

Ministério da Saúde  
Fundação Oswaldo Cruz  
Escola Nacional de Saúde Pública

ESTUDO DA INFLUÊNCIA DE VARIÁVEIS METEOROLÓGICAS NO  
APARECIMENTO DE CASOS GRAVES DE LEPTOSPIROSE EM SALVADOR – BA  
VIA MODELOS DE SÉRIES TEMPORAIS

Aluno: Antonio Guilherme Fonseca Pacheco

**Orientador: Claudio José Struchiner**

**Membros da Banca:**

**Albert Ko**

**Claudio José Struchiner**

**Margareth Portela**

# Índice

<b>ÍNDICE</b> .....	<b>2</b>
<b>LISTA DE TABELAS</b> .....	<b>4</b>
<b>LISTA DE FIGURAS</b> .....	<b>5</b>
<b>ABSTRACT</b> .....	<b>9</b>
<b>ABSTRACT</b> .....	<b>9</b>
<b>RESUMO</b> .....	<b>10</b>
<b>INTRODUÇÃO</b> .....	<b>11</b>
<b>OBJETIVOS</b> .....	<b>15</b>
<b>CONSIDERAÇÕES ÉTICAS</b> .....	<b>16</b>
<b>MATERIAIS E MÉTODOS</b> .....	<b>18</b>
CASOS GRAVES DE LEPTOSPIROSE .....	18
VARIÁVEIS METEOROLÓGICAS.....	20
MODELOS DE SÉRIES TEMPORAIS.....	20
• <i>Ruído Branco</i> .....	20
• <i>Operador “Back Shift”</i> .....	21
• <i>Passeio Aleatório</i> .....	21
• <i>Médias Móveis</i> .....	22
• <i>Autorregressão</i> .....	23
• <i>Modelo ARMA<sub>(p,q)</sub> (Autoregressive-Moving Average)</i> .....	25
• <i>Modelo ARIMA<sub>(p,d,q)</sub> (Autoregressive-Integrated Moving Average)</i> .....	25
• <i>Modelo SARIMA<sub>(p,d,q)×(P,D,Q)</sub> (Seasonal Autoregressive-Integrated Moving Average)</i> .....	26
• <i>Modelos com Co-variáveis Regressionais</i> .....	26
• <i>Modelos com Regressores Dinâmicos</i> .....	27
• <i>Modelos de Análise de Intervenção</i> .....	28
DESCRIÇÃO DAS SÉRIES .....	28
MÉTODOS DE IDENTIFICAÇÃO DE SÉRIES TEMPORAIS .....	29
• <i>Séries Simples</i> .....	29
• <i>Séries Múltiplas</i> .....	31
• <i>Correlogramas</i> .....	32
ESTIMAÇÃO DOS PARÂMETROS DOS MODELOS DE SÉRIES TEMPORAIS .....	33
MÉTODOS PARA AVALIAÇÃO DOS MODELOS DE SÉRIES TEMPORAIS .....	34
PREVISÃO UTILIZANDO MODELOS DE SÉRIES TEMPORAIS .....	36
CORRELAÇÃO ENTRE AS SÉRIES E IDENTIFICAÇÃO DAS FUNÇÕES DE TRANSFERÊNCIA .....	36
AJUSTE DOS MODELOS COM REGRESSORES DINÂMICOS E MODELOS DE INTERVENÇÃO .....	39
MÉTODOS PARA AVALIAÇÃO DOS MODELOS COM REGRESSORES DINÂMICOS E MODELOS DE INTERVENÇÃO ..	40
<b>RESULTADOS</b> .....	<b>41</b>
DESCRIÇÃO DAS SÉRIES .....	41
• <i>Gráficos Temporais das Séries</i> .....	41
• <i>Transformação das Séries</i> .....	42
• <i>Sazonalidade das Séries</i> .....	43
IDENTIFICAÇÃO E AJUSTE DOS MODELOS PARA AS SÉRIES TEMPORAIS .....	44
• <i>A Série de Casos</i> .....	45
• <i>A Série de Pluviosidade</i> .....	47
• <i>A Série de Umidade</i> .....	48
• <i>As Séries de Temperatura Mínima, Média e Máxima</i> .....	49
CORRELAÇÃO ENTRE AS SÉRIES .....	50
IDENTIFICAÇÃO E AJUSTE DOS MODELOS COM FUNÇÕES DE TRANSFERÊNCIA .....	52
• <i>Regressores Dinâmicos e o Modelo ARIMA(1,1,1)</i> .....	54

• <i>Regressores Dinâmicos e o Modelo SARIMA(1,0,1)x(0,1,1)</i> .....	54
• <i>Regressores Dinâmicos e o Modelo SARIMA(1,1,1) x(0,1,1)</i> .....	55
• <i>Comparação dos modelos com e sem Regressores Dinâmicos</i> .....	56
CONSTRUÇÃO DE UM MODELO DE INTERVENÇÃO .....	60
<b>DISCUSSÃO</b> .....	<b>61</b>
<b>CONCLUSÕES</b> .....	<b>71</b>
<b>REFERÊNCIAS BIBLIOGRÁFICAS</b> .....	<b>72</b>
<b>ANEXO 1 – TABELAS</b> .....	<b>76</b>
<b>ANEXO 2 – FIGURAS</b> .....	<b>82</b>

## Lista de Tabelas

- Tabela 1 – Modelos Selecionados para Casos, Pluviosidade e Umidade – Estimativas dos Parâmetros, Variância de AIC dos Modelos. Os números entre parênteses são o erro padrão da estimativa ..... 77**
- Tabela 2 – Modelos Selecionados para Temperaturas Máxima, Mínima e Média – Estimativas dos Parâmetros, Variância de AIC dos Modelos. Os números entre parênteses são o erro padrão da estimativa ..... 78**
- Tabela 3 – Modelos Selecionados com Ruído ARIMA(1,1,1) com Pluviosidade, Umidade e Temperaturas Máxima, Mínima e Média como Regressores Dinâmicos e 3 Diferentes Funções de Transferência – Estimativas dos Parâmetros, Variância de AIC dos Modelos. Os modelos entre parênteses representam a ordem das funções de transferência  $(r,s,b)$ . Os números entre parênteses no corpo da Tabela são o erro padrão das estimativas. Os modelos selecionados estão destacados em negrito. .... 79**
- Tabela 4 – Modelos Selecionados com Ruído SARIMA (1,0,1)x(0,1,1) com Pluviosidade, Umidade e Temperaturas Máxima, Mínima e Média como Regressores Dinâmicos e 3 Diferentes Funções de Transferência – Estimativas dos Parâmetros, Variância de AIC dos Modelos. Os modelos entre parênteses representam a ordem das funções de transferência  $(r,s,b)$ . Os números entre parênteses no corpo da Tabela são o erro padrão das estimativas. Os modelos selecionados estão destacados em negrito..... 80**
- Tabela 5 – Modelos Selecionados com Ruído SARIMA (1,1,1)x(0,1,1) com Pluviosidade, Umidade e Temperaturas Máxima, Mínima e Média como Regressores Dinâmicos e 3 Diferentes Funções de Transferência – Estimativas dos Parâmetros, Variância de AIC dos Modelos. Os modelos entre parênteses representam a ordem das funções de transferência  $(r,s,b)$ . Os números entre parênteses no corpo da Tabela são o erro padrão das estimativas. Os modelos selecionados estão destacados em negrito..... 81**

