáticas (SACHS e MALANEY, 2002). Diversos estudos presentes na literatura têm por objetivo analisar tendências, sazonalidades e preditores bem como propor modelos para o monitoramento dessa doença de notificação compulsória.

Tian *et al.* (2008) exploraram o impacto da variabilidade climática sobre a transmissão da malária na floresta tropical de Mengla County, no sudoeste da China.Os dados coletados compreendem o período de 1971 a 1999 e as covariáveis do estudo foram as temperaturas máxima e mínima e a frequência de dias de nevoeiro, a qual foi pela primeira vez identificada como um preditor de incidência da malária em uma área de floresta tropical. Por meio de modelos $ARIMAX$ os autores concluíram que presença de neblina causa um efeito positivo sobre a transmissão da malária, pois fornece a entrada de água e consequente manutenção de criadouros para mosquitos aquáticos principalmente em estações secas. Assim, esses preditores devem ser considerados na previsão de incidência da malária para áreas florestais tropicais semelhantes e outros fatores associados à hidrologia e ecologia devem ser explorados.

A análise da incidência de malária também por meio de séries temporais é objeto de estudo de outros pesquisadores: Briët *et al.* (2008) com o objetivo de desenvolverem um sistema de previsão de recursos para o controle da malária no Sri Lanka, adicionaram a precipitação de chuva às covariáveis dos modelos estimados; Chowell *et al.* (2009) analisaram os padrões de periodicidade da malária e padrões de persistência da doença em função do tamanho da comunidade e da heterogeneidade espacial no Peru; Lin *et al.* (2009) analisaram, com o uso de modelos $SARIMA$, a distribuição geográfica, padrões demográficos e tendências temporais de malária em regiões endêmicas e não-endêmicas da China; Wangdi *et al.* (2010) usaram o modelo estendido $ARIMAX$ e dados de malária em sete regiões endêmicas do Butão com a adição dos valores das temperaturas máximas no modelo; Hanf *et al.* (2011) incorporaram no modelo $ARIMA$ dados do fenômeno *El Niño* medidos pelo índice de oscilação meridional e obtiveram resultados

satisfatórios na previsão de novos casos de malária na Guiana Francesa; Huang *et al.* (2011) associaram variáveis meteorológicas, tais como temperatura, umidade relativa e precipitação aos modelos multiplicativos $SARIMA$, para uma série de vinte anos com observações mensais de casos de malária em Motuo County (Tibet).

Segundo a Organização Mundial da Saúde (WHO – World Health Organization, 2009) a dengue afeta mais de 50 milhões de pessoas anualmente, portanto, é uma das doenças transmitidas por vetores mais importantes do mundo. A área de transmissão desta doença continua a se expandir devido a muitos fatores diretos e indiretos ligados à expansão urbana, o aumento das viagens e do aquecimento global. Medidas preventivas atuais incluem programas de controle dos mosquitos transmissores (*Aedes aegypti*), mas devido à natureza complexa da doença juntamente com a falta de medidas profiláticas eficientes, o controle da doença e a eliminação do transmissor ainda não são uma realidade. Alguns estudos sobre análise temporal da dengue apresentam modelos epidemiológicos que estimam futuros surtos utilizando informações sobre os fatores de risco da doença.

Silawan *et al.* (2008), estudaram padrões temporais e desenvolveram um modelo de previsão de incidência de dengue no nordeste da Tailândia. Os autores verificaram a presença de sazonalidade na série, observaram que surtos epidêmicos ocorrem a cada dois anos na região e ajustaram modelos de previsão para cada uma das vinte províncias analisadas. Wongkoon *et al.* (2012) ajustaram um modelo $ARIMA$ para a mesma região da Tailândia com o uso de dados mais recentes.

No Brasil, a abordagem de Box-Jenkins foi utilizada para ajustar um modelo $ARIMA$ à incidência de dengue no Rio de Janeiro por Luz *et al.* (2008). O modelo $ARIMA$ pode ser utilizado, segundos os autores, para otimizar a previsão de casos de dengue além de fornecer estimativas sobre a tendência de incidência da doença. Previsões precisas estimadas pelo modelo podem proporcionar uma importante oportunidade para o controle do agravo ou orientar intervenções inclusive no preparo para a demanda hospitalar. Extensões deste modelo também podem ser concebidas para monitorizar e prever outras doenças infecciosas em diferentes áreas geográficas.

Em São Paulo, na cidade de Ribeirão Preto, a incidência da dengue foi monitorada por Martinez *et al.* (2011) com dados compreendidos entre os anos de 2000 a 2008. Um modelo $SARIMA$ foi ajustado e os resultados mostraram que este modelo é capaz de fornecer previsões do número de casos de dengue de forma

muito eficaz e confiável sendo uma ferramenta útil para estratégias de controle da doença e prevenção da doença.

Earnest *et al.* (2012) compararam a eficácia de modelos $ARIMA$ e modelos Knorr-Held (K-H) para prever novos casos de dengue em Singapura e concluíram que a performance dos modelos são similares, mas apontam que modelos K-H são mais difíceis para serem ajustados além da necessidade de um tempo relativamente longo para estimar os parâmetros do modelo.

Na Malásia, o surto de dengue é um risco constante tanto em áreas urbanas quanto rurais. Dom *et al.* (2012) desenvolveram um modelo de previsão para a incidência de casos de dengue na cidade Subang Jaya usando a análises de séries temporais. Com base em dados coletados a partir de 2005 a 2010 um modelo $ARIMA$ foi ajustado por meio de três abordagens diferentes utilizando previsões por períodos de 52, 13 e 4 semanas de antecedência. A previsão de casos feita 4 semanas antes se apresentou mais consistente sendo considerada o modelo com melhor ajuste. No entanto, o poder preditivo do modelo considerando 52 semanas de antecedência foi melhorado com a inclusão de regressores externos como as variáveis climáticas, uma vez que esses modelos auxiliam na observação da tendência da incidência da doença. Os autores concluíram que o uso de modelos $ARIMA$ com variação semanal é uma ferramenta útil para o controle da doença bem como programas de prevenção, tendo em vista que são capazes de prever com eficácia o número de casos de dengue na Malásia.

A febre hemorrágica com síndrome renal (HFRS), causada pelo hantavírus, é uma doença endêmica na China, a qual é responsável por 90% dos casos relatados no mundo. Com o objetivo de desenvolver um modelo de séries temporais univariada para este agravo, especificamente um modelo $ARIMA$ para previsões de curto prazo, Liu *et al.* (2011) empregaram a metodologia de Box-Jenkins na série de casos registrados no período de 1975 a 2008. Os autores elegeram um modelo $ARIMA$ (0,3,1) e apresentaram as previsões para os três anos seguintes com erro absoluto percentual médio (MAPE) igual a 12,2%. Os pesquisadores concluíram que o modelo $ARIMA$ pode ser usado para otimizar a prevenção de HFRS, uma vez que fornece estimativas sobre as tendências de incidência neste país.

A hepatite E é outra doença que vêm se tornando um importante problema de saúde pública na China. Devido ao comportamento linear e não-linear da série histórica da doença, raros são os modelos matemáticos podem ser ajustados. Ren *et*

al. (2013) desenvolveram um modelo matemático combinando o uso de modelos $ARIMA$ às técnicas de redes neurais ($BPNN$) para prever a incidência de hepatite E na cidade de Xangai. A análise de séries temporais sugere um padrão sazonal da doença, e um modelo do tipo $ARIMA - BPNN$ foi ajustado para prever com precisão as infecções da hepatite E. Os autores ressaltam que modelos puramente $ARIMA$ ou $BPNN$ já haviam sido estudados e embora tenham obtido resultados melhores com o modelo combinado, sugerem que outros tipos de modelos ou associações para a análise da doença devem ser investigados.

Chen *et al.* (2012) também propuseram uma combinação de métodos para modelar dados de séries de tempo e alcançar uma maior resolução para inferências estatísticas. Neste estudo os autores, usando uma combinação de métodos bayesianos e $ARIMA$, estimaram o momento e a magnitude da mudança de tendência para os casos de tuberculose nos Estados Unidos.

Recentemente, Zhang *et al.* (2014) publicaram um estudo no qual compararam quatro tipo de modelos de séries temporais aplicados a dados de Vigilância Epidemiológica. Neste trabalho, nove séries de doenças infecciosas, coletadas através de um sistema nacional de vigilância em saúde pública na China continental, foram utilizadas para avaliar e comparar o desempenho dos seguintes métodos: dois métodos de decomposição (de regressão e de suavização exponencial), método $ARIMA$ e máquina de vetor-suporte (SVM - *support vector machine*. As apresentações foram avaliadas com base em três indicadores: erro absoluto médio (MAE), a média de erro percentual absoluto (MAPE) e erro médio quadrático (MSE). A precisão dos modelos estatísticos para previsão das doença epidêmicas provou a eficácia desses métodos em vigilância epidemiológica. Após as comparações, os autores observaram que nenhum método univariado é completamente superior aos demais, mas destacaram que a SVM supera o modelo $ARIMA$ e métodos de decomposição, na maioria dos casos

Outros desfechos têm sido estudados com auxílio de modelos de séries temporais como: internações por hemorragia intracerebral na Sérvia (MILOSEVIC, 2011), mortalidade por cirrose hepática associada ao consumo de álcool nos Estados Unidos (YE e KERR, 2011), a relação entre fatores meteorológicos e a encefalite na China (LIN *et al.*, 2012), variação sazonal de fratura de quadril no Canadá (MODARRES *et al.*, 2012), a ocorrência de oncocercose em regiões endêmicas no México (LARA-RAMIREZ *et al.*, 2013), incidência de febre tifóide na

China (ZHANG *et al*., 2013) mortes associadas à poluição do ar em Londres (BHASKARAN *et al.*, 2013), número de casos de tuberculose na China (CAO *et al.*, 2013).

Neste trabalho foram investigadas as séries de casos de hepatite A e de leptospirose no Rio Grande do Sul por meio de diferentes modelos de séries temporais.

A hepatite A é uma doença infecciosa viral, contagiosa, causada pelo vírus A (HAV) e também conhecida como hepatite infecciosa. Essa doença apresenta distribuição mundial e a principal via de contágio é fecal-oral, por contato inter-humano ou por água e alimentos contaminados. A disseminação está relacionada às condições de saneamento básico, nível socioeconômico da população, grau de educação sanitária e condições de higiene da população.

Esse agravo pode ocorrer de forma esporádica ou em surtos e, devido os sinais e sintomas pouco específicos, pode passar na maioria das vezes despercebida, favorecendo a não identificação da fonte de infecção (BRASIL, 2005).

Já a leptospirose é uma zoonose causada por uma bactéria do gênero leptospira patogênica e também de importância mundial. A doença é transmitida pelo contato com urina de animais infectados ou água e lama contaminadas pela bactéria. Seu quadro clínico varia desde infecção assintomática até quadros graves que levam o paciente ao óbito.

A leptospirose é um importante problema de saúde pública no Brasil e em outros países tropicais em desenvolvimento. Isso se deve à alta incidência nas populações que vivem em aglomerações urbanas sem adequada infraestrutura sanitária e com altas infestações de roedores. Esses fatores, associados às estações chuvosas e inundações propiciam a disseminação da doença uma vez que predispõe o contato do homem com águas contaminadas facilitando a ocorrência de surtos (BRASIL, 2009).

No capítulo a seguir serão abordadas os métodos de modelagem de séries temporais utilizados para o ajuste de diferentes modelos para os agravos estudados.

3 MÉTODOS DE MODELAGEM

Uma série temporal é, segundo Morettin e Toloi (2006), qualquer conjunto de observações ordenadas no tempo e sua característica fundamental é que estas observações geralmente são serialmente correlacionadas.

Com base nessas séries temporais é possível estimar modelos, ou seja, uma representação da realidade estruturada de tal forma que permita compreender o funcionamento total ou parcial do fenômeno observado.

A seguir serão apresentadas três metodologias para o ajuste de modelos de séries temporais: modelos $ARIMA$, modelos $ARIMAX$ e modelos VAR.

3.1 Modelos $ARIMA$

A publicação por Box, Jenkins da obra *Time series analysis: forecasting and control* (1976) conduziu a uma nova geração de ferramentas de previsão. Popularmente conhecida como metodologia de Box-Jenkins, mas tecnicamente como metodologia $ARIMA$, esses modelos resultam da combinação de três componentes também denominados “filtros”: o componente Autorregressivo (AR), o filtro de Integração (I) e o componente de Médias Móveis (MA).

Uma série de tempo pode conter os três componentes ou apenas um subconjunto deles, resultando daí várias alternativas de modelos passíveis de análise pela metodologia de Box-Jenkins.

3.1.1 Processo estacionário

Considera-se que um processo é estacionário quando ao realizar um mesmo deslocamento no tempo de todas as variáveis de qualquer distribuição conjunta finita, resulta que esta distribuição não varia, isto é, a distribuição de *N* observações

a partir de um determinado instante *t* ($X_t, X_{t+1}, \dots, X_{t+N-1}$) tem as mesmas propriedades estatísticas que a distribuição de *N* observações defasadas de *k* instantes ($X_{t+k}, X_{t+k+1}, \dots, X_{t+k+N-1}$). Um processo é não estacionário quando não se verifica a condição enunciada anteriormente. A defasagem no tempo *k* é responsável por uma alteração no valor de um ou dos dois parâmetros do processo (média e variância).

Dessa forma, segue que quando um processo é estacionário, o ajuste da série temporal é realizado por um modelo $ARMA$ (p,q). Se o processo não é estacionário, é necessário aplicar o operador de diferenças no sentido de transformar a variável original *X* em uma variável estacionária *Y* definida, no instante *t*, por $Y_t = \nabla^d X_t$.

A partir do momento em que uma série temporal é estacionária, pode-se modelá-la com uma variedade de formas e algumas delas serão abordadas a seguir.

A maioria dos métodos de análise estatística de séries temporais têm como premissa a estacionariedade do processo, assim, é necessário que os dados originais sejam inicialmente transformados para que possam atender a essa condição e seguir com a modelagem da série conforme será abordado a seguir.

3.1.2 Testes de raiz unitária

Um processo é estacionário se sua media e variância forem constantes ao longo do tempo e se o valor da covariância entre dois períodos de tempo depender apenas da distância ou defasagem entre os dois períodos, e não do período de tempo efetivo que a covariância é calculada (MATOS, 2000).

A ausência de estacionariedade ou a não-estacionariedade constitui, portanto, uma violação de pressuposto, cuja consequência é a possibilidade de se obter inferências e resultados errôneos. As séries que apresentam as características de não estacionariedade possuem, portanto, raiz unitária. Do ponto de vista estatístico, diz-se que uma série Y_t tem raiz unitária se, numa equação que relacione Y_t como variável dependente e seus próprios valores relativos ao período anterior, Y_{t-1}, como variável explicativa, o coeficiente estimado associado a Y_{t-1} for estatisticamente igual à unidade.

Assim, existem alguns procedimentos para que verificar se uma série de tempo possui raiz unitária ou não e a aqui serão apresentados dois testes comumente encontrados na literatura:

a) Dickey-Fuller Aumentado (*Augmented Dickey-Fuller* - ADF)

O teste ADF é expresso pela seguinte especificação:

$$\Delta y_t = \alpha + \beta t + \eta y_{t-1} + \sum_{i=1}^{p-1} \lambda_i \Delta y_{t-1} + \mu \quad (1)$$

em que $\lambda_i = -\sum_{i=1}^{p} \rho_i$.

Esse teste tem na hipótese nula a presença da raiz unitária ou não estacionariedade da série.

H_0: $\rho = 0$, existe raiz unitária, a série é não estacionária.

H_1: $\rho < 0$, a série é estacionária.

Dessa forma, se o teste rejeitar a hipótese nula, há uma série estacionária da série temporal (MADDALA, 2003).

b) Teste KPSS (*Kwaiatkowski, Philips, Schmidt e Shin* – KPSS)

Um dos problemas do teste de raiz unitária desenvolvido por Dickey e Fuller (1979) é seu baixo poder. Isso significa que o teste ADF não consegue rejeitar a hipótese nula para uma infinidade de séries temporais. Por essa razão, outros testes vêm sendo estudados e o KPSS desenvolvido em 1992 por Kwaiatkowski, Philips, Schmidt e Shin é um deles (BUENO, 2008).

As hipóteses desse teste são contrárias as usuais como segue:

H_0: $\rho < 0$, a série é estacionária.

H_1: $\rho = 0$, existe raiz unitária, a série é não estacionária.

A estatística do teste KPSS é baseada nos resíduos da seguinte regressão:

$$y_t = \delta x_t + u_t \tag{2}$$

em que y_t é a variável endógena; x_t são os regressores exógenos ótimos (constante ou constante e tendência), e u_t são os resíduos.

O teste KPSS utiliza uma versão modificada da estatística de máximo verossimilhança (LM) dada por:

$$LM = \frac{\sum_{t=1}^{T} s_t^2}{T^2 / f_0} \tag{3}$$

onde f_0 é o estimador dos resíduos espectrais na frequência zero, e s_t é a função acumulada dos resíduos apresentada por:

$$s_t = \sum \hat{u}_t \tag{4}$$

O teste KPSS seria, segundo seus proponentes, uma forma de complementar a análise dos testes de raiz unitária tradicionais. O quadro abaixo mostra como se deve proceder à análise conjunta dos testes.

KPSS $H_0: \rho < 0$, a série é estacionária.	ADF $H_0: \rho = 0$, a série é não estacionária	
	Aceita	Rejeita
Aceita	Decisão inconclusiva (informações insuficientes)	Decisão conclusiva (estacionariedade)
Rejeita	Decisão conclusiva (não-estacionariedade)	Decisão inconclusiva (integração fracionária)

Quadro 1- Possíveis resultados dos testes de raiz unitária

3.1.3 Autocorrelação

A autocorrelação pode ser definida como a correlação entre integrantes de séries de observações ordenadas no tempo (como as séries temporais) ou no espaço (como nos dados de corte transversal).

Segundo Gujarati e Porter (2011), o modelo clássico de regressão linear pressupõe que o termo do erro relacionado a qualquer uma das observações não é influenciado pelo termo do erro de qualquer outra observação, consequentemente a autocorrelação não existe em termos de erro a_t. Simbolicamente

$$cov\left(a_t, a_{t+1} | x_t, x_{t+1}\right) = E(a_t, a_{t+1}) = 0 \quad (5)$$

Contudo, se for verificada essa dependência, tem-se autocorrelação, ou seja

$$E(a_t, a_{t+1}) \neq 0 \quad (6)$$

Muitas vezes a correlação serial ocorre em estudos de séries temporais quando os erros associados às observações em um dado período de tempo se mantêm por transferência nos períodos futuros. Isso pode acontecer ocasionalmente com dados em corte transversal quando as unidades de observação têm uma ordem natural (VASCONCELLOS e ALVES, 2000) .

Dentre as formas adequadas para verificar a existência de autocorrelação significativa (ou verificar se os dados não são independentes), a Função de Autocorrelação (FAC) e Função de Autocorrelação Parcial (FACP) são alguns dos instrumentos utilizados para esta identificação.

3.1.4 Função de autocorrelação e autocorrelação parcial

a) Função de autocorrelação (FAC): Chama-se autocorrelação de defasamento k ("$lag\ k$") à correlação entre quaisquer duas observações defasadas de k instantes. Esta autocorrelação é definida por meio do coeficiente de correlação ρ_k, ou seja, a autocorrelação entre x_t e x_{t-k} é dado por:

$$\rho_k = \frac{Cov\ (X_t, X_{t-k})}{V(X_t)} \qquad , k = 0,1,2, \ldots \tag{7}$$

onde: $Cov\ (X_t, X_{t-k})$ é a covariância de observações defasadas em k instantes;

$V(X_t)$ é a variância de X.

A sequência de pares (k, ρ_k) é denominada função de autocorrelação. Valores negativos de k não são considerados explicitamente, tendo em vista que $\rho_k = \rho_{-k}$.

O coeficiente de autocorrelação ρ_k envolve parâmetros geralmente desconhecidos. Na prática, é necessário trabalhar com o coeficiente de autocorrelação amostral r_k para estimar ρ_k, expresso por:

$$r_k = \frac{\sum_{t=k+1}^{n}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \tag{8}$$

onde n é o número de observações da série x_t.

O conjunto formado pelos coeficientes de correlação ρ_k, para $k = 0,1,2$ ..., designa-se por FAC.

Para verificar se os dados referentes à variável X são autocorrelacionados, deve-se construir o gráfico da FAC ($\hat{\rho}$ em função de k) conhecido como correlograma. O critério de decisão consiste em verificar se todos os valores de r_k pertencem a um determinado intervalo de confiança, o que corresponde à ausência previsível de autocorrelação dos dados.

Este intervalo de confiança é determinado com base no valor esperado e na variância de r_k. Como r_k tem uma distribuição aproximadamente Normal ($r_k \sim \mathcal{N}(0, Var(r_k))$, o intervalo de confiança, para um nível de significância α, será dado pela equação:

$$-z_{\alpha/2}\sqrt{\hat{v}} \leq r_k \leq z_{\alpha/2}\sqrt{\hat{v}}, \quad \text{onde } \hat{v} = Var\ (r_k) \tag{9}$$

b) Função de autocorrelação parcial

As séries temporais podem ser descritas através de modelos autorregressivos (AR), de médias móveis (MA) ou de uma mistura dos dois ($ARMA$), desde que o processo seja estacionário (processo com média e variância constantes ao longo do tempo e o valor da covariância dependendo apenas da defasagem). A modelação de um processo não estacionário pode ser realizada com recurso aos modelos $ARIMA$.

Segundo Pereira e Requeijo (2008), se o processo for descrito por um modelo AR, a FAC começa a decrescer a partir de determinada ordem de defasamentos, embora nunca atinja o valor zero. Já em modelos MA, a função de autocorrelação anula-se a partir de certa ordem de defasamento. Assim, pode-se perceber que o comportamento da FAC dificulta a escolha do melhor modelo para descrever o processo (baseado na série temporal), sendo necessário o uso de outro instrumento, a Função de Aucorrelação Parcial (FACP).

A autocorrelação parcial com defasagem k é definida como a correlação entre X_t e X_{t+k} com os efeitos das observações ($X_{t+1}, X_{t+2}, \dots, X_{t+k-1}$) removidos. Em notação, o coeficiente de autocorrelação parcial de ordem k é usualmente designado por ϕ_{kk}.

Os coeficientes ϕ_{kk} não são conhecidos e, portanto devem ser estimados. O conjunto formado pelos coeficientes de autocorrelação parcial estimados $\hat{\phi}_{kk}$ constitui a função de autocorrelação parcial.

Para verificar se a autocorrelação parcial para o instante k tem um efeito significativo, é necessário testar se os valores de ϕ_{kk} são significativamente diferentes de zero. Para tanto, calcula-se o intervalo de confiança para um nível de significância α. Este intervalo de confiança é função do valor esperado e da variância de $\hat{\phi}_{kk}$. Se o processo for autorregressivo de ordem p, os coeficientes de autocorrelação seguem aproximadamente uma distribuição Normal com média zero e variância $Var(\hat{\phi}_{kk})$(VASCONSELLOS e ALVES, 2000).

3.1.5 Modelos univariados

3.1.5.1 Modelo autorregressivo de ordem p - AR (p)

De acordo com esse modelo, y_t é descrito apenas por seus valores passados e pelo ruído branco ε_t. A forma mais simples de um modelo AR ocorre quando y_t depende somente de y_{t-1} e de ε_t sendo denominado de processo autorregressivo estocástico de primeira ordem e descrito por:

$$y_t = \phi y_{t-1} + \varepsilon_t \tag{10}$$

onde ϕ é um parâmetro desconhecido, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ e $E(\varepsilon_t \varepsilon_s) = 0$ para $t \neq s$.

Por se tratar de um modelo estacionário, a variância de $y_t(\gamma_0)$ deve ser constante e as autocovariâncias (γ_k) devem ser independentes de t dadas por:

Variância do AR (1)

$$\gamma_0 = E(y_t^2) = E(\phi y_{t-1} + \varepsilon_t)^2 = \phi^2 V(y_{t-1}) + V(\varepsilon_t) \tag{11a}$$

$$\gamma_0 = \phi^2 \gamma_0 + \sigma_\varepsilon^2 \tag{11b}$$

$$\gamma_0 = \frac{\sigma_\varepsilon^2}{1 - \phi^2} \tag{11c}$$

Para que a variância de y_t seja não negativa e finita, é necessário que $|\phi| < 1$. Essa restrição imposta sobre o parâmetro ϕ_1 é chamada condição de estacionariedade.

Autocovariância de ordem k

$$\gamma_1 = E(y_t y_{t-2}) \tag{12a}$$

$$\gamma_1 = \phi^k \gamma_0 \tag{12b}$$

Observa-se, portanto, que as autocovariâncias não dependem de *t* e sim de *k*. Como $|\phi| < 1$, pela condição de estacionaridade, quanto maior for o valor de *k*, ou seja, quanto maior for a distância entre as observações, menor a autocovariância.

O processo autorregressivo genérico, representado por AR (p) pressupõe que seja o resultado da soma ponderada de seus *p* valores passados, além do ruído branco ε_t:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \quad (13)$$

3.1.5.2 Modelo de médias móveis MA (q)

Por este processo, a série y_t resulta da combinação linear de termos de erro de ruído branco ocorridos no período corrente e nos períodos passados. O modelo é dado por:

$$y_t = \varepsilon_t - \theta \varepsilon_{t-1} \quad (14)$$

onde θ é um parâmetro.

A média de y_t é zero, já que ε_t é um ruído branco. Sua variância é:

$$\gamma_0 = E(y_t^2) = E[(\varepsilon_t - \theta \varepsilon_{t-1})^2] = E[\varepsilon_t^2] + \theta^2 E[\varepsilon_{t-1}^2] - 2\theta E[\varepsilon_t \varepsilon_{t-1}] \quad (15a)$$

$$\gamma_0 = \sigma_\varepsilon^2 + \theta^2 \sigma_\varepsilon^2 \quad (15b)$$

As autocovariâncias do MA (1) são definidas da seguinte forma:

$$\gamma_1 = E[y_t y_{t-1}] = E[(\varepsilon_t - \theta \varepsilon_{t-1})(\varepsilon_{t-1} - \theta \varepsilon_{t-2})] \quad (16a)$$

$$\gamma_1 = -\theta \sigma_\varepsilon^2 \quad (16b)$$

$$\gamma_2 = E[y_t y_{t-2}] = E[(\varepsilon_t - \theta \varepsilon_{t-1})(\varepsilon_{t-2} - \theta \varepsilon_{t-3})] \quad (16c)$$

$$\gamma_2 = 0 \quad (16d)$$

Percebe-se então que as autocovariâncias de ordem maior ou igual a 2 são nulas:

$$\gamma_k = 0, \qquad k \geq 2 \quad (17)$$

O modelo de médias móveis genérico envolve q valores defasados de ε e é indicado por MA (q) cuja equação é:

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (18)$$

3.1.5.3 Modelo autorregressivo de médias móveis – $ARMA$ (*p,q*)

Conforme o próprio nome indica, esse modelo é uma combinação dos dois anteriores: y_t apresenta um termos autorregressivos e termos de médias móveis *q* em que a especificação genérica apresenta a seguinte equação:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \quad (19)$$

3.1.5.4 Modelos autorregressivos integrados de médias móveis – $ARIMA$ (*p,d,q*)

Os modelos de séries temporais apresentados até então são baseados na suposição de que as séries temporais envolvidas são (fracamente) estacionárias, ou seja, apresentam de forma geral média e variância constantes e covariância invariante no tempo. Sabe-se, no entanto, que muitas séries temporais são não estacionárias, isto é, são integradas e necessitam ser diferenciadas *d* vezes para tornarem-se estacionárias.

Assim, se y_t é não estacionária, mas $x_t = \nabla y_t = y_t - y_{t-1}$ é estacionária, então y_té dita integrada de ordem 1. Se y_t precisar de duas diferenças para ser estacionarizada, ou seja, se $z_t = \nabla^2 y_t = \nabla(\nabla y_t) = \nabla(y_t - y_{t-1})$ é estacionária, então y_té integrada de ordem 2.

Após diferenciar uma série temporal *d* vezes para torná-la estacionária e aplicar-lhe o modelo $ARMA$ (*p,q*), diz-se que a série temporal original é $ARIMA$ (*p,d,q*), ou seja, ela é uma série temporal autorregressiva integrada de médias móveis em que *p* denota os números de termos autorregressivos, *d* o número de vezes que a série deve ser diferenciada antes de tornar-se estacionária e que *q* o número de termos de média móvel. Esse modelo pode ser representado genericamente por:

$$w_t = \phi_1 w_{t-1} + \cdots + \phi_p w_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \quad (20)$$

onde $w_t = \nabla^d y_t$.

3.1.5.5 Modelos sazonais – $SARIMA$(p,d,q)×(P,D,Q)$_s$

A sazonalidade refere-se às flutuações periódicas dentro do período analisado. As séries sazonais apresentam um importante tipo de correlação verificada entre *s* instantes de tempo, onde *s* é o número de observações contidas em um ano (por exemplo *s* = 12 para dados mensais e *s* = 4 para dados trimestrais).

Os modelos $SARIMA$ são, na verdade, extensões dos modelos $ARIMA$ com uma componente sazonal. Os modelos podem ser puramente sazonais, ou seja, modelos onde correlações existentes ocorrem entre instantes de tempo múltiplos de *s* ou os modelos podem ser apresentados além da correlação entre *t* e *t-s*, *t-2s...* a correlação entre tempos sucessivos recaindo no modelo sazonal multiplicativo geral denominado $ARIMA\ (p,d,q) \times (P,D,Q)_s$, cuja equação geral é dada por:

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s^D y_t = \theta(B)\Theta(B^s)\varepsilon_t \quad (21)$$

onde $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$é o polinômio autorregressivo simples de grau *p*;

$\Phi(B^s) = 1 - \Phi_1 B^s - \cdots - \Phi_P B^{Ps}$ é o polinômio autorregressivo múltiplo sazonal de grau *P*;

$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ é o polinômio de média móvel simples de grau *q*;

$\Theta(B^s) = 1 - \Theta_1 B^s - \cdots - \Theta_Q B^s$ é o polinômio de média móvel múltiplo sazonal de grau *Q*;

$\nabla^d\nabla_s^D y_t = w^t$ é o filtro não-linear aplicado à série original X_t que produz um processo estacionário w^t com $\nabla^d = (1-B)^d$ e $\nabla_s^D = (1-B^s)^D$;

ε_t é o ruído branco do modelo.

3.1.6 Critérios de informação

O critério de informação é uma forma de encontrar o número ideal de parâmetros de um modelo. Essa identificação é a mais difícil das etapas da metodologia de Box-Jenkins e portanto, muitos pesquisadores preferem utilizar um

procedimento mais sistemático como o uso de critérios de seleção de modelos construídos com base na variância estimada de ε_t, no tamanho da amostra e nos valores de p e q (VASCONSCELLOS e ALVES, 2000). Há dois principais critérios de informação. A estatística de Schwartz é dada pela seguinte expressão denotada por BIC (Bayesian Information Criterion):

$$BIC(p,q) = \ln \hat{\sigma}_\varepsilon^2 + \frac{(p+q)\ln(n)}{n} \tag{22}$$

A estatística de Akaike, denotada por AIC (Akaike Information Criterion), é dada por:

$$AIC(p,q) = \ln \hat{\sigma}_\varepsilon^2 + \frac{2(p+q)}{n} \tag{23}$$

A presença de p e q nas fórmulas dos critérios AIC e BIC tem por objetivo "penalizar" os modelos com muitos parâmetros, tendo em vista que modelos mais parcimoniosos devem ser privilegiados por apresentarem menor número de parâmetros a serem estimados.

Também é preciso considerar que quanto mais parâmetros são estimados no mesmo período da amostra menor será o erro estimado, mas isso será penalizado na segunda parcela da estatística. Por isso deseja-se o menor valor de AIC ou BIC possível (BUENO, 2008).