## **Lista de Figuras**

<b>Figura 1 – Percentual da Procedência dos Casos de Leptospirose.....</b>	<b>83</b>
<b>Figura 2 - Gráficos Temporais das Séries Calendário de Casos, Pluviosidade e Umidade.....</b>	<b>84</b>
<b>Figura 3 - Gráficos Temporais das Séries calendário de Temperaturas Máxima, Média e Mínima .....</b>	<b>85</b>
<b>Figura 4 – Série de Casos – Gráficos Temporais da Série Diferenciada e Histogramas Comparando a Série Original e as Transformações Raiz Quadrada e Logarítmica Deslocada de 1 Unidade. ....</b>	<b>86</b>
<b>Figura 5 – Série de Pluviosidade – Gráficos Temporais da Série Diferenciada e Histogramas Comparando a Série Original e as Transformações Raiz Quadrada e Logarítmica com Deslocamento de 1 Unidade. ....</b>	<b>87</b>
<b>Figura 6 – Gráficos Temporais das Séries Diferenciadas e Histogramas das Séries de Umidade e Temperaturas Máxima, Mínima e Média .....</b>	<b>88</b>
<b>Figura 7 – Gráficos da Decomposição Sazonal via LOESS das Séries Temporais de Casos, Pluviosidade, Umidade e Temperaturas Máxima, Média e Mínima. A curva superior é a série original e foi deslocada +20 unidades para casos e pluviosidade. A curva do meio são os resíduos e foram deslocadas +10 para casos e pluviosidade, -20 para Umidade e -10 para as 3 temperaturas. Curva inferior é a sazonalidade não deslocada.....</b>	<b>89</b>
<b>Figura 8 – Box-Plots Sazonais das Séries de Casos, Pluviosidade, Umidade e Temperaturas Máxima, Média e Mínima. Apesar das séries começarem em Março, a semana número 1 neste gráfico corresponde à primeira semana do ano.....</b>	<b>90</b>
<b>Figura 9 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Casos e sua Diferenciação não Sazonal.....</b>	<b>91</b>
<b>Figura 10 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Casos com Diferenciação Sazonal e com Diferenciação Sazonal e não sazonal.....</b>	<b>92</b>
<b>Figura 11 – Saída do Programa S-Plus Mostrando os Gráficos de Diagnóstico de Ajuste para a Série de Casos – Modelo ARIMA (1,1,1) – (a) Gráfico Temporal dos Resíduos Padronizados; (b) Correlograma da ACF dos Resíduos; (c) Correlograma da PACF dos Resíduos; (d) Teste de Portmanteau para Autocorrelação dos Resíduos. ....</b>	<b>93</b>
<b>Figura 12 – Saída do Programa S-Plus Mostrando os Gráficos de Diagnóstico de Ajuste para a Série de Casos – Modelo SARIMA (1,0,1)x(0,1,1) – (a) Gráfico Temporal dos Resíduos Padronizados; (b) Correlograma da ACF dos Resíduos; (c) Correlograma da PACF dos Resíduos; (d) Teste de Portmanteau para Autocorrelação dos Resíduos. ....</b>	<b>94</b>
<b>Figura 13 – Saída do programa S-Plus Mostrando os Gráficos de Diagnóstico de Ajuste para a Série de Casos – Modelo SARIMA (1,1,1)x(0,1,1) – (a) Gráfico Temporal dos Resíduos Padronizados; (b) Correlograma da ACF dos Resíduos; (c) Correlograma da PACF dos Resíduos; (d) Teste de Portmanteau para Autocorrelação dos Resíduos. ....</b>	<b>95</b>
<b>Figura 14 –Periodogramas Acumulados e IC 95% para a Série de Casos e Resíduos dos modelos ARIMA (1,1,1), SARIMA (1,0,1)x(0,1,1) e SARIMA (1,1,1)x(0,1,1).....</b>	<b>96</b>
<b>Figura 15 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Pluviosidade e sua Diferenciação não Sazonal. ....</b>	<b>97</b>
<b>Figura 16 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Pluviosidade com Diferenciação Sazonal e com Diferenciação Sazonal e não sazonal. ....</b>	<b>98</b>

Figura 17 – Periodogramas Acumulados e IC 95% para a Série de Pluviosidade e Resíduos dos modelos ARIMA (3,0,1), ARIMA (0,1,1) e SARIMA (1,0,1)x(0,1,1).....	99
Figura 18 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Umidade e sua Diferenciação não Sazonal.....	100
Figura 19 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Umidade com Diferenciação Sazonal e com Diferenciação Sazonal e não sazonal. ....	101
Figura 20 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Temperatura Média e sua Diferenciação não Sazonal. ....	102
Figura 21 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Temperatura Média com Diferenciação Sazonal e com Diferenciação Sazonal e não sazonal. ....	103
Figura 22 – Correlogramas das CCFs e IC 95% para a Série de Casos e: (a) Pluviosidade; (b) Umidade; (c) Temperatura Média; (d) Temperatura Mínima – Ambas as séries foram diferenciadas apenas.	104
Figura 23 – Correlogramas das CCFs e IC 95% para a Série de Casos e: (a) Pluviosidade – AR (3); (b) Umidade – ARMA (1,1); (c) Temperatura Média – ARIMA (2,1,3); (d) Temperatura Máxima ARIMA (2,1,3) – Ambas as séries foram modeladas por modelos ARIMA de mesma ordem – Pré-branqueamento.....	105
Figura 24 – Correlogramas das CCFs e IC 95% para as Séries Diferenciadas de (a) Pluviosidade e Umidade; (b) Pluviosidade e Temperatura Média; (c) Umidade e Temperatura Média; (d) Temperatura Mínima.....	106
Figura 25 – Diagnósticos para os Modelos com Regressores Dinâmicos – Modelo SARIMA (1,1,1)x(0,1,1) com Pluviosidade como Entrada. Correlogramas e IC 95% para: (a) ACF dos Resíduos do Modelo com Pluviosidade como Entrada; (b) ACF Modelo Pré-Branqueado de Pluviosidade - AR (3) e (c) CCF entre os Resíduos do Modelo com Pluviosidade como Entrada e os Resíduos do Modelo Pré-Branqueado de Pluviosidade - AR (3).....	107
Figura 26 – Diagnósticos para os Modelos com Regressores Dinâmicos – Modelo SARIMA (1,0,1)x(0,1,1) com Umidade como Entrada. Correlogramas e IC 95% para: (a) ACF dos Resíduos do Modelo com Umidade como Entrada; (b) ACF Modelo Pré-Branqueado de Umidade - ARMA (1,1) e (c) CCF entre os Resíduos do Modelo com Umidade como Entrada e os Resíduos do Modelo Pré-Branqueado de Umidade - ARMA (1,1); (d).....	108
Figura 27 – Diagnósticos para os Modelos com Regressores Dinâmicos – Modelo SARIMA (1,0,1)x(0,1,1) com Temperatura Média como Entrada Correlogramas e IC 95% para: (a) ACF dos Resíduos do Modelo; (b) ACF Modelo Pré-Branqueado de Temperatura Média - ARIMA (2,1,3) e (c) CCF entre os Resíduos do Modelo com Temperatura Média como Entrada e os Resíduos do Modelo Pré-Branqueado de Temperatura Média - ARIMA (2,1,3).....	109
Figura 28 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados). Acima: Modelo ARIMA (1,1,1) sem Regressores Dinâmicos. Abaixo: Modelo ARIMA (1,1,1) com Pluviosidade (2,0,0) como Regressor Dinâmico. ....	110
Figura 29 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados).	

- Acima: Modelo ARIMA (1,1,1) com Umidade (2,0,0) como Regressor Dinâmico. Abaixo: Modelo ARIMA (1,1,1) com Temperatura Máxima (0,1,1) como Regressor Dinâmico..... 111
- Figura 30 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados).
- Acima: Modelo ARIMA (1,1,1) com Temperatura Mínima (0,1,1) como Regressor Dinâmico. Abaixo: Modelo ARIMA (1,1,1) com Temperatura Média (0,1,1) como Regressor Dinâmico. .... 112
- Figura 31 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados).
- Acima: Modelo SARIMA (1,0,1)x(0,1,1) sem Regressores Dinâmicos. Abaixo: Modelo SARIMA (1,0,1)x(0,1,1) com Pluviosidade (2,0,0) como Regressor Dinâmico. .... 113
- Figura 32 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados).
- Acima: Modelo SARIMA (1,0,1)x(0,1,1) com Umidade (2,0,0) como Regressor Dinâmico. Abaixo: Modelo SARIMA (1,0,1)x(0,1,1) com Temperatura Máxima (0,1,1) como Regressor Dinâmico. .... 114
- Figura 33 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados).
- Acima: Modelo SARIMA (1,0,1)x(0,1,1) com Temperatura Mínima (0,1,1) como Regressor Dinâmico. Abaixo: Modelo SARIMA (1,0,1)x(0,1,1) com Temperatura Média (0,1,1) como Regressor Dinâmico. .... 115
- Figura 34 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados).
- Acima: Modelo SARIMA (1,1,1)x(0,1,1) sem Regressores Dinâmicos. Abaixo: Modelo SARIMA (1,1,1)x(0,1,1) com Pluviosidade (2,0,0) como Regressor Dinâmico. .... 116
- Figura 35 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados).
- Acima: Modelo SARIMA (1,1,1)x(0,1,1) com Umidade (1,0,1) como Regressor Dinâmico. Abaixo: Modelo SARIMA (1,1,1)x(0,1,1) com Temperatura Máxima (0,1,1) como Regressor Dinâmico. .... 117
- Figura 36 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados).
- Acima: Modelo SARIMA (1,1,1)x(0,1,1) com Temperatura Mínima (0,1,1) como Regressor Dinâmico. Abaixo: Modelo SARIMA (1,1,1)x(0,1,1) com Temperatura Média (0,1,1) como Regressor Dinâmico. .... 118
- Figura 37 – Gráficos da Variação das Estimativas de Pontos Futuros em Relação às Observações Reais em Percentual – Resultado de 18 Diferentes Modelos, Agrupados pelos Diferentes Modelos de Casos.. 119
- Figura 38 – Gráficos dos Comportamentos dos Erros-Padrão em 18 Modelos Diferentes. Gráfico Superior: Agrupados pelos Diferentes Modelos de Casos. Gráfico Inferior: Detalhe dos Modelos sem Regressores e com Pluviosidade. .... 120
- Figura 39 – Histogramas das Séries de Casos e Pluviosidade Transformadas pela Função Logarítmica, Deslocadas de 0.01, 0.1 e 1 Unidade. .... 121
- Figura 40 – Correlogramas das CCFs e IC 95% para: (a) Resíduos dos Casos e Pluviosidade Ajustados por AR(5); (b) Resíduos dos Casos e Umidade Ajustados por ARMA(1,1); (c) Resíduos dos Casos e Temperatura Média Ajustados por ARIMA(2,1,3); (d) Resíduos da Diferenciação da Pluviosidade e

**Umidade. As séries de casos e pluviosidades usadas nestes correlogramas são as originais, não transformadas..... 122**

## **Abstract**

Leptospirosis is an endemic disease in many parts of Brazil, including Salvador – BA. The clinical presentation is broad and, in general, serological confirmation is needed, except for severe cases. From 1996 to 1999 all the severe cases of Leptospirosis diagnosed in an infectious diseases reference hospital in Salvador – BA were identified and demographic and clinical data were collected. Thus, it was possible to set up a daily time series of Leptospirosis cases. Since daily observations of Rainfall, Humidity and maximum, minimum and mean Temperature were also available, it was possible to study the relationship between these variables and the cases of Leptospirosis.

The methods employed for analysis were the classical Box and Jenkins time series models. The meteorological variables were studied using the cross-correlation function and also considered in models with dynamic regressors, to assess the influence of these variables over the behavior of the Leptospirosis cases series.

The results show that the series of cases does not present any significant level of variation during the study period, but a seasonal pattern can be identified. The influence of the meteorological variables over the cases is significant for all of them, and with a lag of 1 or 2 weeks. It is fair to assume that this lag is the incubation period of the disease in this population. The meteorological variables present a strong correlation among them, with no lag. All the meteorological variables and especially Rainfall, proved to be useful when used in a model to forecast Leptospirosis cases.

Despite the reasonable fitness of these models, they are clearly not adequate for counting series, and new classes of models such as Dynamic Generalized Linear Models should be used for refinement of the results.

## **Resumo**

A Leptospirose é uma doença endêmica em várias partes do Brasil, e em particular em Salvador – BA. Sua apresentação clínica é de largo espectro e, em geral, é necessária a confirmação sorológica, exceto em casos graves da doença. No período de 1996 a 1999 todos os casos graves de Leptospirose diagnosticados em hospital de referência em infectologia em Salvador – BA foram identificados e coletados dados demográficos e clínicos. Assim, foi possível montar uma série temporal diária de casos de Leptospirose. Como observações diárias de Pluviosidade, Umidade e Temperaturas máxima, mínima e média também estavam disponíveis, foi possível estudar as relações entre estas variáveis e os casos de Leptospirose.