3.1.7 Diagnóstico de resíduos

Esta etapa da metodologia consiste em verificar se o modelo identificado e estimado é adequado. Em caso positivo pode-se seguir com a previsão e caso contrário, outra especificação deve ser escolhida para modelar a série.

Os resíduos do modelo estimado $\hat{\varepsilon}_t$ são estimativas do ruído branco, ε_t. Assim, devem comportar-se aproximadamente como um ruído branco se o modelo estiver adequadamente especificado (VASCONCELLOS e ALVES, 2000).

Para verificar se isso ocorre, alguns testes e análises podem ser feitos e neste trabalho optou-se por verificar a FAC e FACP e testar a normalidade dos resíduos por meio do teste Jarque-Bera e análise visual do gráfico qq-plot.

A FAC e a FACP dos resíduos estimados devem se mostrar sem qualquer memória. Se a hipótese nula é rejeitada, isso implica dizer que há informação ainda não captada pelo modelo, o que pode gerar previsões ruins. Assim os correlogramas associados à estatística Q (teste de Ljung-Box) serão apresentados.

O teste de Ljung-Box tem como hipóteses:

H_0: os resíduos apresentam autocorrelação nula,

H_1: os resíduos apresentam autocorrelação diferente de zero.

A estatística Q de Ljung-Box é expressa por:

$$Q(K) = n(n+2)\sum_{k=1}^{K}\frac{r_k^2(\hat{\varepsilon})}{n-k} \tag{24}$$

Se o modelo for apropriado, a estatística Q do teste terá uma distribuição qui-quadrado com (k-p-q) graus de liberdade onde k é o número de defasagens tomada, p e q são as ordens do modelo ajustado. Portanto, rejeita-se a hipótese nula se $Q > \chi^2_{1-\alpha,k-p-q}$ com um nível de significância α.

O teste de normalidade de Jarque-Bera (JB), expresso na equação 25, é um teste assintótico e verifica os momentos da série estimada são iguais da normal. Sob essa hipótese, a assimetria é igual a zero e a curtose é igual a 3. As hipóteses do teste são:

H_0: os resíduos se distribuem normalmente,

H_1: os resíduos não seguem a distribuição normal.

$$JB = T\left(\frac{\hat{\alpha}_1}{6} + \frac{(\hat{\alpha}_2 - 3)^2}{24}\right) \tag{25}$$

onde $\hat{\alpha}_1$ e $\hat{\alpha}_2$ são respectivamente os coeficientes amostrais de assimetria e de curtose e T é o tamanho da amostra.

A estatística JB também segue uma distribuição qui-quadrado com 2 graus de liberdade. Assim, rejeitamos a hipótese de normalidade dos erros se $JB > \chi^2_{\alpha,2}$, onde

$\chi^2_{\alpha,2}$ é o quantil de nível $1-\alpha$ da distribuição χ^2 com dois graus de liberdade (BUENO, 2008).

O gráfico Q-Q plot é uma ferramenta útil na inspeção visual da normalidade dos resíduos, através desse gráfico é possível observar o quanto do quantil de probabilidade observado é próximo do esperado.

3.1.8 Previsão e medidas de desempenho

Depois de haver selecionado entre os modelos estimados aquele que se mostrar mais adequado, chega-se a última etapa da metodologia de Box-Jenkins, que consiste na realização de previsões para a série y_t em instantes de tempo posteriores a n.

Em geral, algumas observações finais são suprimidas da série original com o intuito de utilizá-las na avaliação e precisão das previsões por meio de medidas de desempenho. Três dessas medidas são básicas: o erro quadrático médio (MSE – mean square error); o erro absoluto médio (MAE – mean absolute error); e o erro absoluto percentual médio (MAPE – mean absolute percentual error) (BUENO, 2008).

Assim, em uma amostra com $T+H$ observações, deixam-se as últimas H observações fora da amostra e estima-se o modelo com as T observações restantes, e as medidas são calculadas da seguinte forma:

$$MSE_{t,H} = \sqrt{\frac{\sum_{h=1}^{H}(\hat{y}_t - y_t)^2}{H}} \tag{26}$$

$$MAE_{t,H} = \frac{\sum_{h=1}^{H}|\hat{y}_t - y_t|}{H} \tag{27}$$

$$MAPE_{t,H} = 100 \times \sum_{h=1}^{H}\left|\frac{(\hat{y}_t - y_t)}{y_t}\right| / \mathrm{H} \tag{28}$$

Dentre os diversos fatores que podem ser considerados na avaliação de um modelo de previsão como a facilidade de interpretação e uso, a complexidade do

modelo estimado, entre outros, a precisão tem sido usada como o critério principal para comparar modelos de séries temporais (COLLOPY *et al.*, 1992).

3.1.9 Etapas da metodologia Box-Jenkins

Segundo Gujarati e Porter (2011), o método consiste em quatro etapas:

- Identificação. Neste estágio descobre-se os valores apropriados de *p,d* e *q*. Nesta etapa o correlograma e o correlograma parcial auxiliam nessa tarefa.
- Estimação. Depois de identificados os valores apropriados de *p* e *q,* os próximo estágio é estimar os parâmetros dos termos autorregressivos e dos termos de média móvel incluídos no modelo. Às vezes, esse cálculo pode ser feito por mínimos quadrados simples, mas às vezes deve-se lançar mão de métodos de estimação não linear (nos parâmetros). Como essa tarefa é agora rotineiramente feita por pacotes estatísticos, a estimação acaba sendo sistemática.
- Verificação do diagnóstico. Após escolhido um modelo $ARIMA$ ou $SARIMA$ específico, e tendo seus parâmetros conhecidos, deve-se a seguir verificar se o modelo selecionado ajusta-se aos dados razoavelmente bem em detrimento de outros modelos. Esse é o motivo da modelagem $ARIMA$/Box-Jenkins ser mais arte do que ciência; uma habilidade considerável é requerida para escolher o modelo $ARIMA$ adequado. Um teste simples do modelo selecionado é verificar se os resíduos estimados com base nesse modelo são ruídos brancos, se forem, poderemos aceitar o ajuste específico; do contrário, deve-se recomeçar. Portanto, a metodologia Box-Jenkins é um processo iterativo.
- Previsão. Uma das razões da popularidade da modelagem $ARIMA$ é seu sucesso na previsão a qual pode ser pontual ou intervalar e, no segundo caso, é preciso conhecer a distribuição do erro de previsão.

3.2 Modelos $ARMAX$ (p,q,b)

Os modelos $ARMAX$ (Autorregressive Moving Average with Exogenous imputs) são uma extensão dos modelos $ARMA$, pois consideram a possibilidade de inclusão de variáveis exógenas como regressores.

Assim, um modelo $ARMAX$ pode ser escrito da seguinte forma:

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \, \epsilon_{t-j} + \epsilon_t + \sum_{k=1}^{Nx} \beta_k Z_{(t,k)} \tag{29}$$

onde y_t é a variável dependente; α é uma constante; $\phi_1, \dots, \phi_p$ são os coeficientes dos termos autorregressivos; $y_{t-1}, \dots, y_{t-p}$ são os termos autorregressivos; ϵ_t é a componente aleatória do modelo, com $\epsilon_t \sim N(0, \sigma^2)$; $\theta_1, \dots, \theta_q$ são os coeficientes da componente aleatória; $\epsilon_t, \dots, \epsilon_{t-q}$ são as componentes aleatórias defasadas; Z é uma matriz de variáveis exógenas na qual cada coluna é uma série temporal e $Z_{(t,k)}$ é o elemento alocado da t-ésima linha e na k-ésima coluna; e $\beta_1, \dots, \beta_p$ são os coeficientes dos termos de variáveis exógenas.

Segundo Kohn (1979), um modelo $ARMAX$ deve obedecer às seguintes pressuposições básicas para a sua aplicação: os resíduos aleatórios são identicamente distribuídos com média zero e as variáveis exógenas e endógenas devem ser estacionárias.

O modelo $ARMAX$ pode ser alterado com incorporação de um termo integrativo para o sistema descrito, neste caso o modelo é chamado de $ARIMAX$ (I de integração) e é utilizado para descrever sistemas com perturbações lentas.

3.3 Modelos de vetores autorregressivos (VAR)

A metodologia VAR foi desenvolvida por Sims (1980) como uma resposta às críticas ao grande número de restrições impostas às estimações pelos modelos estruturais. Para o autor, se há uma simultaneidade verdadeira entre um conjunto de

variáveis, todas elas devem ser tratadas de forma semelhante, não devendo haver qualquer distinção, *a priori* entre as variáveis endógenas e exógenas.

Dessa forma, a ideia era desenvolver modelos dinâmicos com o mínimo de restrições, nos quais todas as variáveis fossem tratadas como endógenas. Sendo assim, os modelos VAR examinam relações lineares entre cada variável e os valores defasados dela própria e de todas as demais variáveis, impondo como restrições à estrutura do modelo somente a escolha do conjunto relevante de variáveis e do número máximo de defasagens envolvidas nas relações entre elas. Nos modelos VAR o número de defasagens é normalmente escolhido com base em critérios estatísticos, como os de Akaike ou Schwarz (CHAREMZA e DEADMAN, 1997).

Segundo Matos (2000), o modelo VAR apresenta três aspectos positivos fundamentais: o primeiro refere-se à simplicidade da formulação em virtude de não existir preocupação quanto à determinação das variáveis endógenas ou exógenas, pois todas são endógenas, embora variáveis puramente exógenas possam ser incluídas nesse modelo para se considerar, por exemplo, tendências e fatores sazonais. Em segundo lugar, o modelo VAR pode ser facilmente estimado mediante a aplicação do método de mínimos quadrados ordinários em cada equação isoladamente. A terceira virtude do VAR refere-se à qualidade, muitas vezes melhor, da previsão de valores em relação aos complexos modelos de equações simultâneas.

Apesar dessas virtudes, Gujarati e Porter (2011) relatam algumas limitações apontadas pelos críticos da modelagem VAR:

- Diferentemente dos modelos de equações simultâneas, um modelo VAR é ateórico porque utiliza menos informação prévia. Isso porque em modelos de equações simultâneas, a exclusão ou inclusão de certas variáveis tem um papel fundamental na identificação do modelo.
- Devido à sua ênfase na previsão, os modelos VAR são menos adaptados para a análise de políticas econômica, uma vez que modelos de séries temporais são exaustivamente aplicados a essa área.
- Dificuldade na escolha do comprimento das defasagens, além da perda do número do grau de liberdade com o aumento do número daquelas.
- Problemas de transformação de variáveis não estacionárias em estacionárias, quando há, por exemplo, mistura de séries com integrações de diferentes ordens.

- Dado que os coeficientes individuais nos modelos estimados VAR são frequentemente difíceis de interpretar, são usadas funções de impulso-resposta, permitindo a determinação dos impactos provocados numa variável dependente dos choques ou inovações em seu respectivo termo erro. Por exemplo, choques de um desvio-padrão do termo erro relativo a determinada equação no período *t* alterarão o valor corrente da respectiva variável dependente, assim como seus valores futuros, por vários períodos. Como essa mesma variável aparece nas demais equações do sistema, mudanças ou choque iniciais nesse termo erro acarretarão respostas nas demais variáveis. A conclusão é idêntica para choques nos demais termos residuais.

De modo geral, pode-se expressar um modelo autorregressivo de ordem p por um vetor com n variáveis endógenas, X_t, que estão conectadas entre si por meio de uma matriz A, conforme segue:

$$AX_t = B_0 + \sum_{i=1}^{p} B_i X_{t-i} + B\epsilon_t \qquad (30)$$

em que:

A é uma matriz $n \times n$ que define as restrições contemporâneas entre as variáveis que constituem o vetor $n \times 1$,X_t;

B_0 é um vetor de constantes $n \times 1$;

B_i são matrizes $n \times n$;

B é uma matriz diagonal $n \times n$ de desvios-padrão;

ϵ_t é um vetor $n \times 1$ de perturbações aleatórias não correlacionadas entre si, sendo que os erros do modelo seguem características de ruído branco, ou seja, média zero e variância constante.

Na equação 28 pode-se observar as relações entre as variáveis endógenas, frequentemente decorrente de um modelo teoricamente estruturado, e por isso chama-se *forma estrutural* ou *primitiva*. Os choques ϵ_t são denominados choques estruturais porque afetam individualmente cada uma das variáveis endógenas. Os choques estruturais são considerados independentes entre si porque as inter-relações entre um choque e outro são captadas indiretamente pela matriz A. Logo, a independência dos choques dá-se sem perda de generalidade.

Devido à endogeneidade das variáveis, esse modelo é normalmente estimado em sua *forma reduzida*, ou seja, multiplicando ambos os lados da equação 28 por A^{-1} obtém-se a seguinte forma padrão para o modelo de vetores autorregressivos:

$$X_t = A^{-1}B_0 + \sum_{i=1}^{p} A^{-1}B_0 X_{t-i} + A^{-1}B\epsilon_t \tag{31}$$

$$= \Phi_0 + \sum_{i=1}^{p} \Phi_i X_{t-i} + e_t \tag{32}$$

em que

$$\Phi_i \equiv A^{-1}B_i, \quad i = 0,1, \dots, p \quad B\epsilon_t \equiv Ae_t.$$

Para uma melhor visualização, Bueno (2008) apresenta este modelo por meio de um exemplo de um modelo bivariado de ordem 1. A partir dessa simplificação, uma série de resultados intuitivos que valem para modelos de ordem maior pode ser desenvolvida, facilitando o entendimento da metodologia. Assim, segue o modelo bivariado:

$$y_t = b_{10} - a_{12}z_t + b_{11}y_{t-1} + b_{12}z_{t-1} + \sigma_y\epsilon_{yt} \tag{33}$$

$$z_t = b_{20} - a_{21}y + b_{21}y_{y-1} + b_{22}z_{t-1} + \sigma_z\epsilon_{zt} \tag{34}$$

Trata-se de uma especificação inicial bem razoável, pela qual as variáveis são mutuamente influenciadas uma pela outra, tanto contemporaneamente como pelos seus valores defasados. Esse modelo não pode ser estimado diretamente, já que ambas as variáveis contemporâneas z_t e y_t são individualmente correlacionadas com os erros ϵ_{yt} ou ϵ_{zt}, respectivamente. Isso ocorre porque cada uma dessas variáveis depende contemporaneamente da outra (efeito *feedback*). O objetivo do modelo VAR é desenvolver técnicas para evitar esse problema, visando-se encontrar a trajetória da variável de interesse ante um choque nesses erros, ou seja, um choque estrutural.

As hipóteses assumidas para esse modelo são as de que y_te z_t são estacionários, ϵ_{yt} e ϵ_{zt} apresentam características de ruído branco e a covariância entre ϵ_{yt} e ϵ_{zt} é igual a zero.

O modelo apresentado está na *forma estrutural* e pode ser escrito em matrizes:

$$\begin{bmatrix} 1 & a_{12} \\ a_{21} & 1 \end{bmatrix} \begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \sigma_y & 0 \\ 0 & \sigma_z \end{bmatrix} \begin{bmatrix} \epsilon_{yt} \\ \epsilon_{zt} \end{bmatrix} \quad (35)$$

A forma reduzida desse modelo simplificado é dada por:

$$X_t = \Phi_0 + \Phi_1 X_{t-1} + e_t; \quad (36)$$

$$\Phi_0 \equiv A^{-1}B_0;\ \Phi_1 \equiv A^{-1}B_1;\ Ae_t \equiv Be_t \quad (37)$$

A especificação do melhor modelo VAR consiste em determinar o número de defasagens adequado para se obter resíduos com características de ruído branco.

3.3.1 Critérios Penalizadores

Antes da estimação das equações do modelo VAR, é necessário decidir sobre o comprimento máximo de defasagens k. Uma forma de decidir esta questão é utilizar um critério como o Akaike ou o Schwarz que levam em consideração o número de parâmetros utilizados na modelagem sendo denominados critérios penalizadores (MADDALA, 2003).

A versão multivariada dos critérios AIC e BIC é uma generalização da versão univariada da seguinte forma:

$$AIC = -2\left(\frac{l}{T}\right) + 2\left(\frac{k}{T}\right) \quad (38)$$

$$BIC = -2\left(\frac{l}{T}\right) + \frac{k\log(T)}{T} \quad (39)$$

Em que T é o tamanho da amostra; l é o valor da função de log verossimilhança e k é o número de parâmetros estimados.

Esses dois critérios são estruturados em função da variância dos resíduos, incorporando um ajuste para captar a perda de graus de liberdade que advém com a estimação dos parâmetros. Em termos de interpretação, o ideal é que, como na forma univariada, AIC e BIC assumam valor mínimo. Assim, entre n estimativas,

escolhe-se o número de defasagens $k(k = 0, 1, 2, \dots, m)$ que produza o menor valor para essas estatísticas.

Também, para uma decisão mais acertada, outros métodos de determinação do número de defasagens podem ser utilizados, como o logaritmo da função de máxima verossimilhança ($\ln L$), representado algebricamente como:

$$-2 \ln \lambda = 2 \sum_i^n O_i \ln \left(\frac{O_i}{E_i}\right) \quad (40)$$

onde $\ln$ é a máxima verossimilhança, O_i é a frequência observada e E_i é a frequência esperada.

3.3.2 Teste de Causalidade de Granger

Quando se trabalha com modelos de equações simultâneas ou estruturais, é possível avaliar se uma variável y influencia ou ajuda a prever a variável z. Se isso não acontece, então se diz que y não-Granger-causa z.

O teste de causalidade de Granger assume que a informação relevante para a predição das respectivas variáveis y e z está contida apenas nas séries de tempo sobre essas duas variáveis. Dessa forma, uma série de tempo estacionária y causa, no sentido de Granger, uma outra série estacionária z se as melhores predições estatisticamente significantes de z podem ser obtidas ao incluirmos valores defasados de y aos valores defasados de z.

A causalidade deve existir pelo menos em uma das direções: y causa z, z causa y ou ainda bi-causal. Em um contexto usual de séries temporais, a equação de autorregressão bivariada pode ser escrita da seguinte forma:

$$z_t = \sum_{i=1}^{p} \alpha_i \, z_{t-i} + \sum_{i=1}^{p} \beta_i \, y_{t-i} + e_t \quad (41)$$

A hipótese de que y não-Granger-causa z, ou seja:

$$H_o: \beta_1 = \beta_2 = \cdots = \beta_i = 0, i = 1,2, \ldots, p \quad (42)$$

pode ser analisada por meio de um teste F o qual verifica a significância conjunta de parâmetros. A estatística do teste é dada por:

$$\mathrm{S_1} = \frac{(\mathrm{e_r^2} - \mathrm{e_u^2})/\mathrm{p}}{\mathrm{e_u^2}/(\mathrm{T}-2\mathrm{p}-1)} \xrightarrow{\mathrm{d}} \mathrm{F}(\mathrm{p}, \mathrm{T}-2\mathrm{p}-1) \quad (43)$$

em que r representa restrito e u, não restrito. Se $S_1 > F_{5\%}$rejeita-se a hipótese nula de que y não-Granger-causa z (BUENO, 2008).

4 METODOLOGIA

Neste capítulo serão apresentados os procedimentos metodológicos do estudo, assim descritos: métodos e técnicas utilizados, fontes de dados, caracterização e limitação da pesquisa e a estratégia analítica.

4.1 Fontes de dados

Para a realização deste estudo foram utilizados dados mensais de notificação de leptospirose e hepatite A no Rio Grande do Sul (RS), proveniente de dados secundários obtidos no ambiente virtual do Sistema de Informação de Agravos de Notificação (SINAN). Os casos estão compreendidos no período de janeiro de 2008 a dezembro de 2012.

Esses agravos foram eleitos para serem analisados por se tratarem de duas doenças que apresentam vias de contágio semelhantes e por serem doenças de notificação compulsória que se apresentam com frequência no Estado do Rio Grande do Sul.

No tratamento dos dados e aplicação dos testes estatísticos foi utilizado o *software* estatístico *E-VIEWS 7* e os métodos de estimação: $ARIMA$, $ARIMAX$ e VAR.

Este trabalho envolve somente pesquisa bibliográfica e informações originárias de bancos de dados de uso e acesso público (DATASUS), eximindo-se da avaliação do Comitê de Ética.

4.2 Caracterização e limitação da pesquisa

Essa é uma pesquisa transversal, de caráter exploratório e com o objetivo de apresentar uma comparação de diferentes modelos de séries temporais por meio de dados epidemiológicos.

A utilização do SINAN como base de dados local é uma fonte valiosa de informações epidemiológicas, apesar de ainda existir falhas quanto ao preenchimento completo de todos os campos da ficha de notificação/investigação não configurando a realidade dos casos.

A gradual demora na notificação dos casos por parte dos sistemas de vigilância, bem como a falta de publicação de notificações diárias ou semanais também se torna um aspecto limitante do estudo, pois essa falta de alimentação nos bancos de dados específicos dificulta a implementação de métodos de controle e consequentemente, a detecção de surtos ou epidemias em tempo real.

A análise das implicações das doenças aqui estudadas também pode ser um fator limitante por não ter sido realizada de forma aprofundada, já que não era objetivo deste estudo.

4.3 Estratégia Analítica

Com o objetivo de alcançar os objetivos propostos nessa pesquisa, a sequência de procedimentos foi realizada sob a ótica dos três diferentes modelos abordados: $ARIMA$, $ARIMAX$ e VAR.

4.2.1 $ARIMA$

Para se obter um modelo $ARIMA$ com base na metodologia Box-Jenkins, a ordem estratégia consiste em:

1 - Análise da estacionariedade da série. Este passo pode ser feito ao calcular a função de autocorrelação (FAC) e a função de autocorrelação parcial (FACP) ou fazendo uma análise de raiz unitária (testes ADF ou KPSS).

2 - Se a série temporal for não estacionária, deve ser executada a diferenciação uma ou mais vezes até atingir a estacionariedade.

3 - As FAC e FACP da série temporal são calculadas para descobrir se a série é puramente autorregressiva ou puramente do tipo média móvel ou ainda uma mistura das duas.

4 - Um modelo experimental é estimado.

5 - Os resíduos desse modelo experimental são examinados para descobrir se são de ruído branco. Se forem, o modelo experimental será provavelmente uma boa aproximação ao processo estocástico subjacente. Se não forem, o processo será novamente iniciado.

6 - Caso exista mais de um modelo adequado, deve-se usar os critério de seleção do melhor modelo (AIC, BIC, ln L e parcimônia).

7 - O modelo selecionado pode ser utilizado para previsão.

8 - Após serem realizadas as previsões, calculam-se as medidas de desempenho do modelo para posterior comparação com aqueles estimados por meio de outras metodologias.

4.3.2 $ARMAX$

A sequência de passos para estimação de um modelo $ARMAX$ é a mesma anterior acrescentando uma variável exógena. Neste caso optou-se por trabalhar com uma serie de doença como variável endógena e a outra como variável exógena desde que tenham relação de causalidade. Dessa forma, antes de seguir a ordem estratégica descrita anteriormente, é necessário que se faça o teste da causalidade de Granger para verificar a existência de relação entre as variáveis.

4.3.3 VAR

Os passos estatísticos para estimar um modelo de vetor autorregressivo segue a seguinte sequência:

1 - Análise da estacionariedade da série. Este passo pode ser feito ao calcular a função de autocorrelação (FAC) e a função de autocorrelação parcial (FACP) ou fazendo uma análise de raiz unitária (testes ADF ou KPSS).

2 - Se a série temporal for não estacionária, deve ser executada a diferenciação uma ou mais vezes até atingir a estacionariedade.

3 – Definir o melhor número de defasagem em que o modelo VAR apresente melhor resultado baseando-se nos critérios de informação AIC, BIC e ln L.

4 – Estimar o modelo VAR.

5 – Análise dos resíduos, se apresentarem comportamento de ruído branco o modelo pode ser usado para previsão de novos casos.

6 – Após serem realizadas as previsões, calcula-se as medidas de desempenho do modelo.

5 RESULTADOS E DISCUSSÃO

5.1 Análise Descritiva

Uma análise descritiva dos dados foi realizada inicialmente e, nas Tabela 1 e 2, constam, respectivamente, os casos autóctones de hepatite A e leptospirose por mês e ano de ocorrência, acompanhados da média, desvio padrão e coeficiente de variação por período.

Tabela 1- Casos autóctones de hepatite A, segundo o ano e o mês do Estado do Rio Grande do Sul. Brasil, 2008 a 2012.

Mês	Ano					Total	Média	Desvio-padrão	Coef.de Variação
	2008	2009	2010	2011	2012				
Jan	28	27	42	48	23	168	33,6	10,78	0,32
Fev	27	47	102	46	12	234	46,8	34,10	0,73
Mar	38	75	145	61	21	340	68	47,79	0,70
Abr	48	83	100	41	26	298	59,6	30,78	0,52
Mai	42	77	113	36	46	314	62,8	32,23	0,51
Jun	30	46	67	16	23	182	36,4	20,40	0,56
Jul	18	44	53	29	26	170	34	14,20	0,42
Ago	24	39	35	17	20	135	27	9,57	0,35
Set	19	28	56	15	18	136	27,2	16,81	0,62
Out	16	19	40	16	11	102	20,4	11,33	0,56
Nov	31	29	34	19	13	126	25,2	8,84	0,35
Dez	15	22	48	14	5	104	20,8	16,36	0,79
Total	336	536	835	358	244	2309	-	-	-

Fonte: Ministério da Saúde, Secretaria de Vigilância em Saúde, Sistema de Informação de Agravos de Notificação (SINAN).

Pode-se verificar que, apesar de um coeficiente de variação elevado para as duas séries, a média de casos de hepatite A é maior nos meses de fevereiro a maio, enquanto os de leptospirose ocorrem com maior frequência de janeiro a abril.

Tabela 2 - Casos autóctones de leptospirose, segundo o ano e o mês do Estado do Rio Grande do Sul. Brasil, 2008 a 2012.

Mês	Ano					Total	Média	Desvio-padrão	Coef.de Variação
	2008	2009	2010	2011	2012				
Jan	85	51	71	100	37	344	68,8	25,34	0,37
Fev	69	73	78	142	52	414	82,8	34,51	0,42
Mar	54	79	54	77	37	301	60,2	17,68	0,29
Abr	33	40	53	54	14	194	38,8	16,45	0,42
Mai	24	16	38	40	14	132	26,4	12,12	0,46
Jun	10	13	22	9	7	61	12,2	5,89	0,48
Jul	9	14	18	21	9	71	14,2	5,36	0,38
Ago	12	10	27	22	15	86	17,2	7,12	0,41
Set	15	21	22	11	24	93	18,6	5,41	0,29
Out	17	19	15	22	26	99	19,8	4,32	0,22
Nov	33	56	19	25	20	153	30,6	15,24	0,50
Dez	34	68	29	17	29	177	35,4	19,27	0,54
Total	395	460	446	540	284	2125	-	-	-

Fonte: Ministério da Saúde, Secretaria de Vigilância em Saúde, Sistema de Informação de Agravos de Notificação (SINAN).

Dessa forma, partiu-se para a busca de modelos que melhor se ajustassem aos dados dentro das metodologias estipuladas.

5.2 Ajuste do modelo $ARIMA$

Na Figura 1 pode-se observar a série do número de casos de hepatite A e de leptospirose no Rio Grande do Sul entre os anos de 2008 a 2012. A partir de uma análise visual não é possível verificar, com precisão, a existência de tendência ou sazonalidade. Neste gráfico estão apresentadas as 60 observações de cada conjunto de dados, no entanto, as análises posteriores foram feitas suprimindo as seis últimas observações com o intuito de utilizá-las para a análise do ajuste dos modelos.

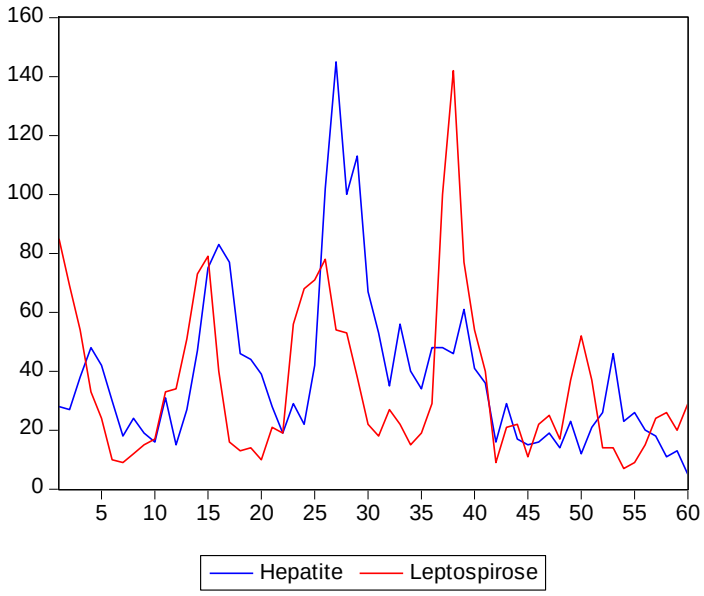


Figura 1 - Gráfico da série de casos mensais de Hepatite A e Leptospirose do Rio Grande do Sul, Brasil, nos anos de 2008 a 2012.

Para verificar a estacionariedade foram realizados os testes ADF e KPSS de raiz unitária nas séries em nível e em primeira diferença. Os resultados obtidos são apresentados nas Tabelas 3 e 4 respectivamente.

Segundo os resultados referentes ao teste de raiz unitária, apresentados na Tabela 3, verifica-se que a série de hepatite A não é estacionária em nível, pois o valor crítico do ADF é maior que o valor calculado para os níveis de confiança, o que leva a crer que essa série apresenta pelo menos uma raiz unitária. No entanto, quando aplicada uma diferenciação na série, ela se torna estacionária enquanto a série de leptospirose é estacionária em nível conforme o teste aplicado.

Com o objetivo de confirmar estas conclusões obtidas pelo teste ADF, empregou-se o teste *Kwaiatkowski, Philips, Schmidt e Shin* (KPSS) e os resultados estão apresentados na Tabela 4. A série da hepatite A em nível apresentou-se estacionária, entrando em contradição com o resultado anterior. A estacionariedade se confirmou para as séries de hepatite A diferenciada e de leptospirose.

Tabela 3 - Teste de *Augmented Dickey-Fuller* (ADF) para séries de casos hepatite A em nível e com a primeira diferença e leptospirose em nível.

Série	Valor crítico ADF	p-valor
Hepatite A	-2,5469	0,1105
	-3,5600 (1%)	
	-2,9177 (5%)	
	-2,5967 (10%)	
Δ Hepatite A	-6,7543	<0,001
	-3,5627 (1%)	
	-2,9188 (5%)	
	-2,5973 (10%)	
Leptospirose	-4,2835	0,0012
	-3,5657 (1%)	
	-2,9188 (5%)	
	-2,5973 (10%)	

Fonte: A autora (2014), a partir do *software* E-views 7; Δ representa a série em primeira diferença.

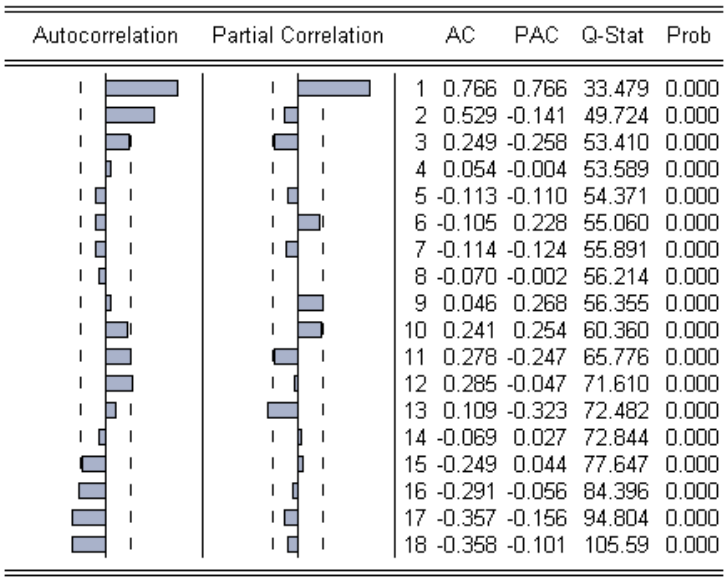
Tabela 4 - Teste de *Kwaiatkowski, Philips, Schmidt e Shin* (KPSS) para séries de casos hepatite A em nível e com a primeira diferença e leptospirose em nível.