Os métodos utilizados para a análise foram os modelos de séries temporais clássicos de Box e Jenkins. As variáveis meteorológicas foram estudadas através da função de correlação cruzada e também foram consideradas em modelos com regressores dinâmicos, a fim de avaliar a influência dessas variáveis sobre a série de casos de Leptospirose.

Os resultados mostram que a série de casos não apresenta variação significativa em sua média, durante o período de estudo, mas pode-se identificar um padrão sazonal. A influência das variáveis meteorológicas sobre os casos foi significativa para todas elas, e com um atraso de 1 a 2 semanas. As variáveis meteorológicas apresentam uma forte correlação entre elas, sem atraso. Todas as variáveis meteorológicas e especialmente a Pluviosidade mostraram-se úteis quando usadas num modelo para prever casos futuros de Leptospirose.

Apesar do ajuste razoável destes modelos, eles são claramente inadequados para séries de contagem e novas classes de modelos como os Modelos Lineares Dinâmicos Generalizados devem ser utilizados com o intuito de refinar a análise.

## **Introdução**

A Leptospirose é uma doença infecciosa, causada por um dos 240 diferentes sorotipos da espiroqueta *Leptospira interrogans*. Clinicamente, se apresenta como uma doença bifásica, composta de uma fase septicêmica, que dura de 4 a 10 dias, e uma fase imune, com duração de 10 a 30 dias (Martins & Castiñeiras 1994; Tappero et al. 2000). Seu espectro clínico é amplo (Perrocheau & Perolat 1997; Murhekar et al. 1998), com início abrupto de febre alta e calafrios, mialgia intensa, notadamente em panturrilhas, geralmente acompanhada de cefaléia e hiperemia conjuntival intensa. Dor abdominal, náuseas e vômitos podem estar presentes. Estes sinais e sintomas persistem por 4 a 7 dias, e em alguns doentes, 2 a 3 dias após, há um recrudescimento (Martins & Castiñeiras 1994; Tappero et al. 2000).

Em cerca de 10% dos pacientes, após o 3º dia de doença, ocorre icterícia e a doença pode evoluir com insuficiência renal aguda e manifestações hemorrágicas. Este quadro mais grave da doença, caracteriza a chamada Síndrome de Weil (Dupont et al. 1997; Trevejo et al. 1998). Por apresentarem características clínicas mais marcantes, e por se tratarem de pacientes com doença mais grave, estes pacientes são mais facilmente diagnosticados e podem servir como marcadores da magnitude da incidência da doença em uma determinada área.

A Leptospirose é uma zoonose cosmopolita, que tem animais silvestres e domésticos como reservatórios. O homem é acometido ao entrar em contato com a urina contaminada destes animais, por penetração ativa pela pele ou passiva pela ingestão de alimentos ou água contaminados (Corwin et al. 1990). Em países desenvolvidos, a população acometida se restringe a profissionais que trabalham em contato com animais, como trabalhadores rurais ou em atividades de lazer (Anderson et al. 1978; CDC 1997). Em países em desenvolvimento, com instalações sanitárias precárias, toda a população está em risco, principalmente a urbana pela proximidade e número do principal reservatório urbano da *L. interrogans*, o rato de esgoto (*Rattus norvegicus*) (Collares-Pereira et al. 1997; Ko et al. 1999). A população mais acometida é a de adultos jovens, especialmente do sexo masculino (Kuriakose et al. 1997; Perrocheau & Perolat 1997), mas crianças podem ser acometidas também (Marotto et al. 1997).

O período de incubação da doença varia de 2 a 20 dias (média de 10), sendo em geral entre 5 e 14 dias. Fatores que influenciam a variabilidade do tempo de incubação bem como a

gravidade da doença ainda não são bem definidos (Martins & Castiñeiras 1994; Tappero et al. 2000).

Fatores meteorológicos influenciam na epidemiologia da doença, pois a espiroqueta necessita de condições ótimas de temperatura, umidade e pH do solo para sobreviver, aumentando a sua capacidade infectante. Além disso, a chuva desempenha um papel fundamental como veículo hídrico para alcançar o homem e é um fator de risco conhecido para epidemias de leptospirose urbana (Douglin et al. 1997; Murhekar et al. 1998).

No Brasil não é incomum a presença de casos de Leptospirose, especialmente em centros urbanos como Salvador – Bahia, cidade localizada no Nordeste brasileiro com cerca de 2 milhões de habitantes, e grande concentração urbana.

A temperatura média e a umidade são elevadas durante todo o ano e a precipitação pluviométrica variável e muitas vezes intensa, principalmente no final do verão. A Leptospirose é uma doença endêmica de incidência também variável com períodos epidêmicos no decorrer do ano (Ko et al. 1999), geralmente concordantes com a época de chuvas.

Um problema interessante a ser estudado é a relação que existe entre a ocorrência de manifestações meteorológicas e o aparecimento de casos de doenças, o que pode ser encontrado com certa frequência na literatura (Saez et al. 1995; Alberdi et al. 1998). Por se tratarem de séries temporais, porém, estes estudos exigem a aplicação de técnicas adequadas a estes tipos de variáveis (Diggle 1992; Chatfield 1996).

Num estudo realizado em Barbados, Everard et al. (1992) utilizaram modelos de análise de séries temporais para demonstrar a associação entre a quantidade de chuva em um determinado mês e o aparecimento de casos graves de leptospirose no mesmo mês. O fato de não ter sido encontrado um atraso entre as duas séries se deve provavelmente ao fato da unidade de agregação (mês) não permitir um refinamento do atraso entre a série de pluviosidade e de casos, o que corresponderia ao tempo de incubação da doença.

Modelos para o estudo de séries temporais foram desenvolvidos inicialmente para aplicações econométricas, e vêm cada vez mais sendo usados por outras disciplinas. Sua aplicação epidemiológica pode ser encontrada com frequência na literatura. Estudos sobre a

relação entre temperatura e mortalidade (Saez et al. 1995; Alberdi et al. 1998), confecção de diagramas de controle de doenças (VanBrackle & Williamson 1999; Williamson & G. 1999), estudos sobre vigilância epidemiológica para Sífilis (Zaidi et al. 1989), Gonorréia (Schnell et al. 1989) e Salmonelose (Watier et al. 1991), análise do comportamento de doenças virais (Helfenstein 1986), de causas externas (Martinez-Schnell & Zaidi 1989) e previsão de demanda para serviços de emergência (Milner 1988) são alguns dos diversos exemplos constantes da literatura.