Série	Valor crítico KPSS	p-valor
Hepatite A	0,1661	<0,001
	0,7390 (1%)	
	0,4630 (5%)	
	0,3470 (10%)	
Δ Hepatite A	0,0583	0,9708
	0,7390 (1%)	
	0,4630 (5%)	
	0,3470 (10%)	
Leptospirose	0,0656	<0,001
	0,7390 (1%)	
	0,4630 (5%)	
	0,3470 (10%)	

Fonte: A autora (2014), a partir do *software* E-views 7; Δ representa a série em primeira diferença.

Para auxiliar na análise da estacionariedade da série de hepatite A em nível, verificou-se o comportamento dos correlogramas dos dados. Na análise da função de autocorrelação (Figura 2) a série pode ser considerada estacionária, pois há um padrão de decaimento rápido. Assim, para dar sequência às metodologias empregadas, optou-se por trabalhar com as séries de hepatite A e leptospirose em nível.

A identificação dos modelos representativos das séries segue com a análise da função de autocorrelação (FAC) e autocorrelação parcial (FACP). Na Figura 2, o gráfico da FAC da série de hepatite A apresenta decaimento exponencial sugerindo a presença de um parâmetro autorregressivo (AR) e a FACP indica o número de parâmetros significativos. Dessa forma, estima-se que o modelo adequado para os casos de hepatite A terá em sua formação pelo menos um componente AR (1). Neste caso, devido a forma de senóide amortecida apresentada na FAC, também deve ser investigado a presença de uma componente sazonal

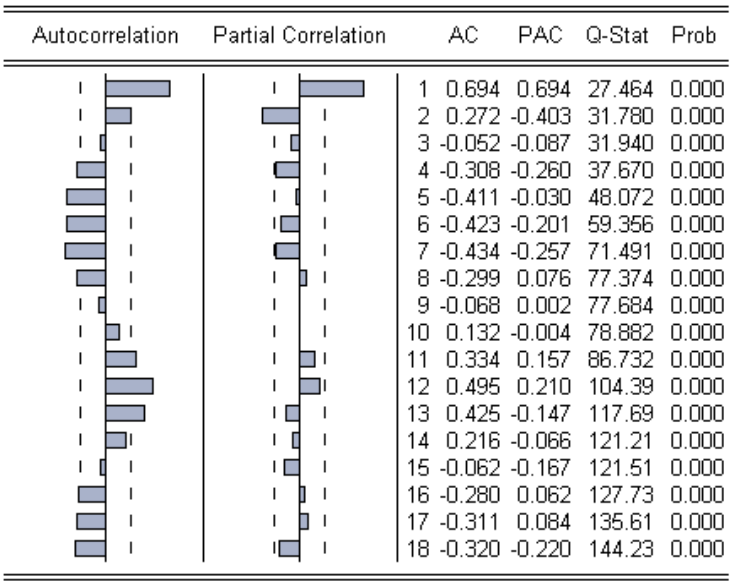


	AC	PAC	Q-Stat	Prob
1	0.766	0.766	33.479	0.000
2	0.529	-0.141	49.724	0.000
3	0.249	-0.258	53.410	0.000
4	0.054	-0.004	53.589	0.000
5	-0.113	-0.110	54.371	0.000
6	-0.105	0.228	55.060	0.000
7	-0.114	-0.124	55.891	0.000
8	-0.070	-0.002	56.214	0.000
9	0.046	0.268	56.355	0.000
10	0.241	0.254	60.360	0.000
11	0.278	-0.247	65.776	0.000
12	0.285	-0.047	71.610	0.000
13	0.109	-0.323	72.482	0.000
14	-0.069	0.027	72.844	0.000
15	-0.249	0.044	77.647	0.000
16	-0.291	-0.056	84.396	0.000
17	-0.357	-0.156	94.804	0.000
18	-0.358	-0.101	105.59	0.000

Figura 2 - Autocorrelação e autocorrelação parcial da série de hepatite A no Rio Grande do Sul, Brasil, 2008 a 2012.

O mesmo procedimento foi realizado para a série de casos de leptospirose conforme observa-se na Figura 3, e o comportamento das funções de autocorrelação e autocorrelação parcial se assemelham aos da série anterior. Assim,

pode-se sugerir a presença de pelo menos um componente autorregressivo de ordem 1 e de sazonalidade na série.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.694	0.694	27.464	0.000
		2	0.272	-0.403	31.780	0.000
		3	-0.052	-0.087	31.940	0.000
		4	-0.308	-0.260	37.670	0.000
		5	-0.411	-0.030	48.072	0.000
		6	-0.423	-0.201	59.356	0.000
		7	-0.434	-0.257	71.491	0.000
		8	-0.299	0.076	77.374	0.000
		9	-0.068	0.002	77.684	0.000
		10	0.132	-0.004	78.882	0.000
		11	0.334	0.157	86.732	0.000
		12	0.495	0.210	104.39	0.000
		13	0.425	-0.147	117.69	0.000
		14	0.216	-0.066	121.21	0.000
		15	-0.062	-0.167	121.51	0.000
		16	-0.280	0.062	127.73	0.000
		17	-0.311	0.084	135.61	0.000
		18	-0.320	-0.220	144.23	0.000

Figura 3 - Autocorrelação e autocorrelação parcial da série de leptospirose no Rio Grande do Sul, Brasil, 2008 a 2012.

Um modelo é considerado adequado quando utiliza, no processo de modelagem, um menor número de parâmetros (princípio da parcimônia), com *p-valores* significativos, valores de critérios de informação de Akaike (AIC) e Bayesiano (BIC) mínimos e logaritmo da verossimilhança (ln L) máximo. Privilegiou-se neste estudo selecionar os modelos que apresentaram o menor número de parâmetros com significância de 5% e que resultaram em estimativas de predição com menor erro.

Com base nos gráficos da função de autocorrelação e autocorrelação parcial da série de hepatite A, alguns modelos puderam ser ajustados (ANEXO A) e com auxílio dos critérios AIC, BIC e ln L e o número de coeficientes significativos, dois modelos foram selecionados: $SARIMA\ (1,0,0) \times (1,0,0)_{12}$ por apresentar o maior valor do logaritmo da verossimilhança e $SARIMA\ (1,0,0) \times (0,0,1)_{12}$ devido aos baixos valores dos critérios AIC e BIC.

A análise dos resíduos dos modelos selecionados foi feita por meio de correlogramas e teste de normalidade (estatística Jarque-Bera). Nas Figuras 4 e 5

pode-se observar a FAC e FACP e o gráfico QQ-plot dos resíduos dos modelos $SARIMA\ (1,0,0) \times (1,0,0)_{12}$ e $SARIMA\ (1,0,0) \times (0,0,1)_{12}$ ajustados para a série de hepatite A. De acordo com os correlogramas e os valores obtidos no teste de Jarque-Bera (p-valor= 0,0016 e p-valor=0,4732, respectivamente) pode-se considerar que os resíduos do segundo modelo apresentam características de ruído branco, sendo este portanto o eleito para representar a série.

No caso da série de leptospirose, um número maior de modelos foi ajustado (ANEXO B) e, dentre esses, dois modelos foram escolhidos por apresentar os melhores critérios de seleção: $SARIMA\ (1,0,0) \times (0,0,1)_{12}$ e $SARIMA\ (2,0,0) \times (0,0,1)_{12}$, sendo o último estimado com a constante.

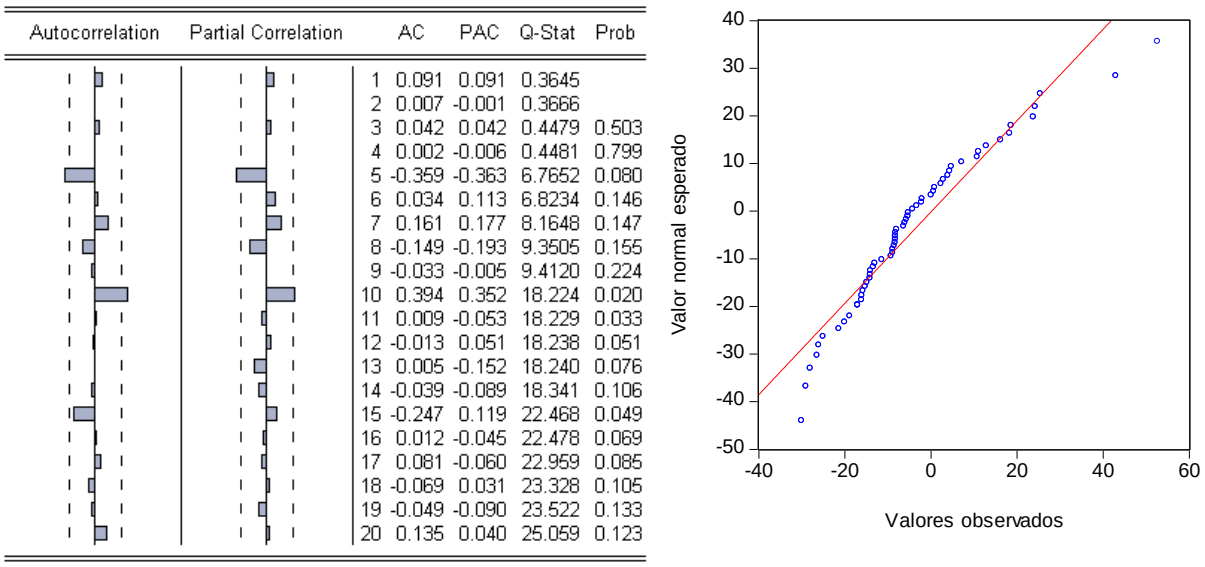
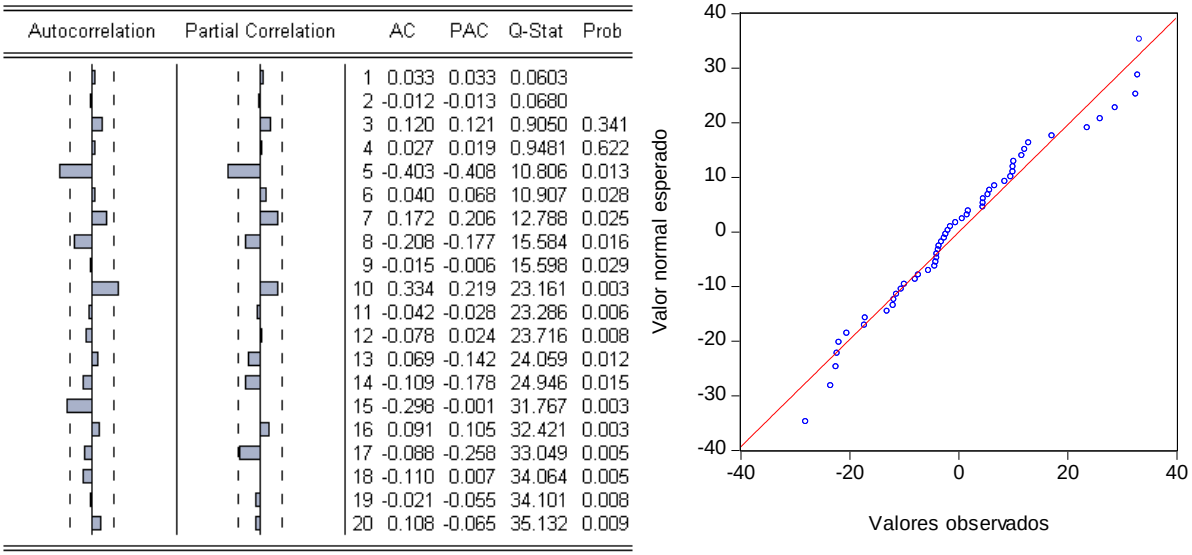


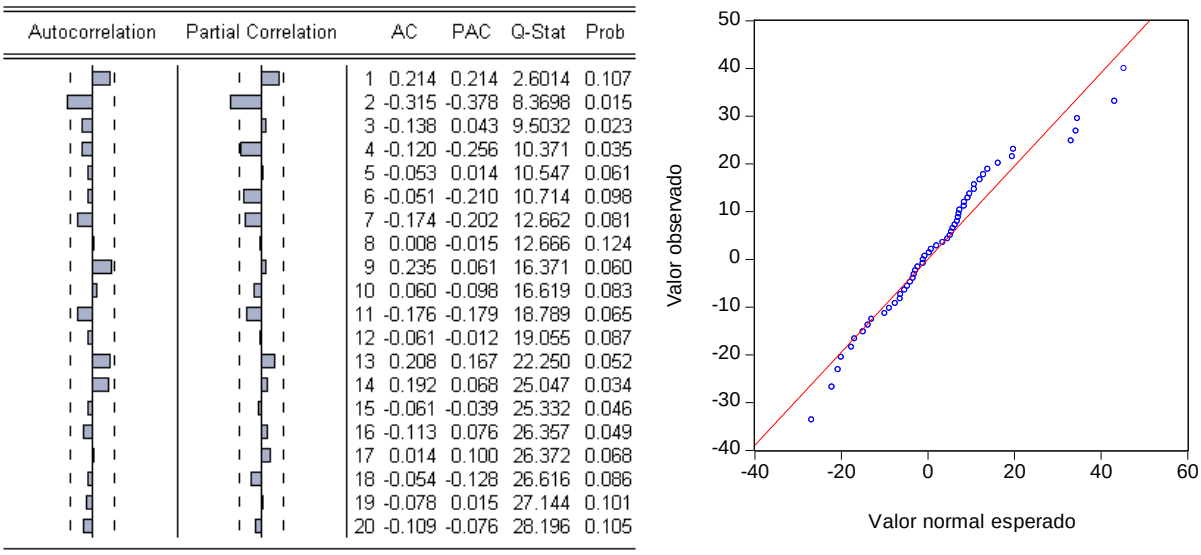
Figura 4 - Função de autocorrelação, autocorrelação parcial e gráfico de normalidade dos resíduos do modelo $SARIMA\ (1,0,0) \times (1,0,0)_{12}$ da série de hepatite A.

Com base nos correlogramas (Figuras 6 e 7) e os valores de p associados ao teste de Jarque-Bera (p-valor= 0,095 e p-valor=0,019, respectivamente) pode-se verificar que apenas os resíduos dos do modelo $SARIMA\ (1,0,0) \times (0,0,1)_{12}$ apresentaram comportamento de ruído branco, sendo selecionado para representar a série de leptospirose.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.033	0.033	0.0603	
		2	-0.012	-0.013	0.0680	
		3	0.120	0.121	0.9050	0.341
		4	0.027	0.019	0.9481	0.622
		5	-0.403	-0.408	10.806	0.013
		6	0.040	0.068	10.907	0.028
		7	0.172	0.206	12.788	0.025
		8	-0.208	-0.177	15.584	0.016
		9	-0.015	-0.006	15.598	0.029
		10	0.334	0.219	23.161	0.003
		11	-0.042	-0.028	23.286	0.006
		12	-0.078	0.024	23.716	0.008
		13	0.069	-0.142	24.059	0.012
		14	-0.109	-0.178	24.946	0.015
		15	-0.298	-0.001	31.767	0.003
		16	0.091	0.105	32.421	0.003
		17	-0.088	-0.258	33.049	0.005
		18	-0.110	0.007	34.064	0.005
		19	-0.021	-0.055	34.101	0.008
		20	0.108	-0.065	35.132	0.009

Figura 5 - Função de autocorrelação, autocorrelação parcial e gráfico de normalidade dos resíduos do modelo $SARIMA(1,0,0) \times (0,0,1)_{12}$ da série de hepatite A.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.214	0.214	2.6014	0.107
		2	-0.315	-0.378	8.3698	0.015
		3	-0.138	0.043	9.5032	0.023
		4	-0.120	-0.256	10.371	0.035
		5	-0.053	0.014	10.547	0.061
		6	-0.051	-0.210	10.714	0.098
		7	-0.174	-0.202	12.662	0.081
		8	0.008	-0.015	12.666	0.124
		9	0.235	0.061	16.371	0.060
		10	0.060	-0.098	16.619	0.083
		11	-0.176	-0.179	18.789	0.065
		12	-0.061	-0.012	19.055	0.087
		13	0.208	0.167	22.250	0.052
		14	0.192	0.068	25.047	0.034
		15	-0.061	-0.039	25.332	0.046
		16	-0.113	0.076	26.357	0.049
		17	0.014	0.100	26.372	0.068
		18	-0.054	-0.128	26.616	0.086
		19	-0.078	0.015	27.144	0.101
		20	-0.109	-0.076	28.196	0.105

Figura 6 - Função de autocorrelação, autocorrelação parcial e gráfico de normalidade dos resíduos do modelo $SARIMA\ (1,0,0) \times (0,0,1)_{12}$ da série de leptospirose.

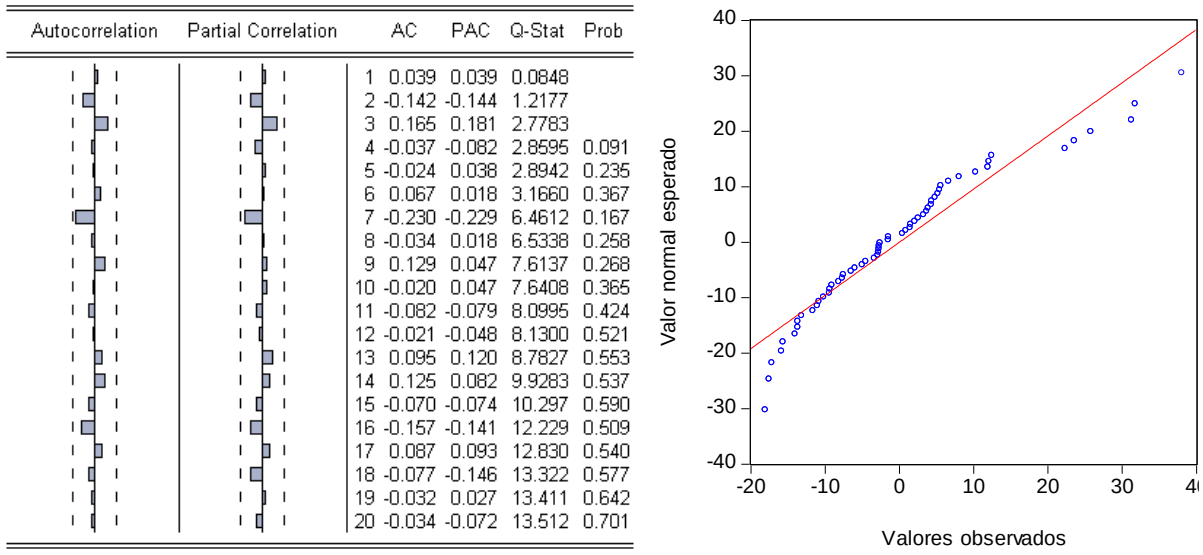


Figura 7 – Função de autocorrelação, autocorrelação parcial e gráfico de normalidade dos resíduos do modelo $SARIMA\ (2,0,0) \times (0,0,1)_{12}$ da série de leptospirose.

Na sequência, as Figuras 8 e 9 representam os gráficos com a série original, a série ajustada por meio do modelo selecionado e os resíduos para cada um dos agravos.

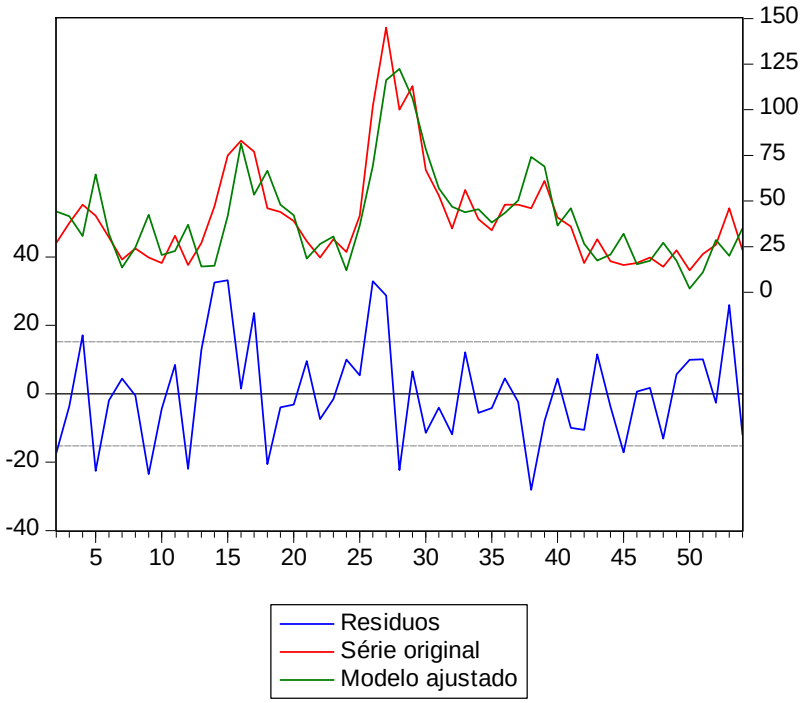


Figura 8 - Série original de hepatite A, modelo ajustado $SARIMA\ (1,0,0) \times (0,0,1)_{12}$ e resíduos.

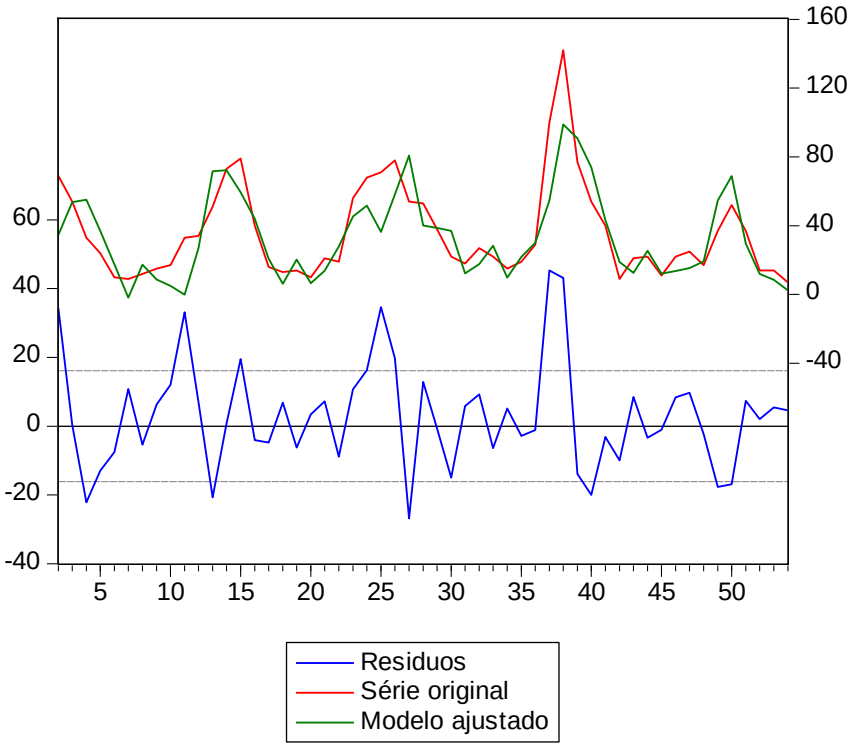


Figura 9 - Série original de leptospirose, modelo ajustado $SARIMA\ (1,0,0) \times (0,0,1)_{12}$ e resíduos.

Após a seleção dos modelos adequados para as séries de hepatite A e de leptospirose, procedeu-se a última etapa da metodologia de Box-Jenkins que consiste na realização de previsão para as séries que, neste caso, foi feita para seis instantes de tempo posteriores à observação 54.

Assim, feitas as previsões com auxílio do *software* estatístico, foram calculadas as medidas de desempenho para cada modelo que estão apresentadas na Tabela 5. Essas medidas serão usadas para comparação com outros modelos ajustados para as séries.

Com base nos procedimentos da metodologia de Box e Jenkins pode-se concluir que as séries estudadas apresentam uma componente sazonal de multiplicidade 12 relacionada ao elevado número de casos, principalmente no período de verão, levando ao ajuste de um modelo $SARIMA$.

Tabela 5 - Medidas de desempenho da previsão dos modelos selecionados para representar as séries de hepatite A e leptospirose do Rio Grande do Sul, Brasil, 2008 a 2012.

Medidas de desempenho	Hepatite A	Leptospirose
	$SARIMA$ $(1,0,0)(0,0,1)_{12}$	$SARIMA$ $(1,0,0)(0,0,1)_{12}$
MSE	8,097	15,856
MAE	7,3891	12,137
MAPE	52,058	47,104

Para a série de hepatite A dois modelos foram destacados dos demais por apresentaram os melhores critérios de seleção, mas os critérios AIC e BIC prevaleceram em detrimento do logaritmo de verossimilhança, e o modelo selecionado apresentou os melhores resultados no diagnóstico dos resíduos.

Quando analisados os modelos estimados para a série de leptospirose, o modelo $SARIMA\ (2,0,0) \times (1,0,0)_{12}$ apresentou o melhor valor para o critério do logaritmo da verossimilhança e também baixos valores para o AIC e BIC, no entanto, após a análise dos resíduos, verificou-se que este modelo não produziu ruídos brancos sendo portanto descartado.

Os modelos estimados permitiram uma previsão de dados de hepatite A e leptospirose para os seis meses suprimidos da série original. No entanto, na prática, a comparação entre os valores estimados e os ocorridos fica prejudicada devido às dificuldades em se trabalhar com dados de vigilância epidemiológica, como a alimentação tardia dos bancos de dados.

Entretanto, a aplicação da metodologia de Box e Jenkins mostra-se bastante útil para descrever séries de dados de vigilância epidemiológica, pois permitem a análise da tendência e previsão de número de casos futuros, a partir de dados passados, possibilitando a avaliação de impacto de intervenções populacionais.

5.3 Causalidade de Granger

Com o propósito de avaliar a existência de causalidade entre os agravos estudados, bem como estimar a direção desta ligação entre as variáveis, foi realizado o teste de Granger.

Na Tabela 6 estão apresentados os resultados referentes à causalidade de Granger entre as doenças, mostrando a direção da relação entre elas.

Hipóteses:

H_0: Doença “A” não-Granger-causa doença “B”

H_1: Doença “A” Granger-causa doença “B”.

Tabela 6 - Causalidade de Granger para as séries doenças hepatite A e leptospirose, Rio Grande do Sul, Brasil, 2008 a 2012.

Hipótese Nula (H_0)	n	Estatística F	p-valor
Hepatite A não-Granger-causa leptospirose	58	0,22064	0,8027
Leptospirose não-Granger-causa hepatite A	58	6,31109	0,0035

Analisando a tabela anterior, pode-se perceber que os resultados expostos revelam a relação de causalidade unidirecional da leptospirose para a hepatite A, ou seja, a leptospirose causa a hepatite A. Na área da saúde sabe-se, no entanto, que essas doenças não apresentam efetivamente relação de causa-consequência, mas devido ao fato de apresentarem vias de transmissão similares, essa causalidade pode ser interpretada da seguinte forma: se há um surto de leptospirose em uma determinada região do Rio Grande do Sul, espera-se que também ocorra um número considerável de casos de hepatite A, já o contrário não é verdadeiro considerando os resultados do teste de Granger.

5.4 Ajuste do modelo $ARMAX$

Com o objetivo de considerar a causalidade unidirecional existente entre os agravos estudados, estimou-se um modelo $ARMA$ ampliado, ou seja, com o acréscimo de uma variável exógena à equação. Neste caso, pode-se construir um modelo utilizando o número de casos de hepatite A (endógena) associado ao número de casos de leptospirose (exógena).

Durante o processo de modelagem foram utilizadas as iterações dos processo autorregressivo (AR) e de médias móveis (MA) variando de 0 a 6, os retornos da variável exógena e suas defasagens que também teve sua inclusão variando de 1 a 6.

Os modelos estimados, os quais apresentaram coeficientes significativos, se encontram no ANEXO C, e observando os critérios de eleição, o selecionado é composto por um parâmetro autorregressivo e a variável exógena defasada em uma e três vezes. No entanto este modelo foi descartado porque não apresentou boas propriedades na verificação dos resíduos. Assim procedeu-se a análise do segundo modelo que apresentou melhores critérios de seleção. Na Tabela 7 constam os principais resultados do modelo eleito.

Tabela 7 - Coeficientes do modelo $ARMAX$ selecionado.

Termo	Coeficiente	Estatística T	p-valor
MA(1)	0,9273	7,2585	<0,001
MA(2)	0,5002	3,8907	<0,001
X: L(-1)	0,3210	2,5112	0,0156
X: L(-3)	0,3240	2,5450	0,0143
C	17,2577	1,9337	0,0593

A partir do modelo selecionado procedeu-se a análise dos resíduos por meio da análise da função de autocorrelação e autocorrelação parcial, além do e gráfico da normalidade dos resíduos conforme constam na Figura 10.

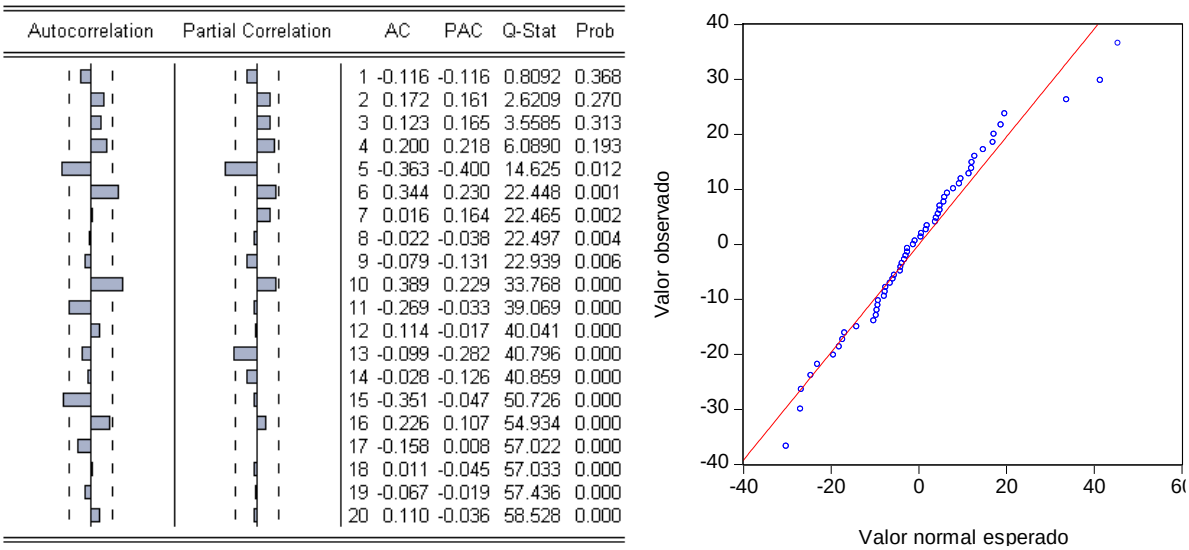


Figura 10 - Função de autocorrelação, autocorrelação parcial e gráfico de normalidade dos resíduos do modelo $ARMAX$da série de hepatite A.

Com base nos gráficos e no teste de Jarque-Bera (p-valor=0,080) pode-se considerar que os resíduos do modelo escolhido é adequado.

Na Figura 11 estão apresentadas a série original de leptospirose, os valores preditos pelo modelo e os respectivos resíduos.