Outra aplicação destes modelos é o estudo da correlação entre variáveis “externas” e uma série temporal de interesse, chamado de sistema linear aberto, onde uma série é a entrada – *input* – e a outra série é a saída – *output* (Chatfield 1996). Esta abordagem é feita através das funções de transferência (Helfenstein 1991; Helfenstein 1996) e pode ser encontrada também na literatura com aplicações em epidemiologia (Stroup et al. 1988; Fernandez-Perez et al. 1998). Além disso, análises de intervenção podem ser também utilizadas para determinar o efeito de alguma medida sobre uma determinada série temporal (Helfenstein et al. 1991).

Os modelos de séries temporais clássicos podem se divididos em duas grandes famílias: os modelos do domínio da frequência, e os modelos do domínio do tempo. O primeiro grupo de modelos são baseados no estudo da decomposição de frequências de uma série temporal, que pode ser representada por uma integral de Fourier invertida (Diggle 1992; Chatfield 1996).

Já no domínio do tempo, são utilizados os modelos conhecidos como Autorregressivos-Médias Móveis (ARMA) e suas variantes, ARIMA (Autorregressivo-Integrado-Médias Móveis) e SARIMA (Autorregressivo-Integrado-Médias Móveis-Sazonal), sendo estes últimos também conhecidos como modelos de Box e Jenkins (Box et al. 1994). São baseados na idéia de que todas as séries temporais estacionárias podem ser representadas por uma estrutura conhecida acrescida de um componente aleatório. Esta estrutura conhecida é formada por pontos cujos valores são dependentes das observações anteriores (componente autorregressivo) e também dos erros aleatórios passados (componente de médias móveis).

Com o uso destes modelos é possível não só estabelecer uma relação entre a precipitação pluviométrica e o surgimento de casos de leptospirose, como também de outras variáveis meteorológicas, como temperatura e umidade. Além disso, é possível a construção

de um modelo preditivo, que leve em conta estas variáveis, bem como um estudo de intervenção, onde é possível estabelecer se medidas externas influenciaram ou não no aparecimento de casos da doença.

## **Objetivos**

Estimar o tempo de incubação de casos graves de Leptospirose em Salvador – Bahia, através do estudo da correlação cruzada entre as séries temporais de casos graves e variáveis meteorológicas.

Modelar uma série temporal de casos graves de Leptospirose em Salvador – Bahia, para construir um modelo preditivo, com o auxílio de séries temporais meteorológicas que sejam explicativas, considerando o tempo de incubação da doença.

Usar este modelo para realizar estudos de intervenção para determinar se medidas de desratização tomadas em Salvador – BA a partir da metade da série observada, i.e. a partir de 1998, foram capazes de diminuir significativamente o número de casos graves de Leptospirose na cidade.

## **Considerações Éticas**

O objeto desta dissertação de mestrado envolve seres humanos na medida em que são computados os casos graves de Leptospirose, diagnosticados por critérios clínicos em Salvador – BA. Estes dados são coletados e digitalizados pelos profissionais do Hospital Couto Maia e da FIOCRUZ, em Salvador.

As informações usadas por estes profissionais são mais completas e englobam variáveis clínicas e laboratoriais, colhidas em entrevistas feitas com os pacientes, sob consentimento livre e esclarecido e ainda com revisão do prontuário dos mesmos, o que exige evidentemente a observância dos preceitos éticos contidos na Resolução 196/96 do Conselho Nacional de Saúde.

Na presente pesquisa, o acesso aos dados se dá apenas através do banco de dados digitalizado fornecido pelo grupo de Salvador, e claro que merece também atenção aos aspectos éticos, ainda que o pesquisador não tenha acesso aos pacientes envolvidos. O princípio da confidencialidade deve ser observado com bastante rigor, evitando a todo custo a identificação por meios diretos ou indiretos dos pacientes.

A identificação direta é impossível neste caso, pois o banco de dados em questão não possui nomes nem endereços dos pacientes ou de parentes destes. A única informação é a cidade de residência do paciente, o que não permite a identificação (nem indireta) de qualquer paciente. Apesar de outras características como sexo, idade e diagnóstico laboratorial estarem disponíveis, estas informações não serão utilizadas ao nível individual, pela natureza do estudo que é apenas descrever o comportamento temporal do aparecimento dos casos.

Em termos de custo-benefício, a pesquisa aproveita um banco de dados que já existe, portanto trata-se de uma base secundária, o que não representa custo adicional algum aos pacientes. Em termos de benefícios que a pesquisa pode trazer, são o melhor entendimento da influência de fatores meteorológicos sobre o aparecimento de casos da doença e a capacidade de previsão da magnitude que estes casos podem alcançar, dadas as condições meteorológicas prévias.

Este conhecimento poderá permitir prever, no futuro, o aparecimento de uma epidemia da doença, por exemplo, e alertar as unidades de saúde locais sobre este fato com antecedência.

Isto, porém não faz parte dos objetivos desta pesquisa, e ocorreria apenas se autoridades sanitárias locais se mostrassem interessadas na utilização de um modelo preditivo desta natureza.

A estimação do tempo de incubação da doença, que nesse caso seria medido em função do atraso entre a manifestação de uma variável externa (“entrada” ou “input” – como a pluviosidade por exemplo) e seu efeito sobre o sistema (aparecimento dos casos de Leptospirose) é feito através da utilização das correlações cruzadas entre as séries pré-branqueadas, como explicitado na seção de Métodos. Além disso, essa informação é utilizada para a construção do modelo com co-variáveis e função de transferência.

## **Materiais e Métodos**

### **Casos Graves de Leptospirose**

Os casos de Leptospirose foram identificados pelos profissionais médicos do Hospital Couto Maia, que é um hospital de referência em infectologia, e responsável por mais de 90% das notificações de Leptospirose na Região Metropolitana de Salvador.

Os critérios clínicos de inclusão de casos graves de leptospirose no estudo consistiam de pacientes com doença febril de duração menor de 21 dias, acompanhado de 1. sufusão conjuntival; 2. icterícia com aminotransferases <1000U/l; ou 3. icterícia com creatinina maior que 1,5 mg/dl ou Uréia > 75 mg/dl.

Sempre que possível foi realizada a confirmação sorológica dos casos, usando-se um dos critérios:

- Aumento de 4 vezes no título da Micoaglutinação (MAT) – quando o pareamento foi possível.
- Título da MAT maior que 1:800 em única aferição.
- Cultura de urina ou sangue positiva.

Além disso, um título único de MAT > 1:100 foi considerado como caso suspeito.

Utilizando-se apenas os critérios clínicos, foram incluídos no estudo, a princípio, 1294 pacientes identificados entre o período de 21 de março de 1996 e 14 de janeiro de 2000. O ideal seria descartar todos os casos sem confirmação sorológica, para se ter a certeza que estaremos avaliando apenas casos de leptospirose e não outras doenças com quadro clínico semelhante. Infelizmente porém, muitas vezes é difícil conseguir a confirmação, principalmente quando se tratam de amostras pareadas de soro. Assim, dos 1294 casos, apenas 50% foram confirmados (504) ou suspeitos (143).

Como a exclusão destes casos não confirmados representaria uma perda significativa (50%) do número de casos disponíveis para o trabalho e pelo fato de, num trabalho anterior realizado apenas com os casos até novembro de 1996 (Ko et al. 1999), o valor preditivo

positivo do diagnóstico clínico dos casos graves de leptospirose foi de 91%, optou-se por utilizar apenas os critérios clínicos para a inclusão inicial no estudo.

A data utilizada para representar o início dos sintomas nos casos foi obtida pela subtração do dia do diagnóstico pelo número de dias do aparecimento dos sintomas, informação que foi obtida diretamente com os pacientes. Sendo assim, foram excluídos os casos que não tinham a informação de dias com sintomas, o que ocorreu em 15 casos.

Outro problema que foi encontrado diz respeito ao local de residência dos casos. Apesar da maioria dos casos (88%) serem provenientes de Salvador, muitos deles tinham local de residência fora da região metropolitana de Salvador, que é composta pelas cidades de Camaçari, Candeias, Catú, Dias D'Ávila, Lauro de Freitas, Nova Candeias, Simões Filho e Salvador. Portanto, 141 casos foram também excluídos por não serem residentes em uma destas cidades. O motivo desta restrição geográfica é que, como será comentado adiante, as variáveis meteorológicas foram obtidas a partir de uma única estação meteorológica, em Salvador. A distribuição da procedência dos casos pode ser observada na Figura 1.

Por fim, como as aferições meteorológicas incluídas no estudo só estavam disponíveis até 31 de dezembro de 1999, mais 5 casos (de janeiro de 2000) foram excluídos.

Assim, a série temporal de casos graves de Leptospirose é composta de 1133 casos, que tiveram seus primeiros sintomas percebidos (já feita a subtração da data de diagnóstico pelos dias de início de sintomas) no período entre 15 de março de 1996 e 31 de dezembro de 1999.

Para a modelagem da série temporal de casos, estes foram agrupados por semanas, a partir da soma total dos casos ocorridos em cada uma das semanas. Os 4 últimos pontos da série já agrupada foram separados para permitir a avaliação do poder preditivo dos modelos *a posteriori*.

## Variáveis Meteorológicas

Todas as variáveis meteorológicas foram colhidas no único centro meteorológico de Salvador, localizado no bairro de Ondina. O período de coleta é de 1 de janeiro de 1996 a 31 de dezembro de 1999, porém somente o período de 15 de março de 1996 a 31 de dezembro de 1999 foi utilizado para o devido pareamento temporal com a série de casos graves de Leptospirose. Posteriormente, foram agrupadas convenientemente por semana, e também separados os 4 últimos pontos das séries.

A mensuração da Pluviosidade é feita diariamente pela leitura da quantidade acumulada de água da chuva, em milímetros, no final de cada dia. É a mensuração mais fácil de ser feita e sujeita a menos erros. A pluviosidade semanal é resultado da soma das pluviosidades de cada um dos dias da semana.

A Umidade é aferida em vários períodos distintos do dia e sua média é calculada para produzir uma medida do dia. No presente trabalho, foi feita uma média dos 7 dias que compõem uma semana, para obter uma média semanal.

As Temperaturas Máxima, Mínima e Média são obtidas também a partir de várias mensurações diárias após o que são computadas as marcas máxima, mínima e uma média entre todas as medidas é calculada. Para a obtenção de valores semanais, foram calculadas as médias entre as observações diárias dentro de cada semana para as 3 temperaturas.

## Modelos de Séries Temporais

As informações contidas nesta seção podem ser encontradas com detalhes em várias referências sobre séries temporais (Diggle 1992; Chatfield 1996; MathSoft Data Analysis Products Division 1999; SAS Institute Inc 1999; Venables & Ripley 1999).

- **Ruído Branco**

Inicialmente, vamos definir um processo estritamente aleatório (também conhecido como "ruído branco"), que consiste numa seqüência de variáveis aleatórias ( $Z_t$ ), mutuamente independentes e identicamente distribuídas (iid). Deste modo, este processo possui média e variância constantes; além disso,

$$\gamma(k) = \text{cov}(Z_t, Z_{t+k}) = 0, \text{ para } k = \pm 1, \pm 2, \dots$$

E o processo é então estacionário em primeira e segunda ordens. Além disso:

$$\rho(k) = \begin{cases} 1 & k = 0 \\ 0 & k = \pm 1, \pm 2, \dots \end{cases}$$

onde  $\rho(k)$  é a função de autocorrelação (ACF), que é definida como a correlação entre uma série  $X_t$  e ela mesma, atrasada de  $k$  passos ou *lags*. Assim, como será melhor descrito adiante:

$$\rho(k) = \text{corr}(X_t, X_{t+k})$$

- **Operador “Back Shift”**

Para facilitar a notação dos modelos, vamos definir um operador especial, que permitirá escrever equações abreviadamente e melhor representar alguns cálculos. Trata-se do operador “back shift” ( $B$ ), definido da seguinte maneira: Seja  $X_t$  uma série temporal, então,

$$B^j X_t = X_{t-j},$$

onde  $j$  é o atraso (*lag*) da série  $X_t$ ; e o índice  $j$  não representa uma potência, mas apenas a ordem do atraso (i.e. quantos passos a série foi atrasada).

- **Passeio Aleatório**

É um processo formado por uma série temporal atrasada 1 *lag* e um ruído branco, assim:

$$X_t = X_{t-1} + Z_t, \text{ que pode ser reescrito usando-se o operador } B:$$

$$X_t = BX_t + Z_t$$

- **Médias Móveis**

O processo de Médias móveis de grau  $q$  –  $MA_{(q)}$  – é aquele onde os valores atuais de uma série temporal  $X_t$  depende dos erros aleatórios, representados por ruídos brancos passados, em  $t-1, t-2 \dots$  Assim,

$$X_t = Z_t - \beta_1 Z_{t-1} - \beta_2 Z_{t-2} - \dots - \beta_q Z_{t-q}$$

Este processo possui algumas características especiais. A sua função de autocorrelação (ACF), que consiste nas sucessivas correlações entre uma série temporal e ela mesma atrasada  $k$  lags, e será melhor definida mais adiante, tem um ponto de corte no lag  $q$ . Isto é, a partir do lag  $q$ , a ACF de um  $MA_{(q)}$  vale zero. Assim:

$$\rho(k) = \begin{cases} 1 & k = 0 \\ \frac{\sum_{i=0}^{q-k} \beta_i \beta_{i+k}}{\sum_{i=0}^q \beta_i^2} & k = 1, \dots, q \\ 0 & k > q \end{cases}$$

Além disso, sua ACF é uma função par, i.e.  $\rho(-k) = \rho(k)$

Usando o operador  $B$  é possível reescrever o  $MA_{(q)}$ :

$$X_t = \theta(B)Z_t$$

onde  $\theta(\cdot)$  é um polinômio de ordem  $q$  com  $\theta(0) = 1$  e:

$$\theta(B) = 1 - \sum_{i=1}^q \beta_i B^i$$

Uma condição importante que deve ser respeitada nos processos de séries temporais é a garantia da inversibilidade do processo. Isto significa que é possível identificar um e somente um processo através das suas características de primeira e segunda ordens. Este critério será alcançado se e somente se todas as raízes do polinômio  $\theta(B) = 0$  caírem fora do círculo unitário, i.e se todas as raízes reais ou complexas forem em módulo maiores que 1.