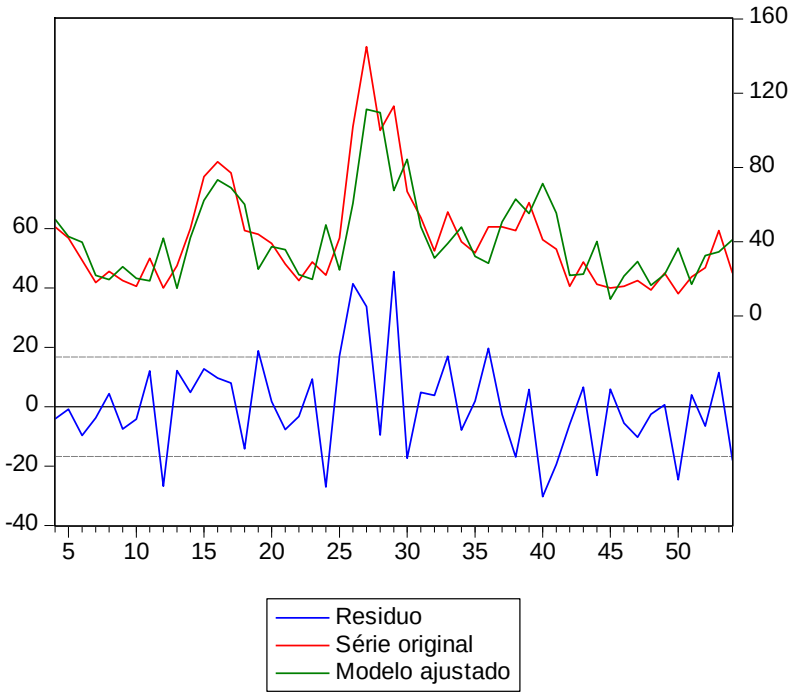


Figura 11 - Série original de leptospirose, modelo ajustado $SARIMA\ (1,0,0) \times (0,0,1)_{12}$ e resíduos.

As medidas de desempenho dos valores previstos para as 6 últimas observações da série de hepatite A são: MSE=10,182; MAE=9,543 e MAPE=93,570.

Uma vez que o teste de causalidade de Granger aponta para existência unidirecional entre os agravos, não foi possível estimar um modelo para a leptospirose utilizando a série de hepatite A como variável exógena.

5.5 Ajuste do modelo VAR

A análise da estacionariedade das séries já foi investigada anteriormente, segue-se então para próxima etapa que consiste em definir o número de defasagens para que o VAR apresente as melhores características de previsão. Assim, como se observa na Tabela 8 e de acordo com os critérios AIC e BIC, o comprimento apropriado para a defasagem do modelo autorregressivo vetorial bivariado (VAR) é de um lag.

Tabela 8 - Definição do número de defasagens do modelo VAR.

Lag	AIC	BIC	ln L
1	8,4641*	8,5756*	-221,2983*
2	8,5310	8,7185	-216,8037
3	8,5999	8,8651	-212,2978
4	8,9795	9,0207	-207,9137
5	8,7592	9,1839	-203,5997

* Melhor lag Segundo os critérios penalizadores, AIC: Akaike information criterion; BIC: Bayesian information criterion; ln L: Log Likelihood. Fonte: a autora (2014), a partir do software E-views 7.

Na Tabela 9 observa-se os coeficientes do modelo VAR onde H(-1) e L(-1) representam as variáveis defasadas da série de hepatite A e leptospirose respectivamente. O erro padrão, estatística t e o p-valor estão apresentados ao lado do correspondente coeficiente para um nível de significância de 5%.

O modelo de vetor autorregressivo consiste em um sistema de equações, em que cada variável que compõem o sistema está em função dos valores das demais variáveis defasadas no tempo mais o termo de erro. Após o ajuste do modelo segue-se com a análise dos resíduos. Os testes de diagnóstico de resíduos dos modelos univariados podem ser generalizados para o caso multivariado e segundo Bueno (2008) essa generalização é praticamente direta e evidente.

Tabela 9 - Estimativa do modelo de vetor autorregressivo.

Série	Variáveis	Coeficiente	Erro Padrão	Estatística t	p-valor
H	H(-1)	0,6937	0,0849	8,1970	<0,01
	L(-1)	0,2807	0,0840	3,3415	<0,01
	C	2,0108	4,5996	0,4372	0,663
L	H(-1)	-0,1337	0,0992	-1,3485	0,181
	L(-1)	0,7469	0,0981	7,6156	<0,01
	C	13,5870	5,3698	2,5303	0,013

A análise de resíduos foi realizada baseada no teste de autocorrelação de Ljung-Box e de normalidade individual e conjunta, que tem a mesma lógica do modelo univariado, sendo necessário verificar se as autocorrelações multivariadas são nulas. Alternativamente, usou-se a estatística de Ljung-Box ajustada, a qual apresenta melhores propriedades para pequenas amostras. Na Figura 12 estão apresentados a Função de Autocorrelação conjunta do modelo estimado. O símbolo df representa o número de graus de liberdade para a distribuição χ^2.

Dessa forma, foi possível constatar que as autocorrelações multivariadas decaem rapidamente, de modo que o teste de Ljung-Box rejeita a hipótese nula de existência de autocorrelação serial.

Lags	Q-Stat	Prob.	Adj Q-Stat	Prob.	df
1	5.957581	NA*	6.072150	NA*	NA*
2	7.632981	0.1060	7.813252	0.0987	4
3	8.798926	0.3595	9.049154	0.3382	8
4	12.28944	0.4227	12.82461	0.3819	12
5	17.37180	0.3619	18.43638	0.2990	16
6	19.52603	0.4879	20.86562	0.4051	20
7	23.18144	0.5091	25.07728	0.4016	24
8	24.04165	0.6793	26.09042	0.5681	28
9	25.88801	0.7685	28.31444	0.6537	32
10	37.19050	0.4140	42.24542	0.2192	36
11	50.89155	0.1160	59.53484	0.0240	40
12	57.33881	0.0855	67.86911	0.0119	44

Figura 12 – Função de autocorrelação multivariada para os resíduos do modelo VAR.

Quanto às autocorrelações individuais apresentadas nas Figuras 13 e 14 para a série de hepatite A e leptospirose, respectivamente, julgou-se que o modelo seria aceitável apenas para primeira série

No que diz respeito à normalidade dos resíduos, foi possível constatar que tanto individualmente quanto conjuntamente os resíduos não são normais, rejeitando em todos os casos a hipótese nula de normalidade uma vez que o p-valor para o teste Jarque-Bera foi menor do que 1% nas diferentes situações.

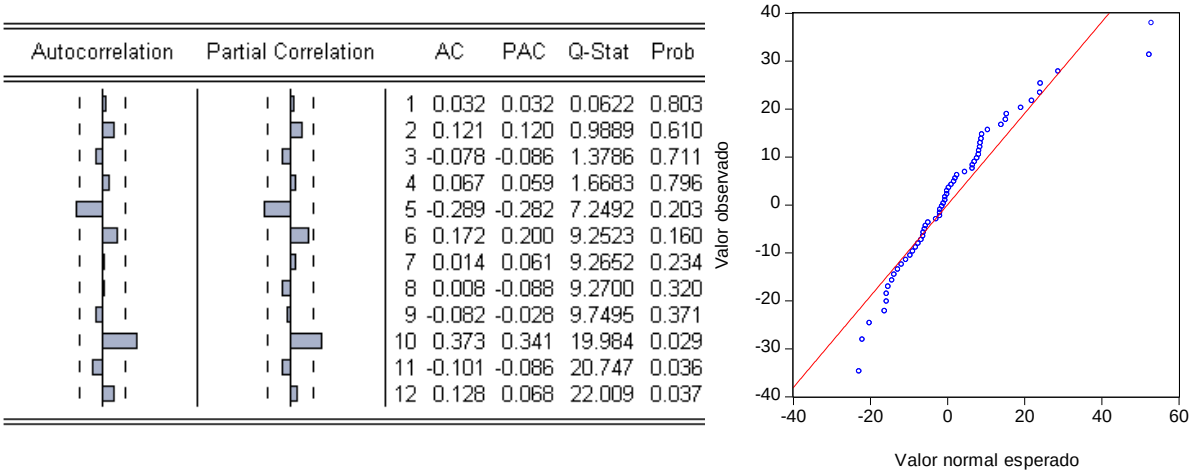


Figura 13 - Função de autocorrelação e autocorrelação parcial dos resíduos da equação da variável hepatite A do modelo VAR.

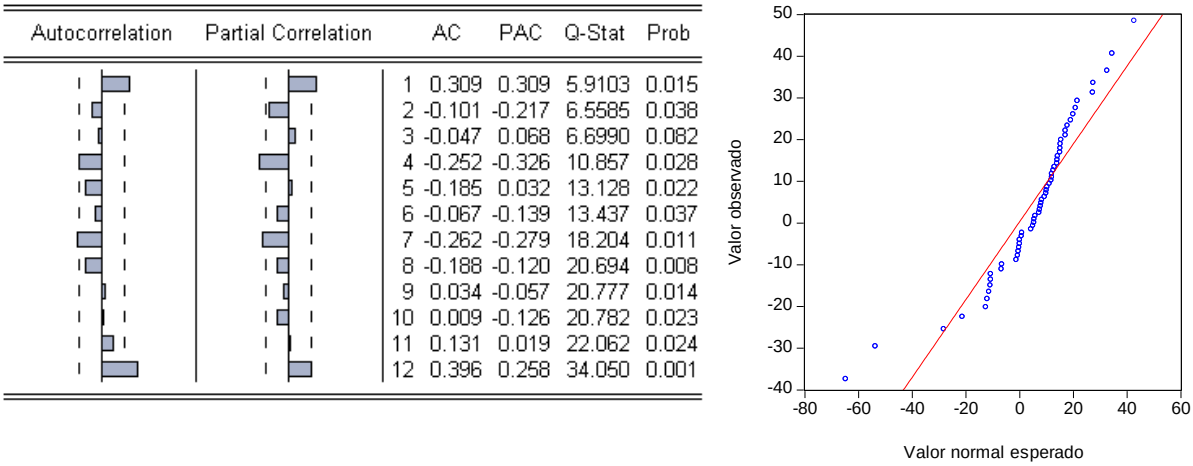


Figura 14 - Função de autocorrelação e autocorrelação parcial dos resíduos da equação da variável leptospirose do modelo VAR.

Embora os resíduos não tenham apresentado boas propriedades, foram feitas as previsões e estimados os valores das medidas de desempenho para a série de hepatite A. Os resultados estão apresentados na Tabela 10 juntamente com as medidas calculadas anteriormente.

Tabela 10 - Medidas de desempenho da previsão dos diferentes modelos estimados e selecionados para representar as séries de hepatite A do Rio Grande do Sul, Brasil, 2008 a 2012.

Medidas de desempenho	$SARIMA$	$ARMAX$	VAR
MSE	8,097	10,182	6,475
MAE	7,3891	9,543	5,234
MAPE(%)	52,058	93,570	57,800

Neste estudo optou-se por analisar as medidas de desempenho tradicionalmente utilizadas para avaliar a qualidade das previsões em séries temporais. Essas medidas são baseadas no erro entre os valores previstos e os observados. Assim, qualquer dessas medidas de desempenho compara uma série de pares de valores (previstos e observados) e pondera as diferenças de uma forma particular. Os modelos que geram previsões com os menores valores de erro são considerados melhores.

Os modelos selecionados apresentaram valores elevados para as medidas de desempenho. Entre os modelos $SARIMA$ e $ARMAX$ observam-se valores próximos para as medidas MSE e MAE, já o MAPE expressa uma diferença maior entre os modelos, indicando que o $SARIMA$ tem o melhor desempenho de previsão dentre os modelos estudados.

O modelo $ARMAX$ pode ter sua capacidade de previsibilidade melhorada após a inclusão de outras variáveis exógenas previamente avaliadas. Neste caso em específico, série de índices pluviométricos ou séries de outras doenças, como cólera e dengue, são passíveis de serem testadas para incorporar no modelo como variáveis explicativas.

O modelo VAR apresentou valores de MSE e MAE menores do que o $SARIMA$, e o MAPE bastante similar ao modelo anterior. Porém, conforme os resultados apresentados, o modelo não gerou resíduos com características de ruído branco e as medidas de desempenho foram avaliadas somente para a equação correspondente à variável hepatite A. Esse modelo é comumente aplicado à área da economia, pois sua forma estrutural permite que se expressem modelos econômicos completos. De forma análoga, pode ser bastante útil na área epidemiológica e necessita de estudos aplicados para que discussões a respeito de seu desempenho possam ser realizadas.

Rothman e Greenland (1998), afirmam que a comparação de séries temporais na área da saúde, com o propósito de se verificar a possível relação entre as mesmas é de suma importância no campo da Epidemiologia. Salientam também que, embora a maior perspectiva seja no sentido de elucidar relações de causalidade, a descrição do comportamento simultâneo de séries históricas epidemiológicas tem grande importância, até mesmo como um indicativo de possíveis associações, mesmo que não necessariamente causais.

6 CONCLUSÃO E CONSIDERAÇÕES FINAIS

A análise de dados epidemiológicos auxilia, substancialmente, a avaliação de serviços ou políticas de promoção da saúde. A metodologia empregada nessas análises deve incorporar a dinamicidade das ocorrências dos casos e também considerar algumas limitações, como por exemplo, a demora na divulgação das informações a cerca das doenças.

As séries aqui estudadas foram obtidas em banco de dados do DATASUS (Ministério da Saúde) as quais são estruturadas em unidades de informação mensal. As séries de casos de hepatite A e leptospirose do Rio Grande do Sul, compreendidas no período de 2008 a 2012, apresentaram-se estacionárias, porém com comportamento sazonal.

A sazonalidade da série deve-se ao fato de que essas doenças de veiculação hídrica ocorrem com maior frequência no verão devido ao alto índice pluviométrico da estação (VILLAR *et. al*, 2002).

O estudo mostra que é possível estabelecer relação de causa entre os agravos estudados e, embora efetivamente não exista associação de causa-consequência na área da saúde entre as doenças, esse tipo de informação é de suma importância em investigações epidemiológicas.

A questão da causalidade em Epidemiologia é um assunto amplamente discutido na literatura e, ao longo dos anos, passou por algumas modificações filosóficas (CZERESNIA e ALBUQUERQUE, 1995; BARATA, 1997; ROTHMAN e GREENLAND, 2005; BONITA *et. al*, 2010; LUIZ e STRUCHINER, 2012). Neste trabalho, a causalidade encontrada remete-se apenas a um achado estatístico e que, se estudada de forma aprofundada, pode sugerir uma relação de causa no sentido de que uma pessoa que se encontra em ambiente/região que apresente elevado número de casos de leptospirose está também exposta a fatores de risco para a hepatite A.

A relação de causalidade unidirecional permitiu o ajuste de modelos $ARIMAX$ e VAR para a série de hepatite A e modelos $SARIMA$ foram ajustados para as duas séries. Dentre os modelos analisados, o modelo $SARIMA$ apresentou as melhores medidas de desempenho para a previsão de novos casos para a série de hepatite A.

A adoção de metodologias de séries temporais em contextos epidemiológicos pode constituir um importante suporte de orientação e apoio para o monitoramento de dados de vigilância epidemiológica. Neste estudo foi possível verificar, com base na revisão de literatura, um número crescente de trabalhos com enfoque aplicado à prática da análise de dados de vigilância. A maioria dessas investigações utiliza os modelos univariados, que embora sejam importantes, pode prejudicar a dinamicidade, a possibilidade da intervenção radical ou ainda omitir fenômenos de causalidade nos processos saúde-doença.

É importante ressaltar que, em vigilância de saúde, o tempo é essencial para a detecção precoce de epidemias/surtos para que as medidas de controle sejam adotadas oportunamente, de modo que grande número de casos e óbitos possam ser prevenidos.

Para pesquisas futuras sugere-se:

- Estimar modelos com intervenção para analisar se alguns pontos discrepantes estão influenciando no ajuste adequado e nas previsões de novos casos das doenças.

- Avaliar se modelos com graus de diferenciação fracionários (ARFIMA) apresentam melhores ajustes para as séries de doenças citadas neste estudo.

- Investigar se outras variáveis exógenas, tal como índice pluviométrico, podem ser adicionadas ao modelo das doenças aqui trabalhadas.

- Verificar se resultados semelhantes ocorrem em outras regiões identificando se na relação causal a leptospirose pode ser entendida como fator reforçador da ocorrência de hepatite A.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, R.M.; GRENFELL, B.T.; Oscillatory fluctuations in the incidence of infectious disease andthe impact of vaccination: time series analysis. **Journal of Hygiene (Cambridge)**, v. 93, p. 587-608, 1984.

ANTUNES, J.L.F. *et al.* Effectiveness of influenza vaccination and its impact on health inequalities. **International Journal of Epidemiology**, v. 36, p.1319-1326, 2007.