- **Autorregressão**

Um processo de autorregressão de ordem  $p$  –  $AR_{(p)}$  - é aquele onde os valores atuais dependem dos valores passados da própria série, junto com uma perturbação aleatória:

$$X_t = \alpha_1 X_{t-1} + \Lambda + \alpha_p X_{t-p} + Z_t$$

onde  $Z_t$  é um processo puramente aleatório de média zero e variância  $\sigma_Z^2$ .

Neste caso, estaremos diante de uma regressão linear múltipla onde as variáveis explicativas são valores passados de  $X_t$  (é a série atrasada de  $p$  lags) e o resíduo ( $Z_t$ ) é um processo puramente aleatório, e portanto estacionário (inclusive em segunda ordem).

Além disso, uma média  $\mu$  qualquer pode ser convenientemente adicionada a este processo.

Com a utilização do operador  $B$ , pode-se também reescrever o processo de autorregressão da seguinte forma:

$$\phi(B)X_t = Z_t, \text{ ou } X_t = \frac{Z_t}{\phi(B)}$$

onde  $\phi(\cdot)$  é um polinômio de ordem  $p$  com  $\phi(0) = 1$  e:

$$\phi(B) = 1 - \sum_{i=1}^p \alpha_i B^i$$

Uma característica importante deste processo é a ambigüidade que ele guarda com o processo de  $MA_{(q)}$ . Um processo finito  $AR_{(p)}$  pode ser escrito como um processo infinito  $MA_{(q)}$ . Para exemplificar este fato, vamos usar um  $AR_{(1)}$  por uma questão de simplicidade. Assim, a  $AR_{(1)}$

$$X_t = \alpha X_{t-1} + Z_t \Rightarrow Z_t = (1 - \alpha B)^{-1} X_t$$

pode ser reescrita, por substituições recursivas como:

$$X_t = Z_t + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \dots$$

Para um caso geral o processo  $AR_{(p)}$  pode ser reescrito como:

$$X_t = \left( \sum_{j=0}^{\infty} \alpha_j B^j \right) Z_t$$

que é a sua representação como um processo linear geral.

O processo de autorregressão precisa sofrer uma restrição para garantir a sua estacionariedade em segunda ordem. A condição é bem parecida com a condição de inversibilidade das médias móveis. Para um  $AR_{(p)}$  ser estacionário é necessário que todas as raízes do polinômio  $\phi(B) = 0$  caiam fora do círculo unitário.

Todas estas ambigüidades entre o processo autorregressivo e de médias móveis serão muito úteis quando forem discutidos os processos mistos.

No modelo autorregressivo é bem mais interessante a avaliação da autocorrelação parcial (PACF), e não a total, devido à sua natureza recursiva. A autocorrelação parcial seria a contribuição da correlação em um determinado *lag* dada a presença dos coeficientes dos demais *lags*. Assim, os  $\rho_{(k)}$  são escritos em função (ou na presença) dos demais  $\rho_{(k-p)}$ , para uma  $AR_{(p)}$ . Como no processo de médias móveis, caracteristicamente a PACF de um  $AR_{(p)}$  também se iguala a zero quando  $k > p$

$$\rho(k) = \begin{cases} \sum_{j=1}^p \alpha_j \rho_{k-j} & k = 1, \dots, p \\ 0 & k > p \end{cases}$$

A Expressão  $\sum_{j=1}^p \alpha_j \rho_{k-j}$  é conhecida como as equações de Yule-Walker.

- **Modelo ARMA<sub>(p,q)</sub> (Autoregressive-Moving Average)**

Agora, tirando proveito das ambigüidades dos processos de autorregressão e médias móveis, pode-se obter um modelo misto:

$$X_t = \sum_{j=1}^p \alpha_j X_{t-j} + Z_t + \sum_{i=1}^q \beta_i Z_{t-i}, \text{ ou}$$

$$\phi(B)X_t = \theta(B)Z_t$$

Novamente as condições de estacionariedade e de inversibilidade do processo são garantidas se as raízes dos polinômios  $\theta(B) = 0$  e  $\phi(B) = 0$  caírem fora do círculo unitário.

As propriedades de segunda ordem dos modelos ARMA compartilham aquelas citadas para os seus componentes (autorregressivo e médias móveis) e em se tratando de um modelo misto, características de ambos os modelos serão encontradas. Portanto, como no processo de AR, uma média  $\mu$  qualquer pode ser acrescentada ao processo ARMA.

- **Modelo ARIMA<sub>(p,d,q)</sub> (Autoregressive-Integrated Moving Average)**

A maioria das séries temporais do mundo real no entanto não apresentam estacionariedade em primeira ordem, ou seja, elas apresentam alguma tendência, sua média aumenta ou diminui em algum grau com o passar do tempo. Para dar conta deste fato, pode-se incorporar nos modelos ARMA um termo chamado de “integrado”, gerando os modelos ARIMA.

Na verdade, trata-se de uma transformação aplicada à série para que ela estabilize a sua média. A série é subtraída dela mesma atrasada  $k$  lags, até que ela estabilize a sua média. Na prática, como se verá mais adiante, um *lag* de 1 geralmente faz o trabalho.

Um exemplo bastante interessante de como isto funciona foi mencionado anteriormente quando foi definido o passeio aleatório, que nada mais é do que um modelo ARIMA apenas com o componente de diferença, com  $k = 1$ . Relembrando:

$$X_t = BX_t + Z_t \Leftrightarrow Z_t = (1 - B)X_t$$

Observe que a série resultante neste casos é um ruído branco, estacionário em primeira (e segunda) ordem. Usando-se então o operador “backshift” para simplificar a notação, podemos escrever um modelo  $ARIMA_{(p,d,q)}$ , na forma:

$$\phi(B)(1 - B)^d X_t = \theta(B)Z_t$$

Como se trata apenas de um modelo  $ARMA_{(p,q)}$  que sofreu uma substituição do componente  $\phi(B)X_t$  pelo componente  $\phi(B)(1 - B)^d X_t$  que é uma transformação daquele, o modelo  $ARIMA_{(p,d,q)}$  goza das mesmas propriedades do modelo  $ARMA_{(p,q)}$ .