BARATA, R.B. Causalidade e Epidemiologia. **História, Ciências, Saúde - Manguinhos**, v.4, n.1, p.31-49, 1997.

BHASKARAN, K. *et al.* Time series regression studies in environmental epidemiology. **International Journal of Epidemiology**, v. 42, p.1187-1195, 2013.

BELL, M. L.; SAMET, J. M.; DOMINICI, F. Time-series studies of particulate matter. **Annual Review Public Health**, v. 25, p.247-280, 2004.

BONITA, R.; BEAGLEHOLE, R.; KJELLSTRÖM, T. **Epidemiologia Básica**. 2.ed., São Paulo: Livraria Editora Santos, 2010. 213p.

BOX, G.E.P., JENKINS, G.M. **Time series analysis-forecasting and control**, San Francisco: Holden-Day, 1976. 575p.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância Epidemiológica**. A, B, C, D, E de hepatites para comunicadores**. Brasília: MS, 2005. 24 p. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/hepatites_abcde.pdf> Acesso em 14 de out. 2013.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Guia de Vigilância Epidemiológica**. Brasília - DF: Ministério da Saúde, 2005

BRASIL, Ministério da Saúde. **Guia Leptospirose: Diagnóstico de Manejo Clínico** São Paulo: MS, 2009, 34 p. Disponível em: <ftp://ftp.cve.saude.sp.gov.br /doc_tec/ZOO/LEPTO09_GUIA_MANEJO.pdf> Acesso em 14 de out. 2013

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Sinan** [dados na Internet]. Brasília: MS. Disponível em: <http://dtr2004.saude.gov.br/sinanweb/>. Acesso em: 10 de set. 2013.

BRIËT, O.J.T. *et al.* Models for short term malaria prediction in Sri Lanka. **Malaria Journal**, v. 7, n. 76, p.1-11, 2008.

BUENO, R.L.S. **Econometria de séries temporais** São Paulo: Cengage Learning, 2008. 299p.

CAO *et al*. A hybrid seasonal prediction model for tuberculosis incidence in China. **PLoS One**, v. 13, n. 56, 2013.

CHARENZA, W.W.; DEADMAN, D.F. **New directions in econometric practice: general to specific modeling, cointegration, and vector autorregression**. 2 ed. United Kingdom: Edward Elgar Publishing, 1997.344p.

CHEN, M.P. *et al.* A Bayesian analysis of the 2009 decline in tuberculosis morbidity in theUnited States. **Statistics in Medicine**, v. 31, p.3278–3284, 2012.

CHOI, K.; THAKER, S.B. An evaluation of influenza mortality surveillance 1962-1979.Times series forecasts of expected pneumonia and influenza death. **American Journal Epidemiology**, v. 113, p. 215-226, 1981.

CHOWELL, G. *et al.* The spatial and temporal patterns of falciparum and vivax malaria in Perú: 1994–2006. **Malaria Journal**, v. 8, n. 142, p. 1-19, 2009.

COLLOPY F.; ADYA M.; ARMSTRONG, J.S. Principles of Examining Predictive Validity: The Case of Information Systems Spending Forecasts. **Information Systems Research**, v. 5, p. 170-179, 1994.

CZERESNIA, D.; ALBUQUERQUE, M.F.M. Modelos de inferência causal: análise crítica da utilização da estatística na Epidemiologia. **Revista de Saúde Pública**, v. 29, n. 5, p. 415-423, 1995.

DOM, N.C. *et al.* Generating temporal model using climate variables for the prediction of dengue cases in Subang Jaya, Malaysia. **Asian Pacific Journal of Tropical Disease**, v. 3, n. 5, p. 352-361, 2012.

DOMINGUEZ, A. Monitoring mortality as an indicator of influenza in Catalonia, Spain. **Journal of Epidemiology and Community Health**, v. 50, p. 293-298, 1996.

EARNEST, A. *et al.* Comparing Statistical Models to Predict Dengue Fever Notifications. **Computational and Mathematical Methods in Medicine**, v. 2012, p. 1-6, 2012.

FERNÁNDEZ-PÉREZ, C.; TEJADA, J.; CARRASCO, M. Multivariate time series analysis in nosocomial infection surveillance: a case study. **International Journal of Epidemiology**, v. 27, p. 282-288, 1998.

FONSECA, M.G. *et al.* AIDS e grau de escolaridade no Brasil: evolução temporal de 1986 a 1996. **Cadernos de Saúde Pública**, v. 16, sup. 1, p. 77-87, 2000.

GIRARD, D.Z. The cost of epidemiological transition: A study of a decrease in pertussis vaccination coverage. **Health Policy**, v. 74 p. 287-303, 2005.

GOLDSTEIN, E. *et al.* Reconstructing influenza incidence by deconvolution of daily mortality time series. **PNAS**, v. 106, n. 51, p. 21825-21829, 2009.

GUJARATI, D.N.; PORTER, D. C. **Econometria Básica**, 5. ed. Bookman: Porto Alegre, 2011. 924 p.

HANF, M. *et al.* The role of El Niño southern oscillation (ENSO) on variations of monthly *Plasmodium falciparum* malaria cases at the cayenne general hospital, 1996-2009, French Guiana. **Malaria Journal**, v. 10, p. 100, p. 1-4, 2011.

HUANG, F. *et al.* Temporal correlation analysis between malaria and meteorological factors in Motuo County, Tibet. **Malaria Journal**, v. 10, n. 54, p. 1-8, 2011.

JEKEL J.F. **Epidemiologia, Bioestatística e Medicina Preventiva**. Porto Alegre: Artes Médicas, 1999.

LARA-RAMÍREZ, E.E. *et al.* Time Series Analysis of Onchocerciasis Data from Mexico:A Trend towards Elimination. **PLoS Neglected Tropical Diseases**, v. 7, n. 2, 2013.

LIN, H. *et al.* Spatial and temporal distribution of falciparum malaria in China. **Malaria Journal**, v. 8, n. 130, p. 1-9, 2009.

LIN, H. *et al.* Time series analysis of Japanese encephalitis and weather in Linyi City, China. **International Journal of Public Health**, v. 57, p. 289-296, 2012.

LIU, Q. *et al.* Forecasting incidence of hemorrhagic fever withrenal syndrome in China using $ARIMA$ model. **BioMedCentral Infectious Diseases**, v. 11, n. 218, 2011.

LUIZ, R.R.; STRUCHINER, C.J. **Inferência causal em Epidemiologia: o modelo de respostas potenciais** [online]. Rio de Janeiro: Editora FIOCRUZ, 2002.

LUZ, P.M. *et al.* Time Series Analysis of Dengue Incidence in Rio de Janeiro, Brazil. **American Journal of Tropical Medicine and Hygiene**, v. 79, n. 6, p. 933–939, 2008.

MADDALA, G.S. **Introdução à Econometria**. 3.ed. Rio de Janeiro: LTC, 2003. 345p.

MATOS, O.C. **Econometria Básica: teoria e aplicações**. 3.ed. São Paulo: Atlas, 2000. 300p.

MARTINEZ, E.Z.; SILVA, E.A.S. Predicting the number of cases of dengue infection in Ribeirão Preto, São Paulo State, Brazil, using a $SARIMA$ model. **Cadernos de Saúde Pública**, v. 27, n. 9, p. 1809-1818, 2011.

MILOSEVIC, V. *et al.* Hospitalizations due to spontaneous intracerebral hemorrhage in the region of Nis (Serbia): 11-year time-series analysis. **Clinical Neurology and Neurosurgery**, v. 113, p. 552- 555, 2011.

MODARRES, R. *et al.* Modeling seasonal variation of hip fracture in Montreal, Canada. **Bone**, v. 50, p. 909-916, 2012.

MORETTIN, P.A.; TOLOI, C.M.C. **Previsão de Séries Temporais. São Paulo**: Atual Editora, 2006.

MUÑOZ-TUDURÍ, M.; GARCÍA-MORO, C.; WALKER, P.L. Time Series Analysis of the Epidemiological Transition in Minorca, 1634–1997. **Human Biology**, v. 78, n. 5, p. 619-634, 2006.

OTERO, U.B.; ROZENFELD, S.; GADELHA, A.J. Óbitos por desnutrição em idosos, São Paulo e Rio de Janeiro. Análise de séries temporais. 1980-1996. **Revista Brasileira de Epidemiologia**, v. 4, n. 3, p. 191-205, 2001.

PORTA, M.; GREENLAND, S.; LAST, J.M.A. **Dictionary of epidemiology**. 5 ed. New York: Oxford University Press, 2008.

PEREIRA, Z.L.; REQUEIJO, J.G. **Qualidade: Planejamento e Controle Estatístico de Processo.** Lisboa: Prefácio, 2008.

REN, H. *et al.* The development of a combined mathematical model to forecast the incidence of hepatitis E in Shanghai, China. **BioMed Central Infectious Diseases**, v. 13, n. 421, 2013.

ROTHMAN, K.J.; GREENLAND, S. Causation and Causal Inference in Epidemiology **American Journal of Public Health**, v. 95, n. S1, p. 144-150, 2005.

SACHS, J.; MALANEY, P. The economic and social burden of malaria. **Nature** v. 415, n. 6972, p. 680-685, 2002.

SILAWAN, T. *et al.* Temporal patterns and forecast of dengue Infection in northeastern Thailand. **Southeast Asian Journal Tropical Medicine Public Health**, v. 39, n. 1, p. 90-98, 2008.

SIMS, C. Macroeconomics and Reality. **Econometrica**, v. 48, p. 1-48, 1980.

SOEBIYANTO, R.P.; ADIMI, F.; KIANG. R.K. Modeling and Predicting Seasonal Influenza Transmission in Warm Regions Using Climatological Parameters. PLoS ONE, v. 5, n. 3, p. 1-10, 2010.

TIAN, L. *et al.* One-year delayed effect of fog on malaria transmission: a time-series analysis in the rain forest area of Mengla County, south-west China. **Malaria Journal**, v. 7, n. 110 2008, p. 1-9, 2008.

UPSHUR, R.E.G.; KNIGHT, K.; GOEL, V. Time-Series Analysis of the Relation between Influenza Virus and Hospital Admissions of the Elderly in Ontario, Canada, for Pneumonia, Chronic Lung Disease, and Congestive Heart Failure. **American Journal of Epidemiology**, v. 149, n. 1, p. 85-92, 1999.

VASCONCELLOS, M.A.S.; ALVES, D. (Orgs.). **Manual de Econometria**. 1 ed. São Paulo: Atlas, 2000.

VILLAR, L.M.; DE PAULA, V.S.; GASPAR A.M.G. Seasonal variation of hepatitis a virus infection in the city of Rio de Janeiro, Brazil. **Revista do Instituto de Medicina Tropical de São Paulo**, v. 44, n. 5, p. 289-292, 2002.

ZHANG, X. *et al.* Comparative study of four time series methods in forecasting typhoid fever incidence in china. **PLoS One**, v. 8, n. 5, 2013.

ZHANG, X. *et al.* Applications and comparisons of four time series models in epidemiological surveillance data. **PLoS One**, v. 9, n. 2, 2014.

WANGDI, K. *et al.* Development of temporal modelling for forecasting and prediction of malaria infections using time-series and $ARIMAX$ analyses: A case study in endemic districts of Bhutan. **Malaria Journal**, v. 9, n. 251, 2010.

WILLIAMSON, G.D.; HUDSON, G.W. A monitoring system for detecting aberrations in public health surveillance reports. **Statistics in Medicine**, v. 18, p. 3283-3298,1999.

WHO. **Dengue guidelines for diagnosis, treatment, prevention and control**. France: WHO and TDR Publication; 2009. Disponível em: <http://whqlibdoc.who.int/publications/2009/9789241547871_eng.pdf> Acesso em 06 de Outubro de 2013.

WONGKOON, S.; JAROENSUTASINEE, M.; JAROENSUTASINEE, K. Development of temporal modeling for prediction of dengue infection in Northeastern Thailand. **Asian Pacific Journal of Tropical Medicine**, v. 2012, p. 249-252, 2012.

YE, Y.; KERR, W.C. Alcohol and Liver Cirrhosis Mortality in the United States: Comparison of Methods for the Analyses of Time-Series Panel Data Models. **Alcoholism: Clinical and Experimental Research**, v. 35, n. 1, p. 108-115, 2011.

ANEXOS

Anexo A - Parâmetros significativos e critérios de informação dos modelos estimados para série de dados observados de hepatite A, Rio Grande do Sul, Brasil, 2008 a 2012.

Parâmetros	AR (1)	AR (1)	$SARIMA$ $(1,0,0)(1,0,0)_{12}$
Termo AR(1)	0,9294	0,7723	0,8745
Termo AR(12) sazonal	-	-	0,4767
Termo MA(12) Sazonal	-	-	-
Constante	-	40,9630	-
AIC	8,673	8,628	8,687
BIC	8,711	8,702	8,771
ln L	-228,848	-226,640	-176,085

Parâmetros	$SARIMA$ $(1,0,0)(1,0,0)_{12}$	$SARIMA$ $(1,0,0)(0,0,1)_{12}$	$SARIMA$ $(1,0,0)(0,0,1)_{12}$
Termo AR(1)	0,7911	0,8773	0,7759
Termo AR(12) sazonal	0,4244	-	-
Termo MA(12) Sazonal	-	0,8629	0,8635
Constante	44,0400	-	37,8689
AIC	8,688	8,836	8,338
BIC	8,814	8,434	8,450
ln L	-175,110	-219,540	-217,958

Anexo B - Parâmetros significativos e critérios de informação dos modelos estimados para série de dados observados de leptospirose, Rio Grande do Sul, Brasil, 2008 a 2012.

Parâmetros	$AR(1)$	$AR(1)$	$ARMA$ (1,1)	$SARIMA$ $(1,0,0)(1,0,0)_{12}$
Termo AR (1)	0,8734	0,7103	0,7801	0,7566
Termo AR (12) sazonal	-	-	-	0,4086
Termo MA (12)	-	-	0,5825	-
Termo MA (12) Sazonal	-	-	-	-
Constante	-	32,5611	-	-
AIC	8,819	8,772	8,677	8,768
BIC	8,856	8,846	8,751	8,844
ln L	-232,704	-230,450	-227,931	-225,980

Parâmetros	$SARIMA$ $(1,0,0)(0,0,1)_{12}$	$SARIMA$ $(2,0,0)(1,0,0)_{12}$	$SARIMA$ $(2,0,0)(0,0,1)_{12}$	$SARIMA$ $(0,0,1)(1,0,0)_{12}$
Termo AR (1)	0,8100	0,8958	0,9248	-
Termo AR (2)	-	-0,3860	-0,3854	-
Termo AR (12) sazonal	-	0,4352	-	0,3907
Termo MA (1)	-	-	-	0,9408
Termo MA (12) Sazonal	0,9020	-	0,8944	-
Constante	-	34,1888	30,5603	35,8680
AIC	8,438	8,665	8,126	8,580
BIC	8,512	8,883	8,276	8,704
ln L	-221,612	-169,303	-207,28	-177,172

Anexo C - Parâmetros significativos e critérios de informação dos modelos estimados para série de dados observados de hepatite A utilizando a série leptospirose como variável exógena.

Variáveis	AIC	BIC	ln L
AR (1), X: L (-1)	8,6378	8,7128	-222,5818
AR (1), X: L (-1), com constante	8,5732	8,6858	-219,9033
AR (1), X: L (-1), L (-3)	8,5332	8,6479	-210,3296
AR (1), X: L (-2), com constante	8,5983	8,7120	-216,2562
MA (1), X: L (-1)	9,0705	9,1448	-238,3671
MA (1), X: L (-1), com constante	8,8523	8,9638	-231,5852
MA (1), X: L (-1) L (-2)	8,9440	9,0567	-229,5452
MA (1) MA (2), X: L (-1), L (-2)	8,7695	8,9196	-224,0056
MA (1), MA (2), X: L (-1 L (-3)	8,6066	8,7581	-215,4678
MA (1), MA (2), X: L (-1), L (-3), com constante	8,5688	8,7582	-213,5039
MA (1), MA (2), X: L (-1), L (-4), com constante	8,6190	8,8103	-210,4763