- **Modelo  $SARIMA_{(p,d,q) \times (P,D,Q)}$  (Seasonal Autoregressive-Integrated Moving Average)**

Muitas séries temporais também apresentam um componente sazonal importante e pode ser preciso modelar este componente também para se obter um modelo mais fidedigno. Para tal foram desenvolvidos os modelos  $ARIMA$  sazonais, conhecidos como  $SARIMA_{(p,d,q) \times (P,D,Q)_s}$  sendo  $s$  o período correspondente à sazonalidade.

A inclusão do componente sazonal de ao modelo se dá de maneira multiplicativa, e a representação de um modelo  $SARIMA_{(p,d,q) \times (P,D,Q)_s}$  pode ser escrita, de forma compacta:

$$\Phi(B^s)\phi(B)(1 - B)^d (1 - B^s)^D X_t = \Theta(B^s)\theta(B)Z_t, \text{ onde}$$

$\Phi(B^s)$  representam os coeficientes sazonais da autorregressão,  $\Theta(B^s)$  representam os coeficientes sazonais das médias móveis, e  $(1 - B^s)^D$  é o operador diferença de ordem  $D$ , para a diferenciação sazonal da série.

É importante notar que as séries agora podem ser diferenciadas tanto no seu componente não sazonal, quanto no seu componente sazonal (ou em ambos).

- **Modelos com Co-variáveis Regressionais**

A utilização de co-variáveis regressionais num modelo de séries temporais tem dois objetivos básicos. O primeiro é estudar a relação entre a série modelada e as co-variáveis, que podem ser uma outra série temporal atrasada no tempo, ou ainda uma variável indicadora numa análise de intervenção. O segundo objetivo seria melhorar a capacidade de previsão do modelo quando estas co-variáveis (as que têm capacidade explicativa sobre a série modelada) estão presentes, o que deve ser diagnosticado segundo as técnicas de correlação cruzada entre as séries, como comentado adiante. De maneira geral, o processo pode ser expresso da seguinte forma:

$$Y_t = z_t' \boldsymbol{\beta} + X_t,$$

onde  $Y_t$  é um processo  $ARIMA_{(p,d,q)}$  com co-variáveis conhecidas,  $z_t'$  é um vetor com  $m$  elementos conhecidos,  $\boldsymbol{\beta}$  é um vetor de parâmetros a serem estimados e  $X_t$  é um processo  $ARIMA_{(p,d,q)}$ . Claro que modelos sazonais também podem ser aplicados aqui.

- **Modelos com Regressores Dinâmicos**

As relações entre duas séries temporais discretas,  $Y_t$  e  $X_t$ , onde a primeira série é a saída de um sistema dinâmico e a segunda a entrada, podem ser representadas por equações diferenciais lineares gerais, que escritas com a ajuda do operador  $B$ , podem ser expressas da seguinte forma (Box et al. 1994):

$$(1 - \delta_1 B - \Lambda - \delta_r B^r) Y_t = (\omega_0 - \omega_1 B - \Lambda - \omega_s B^s) X_{t-b}, \text{ ou ainda em uma forma reduzida}$$

$$Y_t = \frac{\omega(B)}{\delta(B)} X_{t-b}$$

Este tipo de processo é chamado de modelo de função de transferência de ordem  $(r,s,b)$ , onde  $r$  é a ordem do numerador,  $s$  do denominador e  $b$  o atraso da entrada em relação ao efeito na série de saída.

Obviamente este modelo deve também levar em conta o processo que está por trás da série de saída, referido como ruído, mas que neste caso será um processo  $ARIMA/SARIMA$ , que será somado à função de transferência adotada, ficando o modelo geral do tipo:

$Y_t = \frac{\omega(B)}{\delta(B)} X_{t-b} + \frac{\theta(B)}{\phi(B)(1-B)^d} Z_t$ , que freqüentemente é escrito em uma forma mais

reduzida, fazendo  $\Omega(B) = \omega(B)B^b$  e  $\varphi(B) = \phi(B)(1-B)^d$ :

$$Y_t = \frac{\Omega(B)}{\delta(B)} X_t + \frac{\theta(B)}{\varphi(B)} Z_t$$

- **Modelos de Análise de Intervenção**

De maneira parecida com o conceito das funções de transferência, pode-se fazer também análises do tipo intervenção nas séries temporais, que podem ser basicamente de dois tipos: “degrau”, onde o efeito não é observado até o ponto  $y_t$  e passa a ser observado a partir de então, ou pontual, onde o efeito é observado em determinados pontos na série. De maneira geral, um modelo desse tipo pode ser representado, segundo Box et al (1994) da seguinte forma:

$$Y_t = \frac{\omega(B)}{\delta(B)} \xi_{t-b} + N_t,$$

onde  $N_t$  representa um processo ARIMA ou SARIMA e  $\xi_{t-b}$  é um componente determinístico do modelo, representado pela intervenção em questão, que no nosso caso será representado por uma função indicadora para processos do tipo “degrau”, no tempo  $T$ :

$$S_t^{(T)} = \begin{cases} 0 & t < T \\ 1 & t \geq T \end{cases}$$

## Descrição das Séries

Todos os procedimentos realizados neste trabalho foram baseados nas rotinas dos programas S-Plus® 2000 (MathSoft Data Analysis Products Division 1999; Venables & Ripley 1999) e SAS® para Windows® (SAS Institute Inc 1999), sendo que este último foi utilizado para o ajuste de modelos com regressores dinâmicos e de intervenção. Além disso, várias implementações gráficas foram feitas com o programa R 1.2.3. (Venables et al. 2001) para Windows® e Excel®. Detalhes sobre estes programas podem ser encontrados nas referências citadas.

A descrição das séries foi feita usando-se os métodos usuais, iniciando pela visualização em gráficos temporais, para a verificação de tendências, sazonalidades, ciclicidades e variabilidade. Foram usados ainda os Box-Plots sazonais que consistem em juntar num mesmo ponto todos os pontos pertencentes àquele determinado período no tempo dentro do ciclo das séries; como neste estudo os pontos são semanas, os Box-Plots possuem 52 pontos, representando todas as semanas de um ciclo de um ano.

A visualização dos correlogramas – gráficos que correspondem à Função de Autocorrelação e a Função de Autocorrelação Parcial – também são utilizados para a descrição. Porém estes gráficos têm um papel bem mais importante, como será comentado adiante.

## Métodos de Identificação de Séries Temporais

- **Séries Simples**

Primeiramente, será definido o coeficiente de correlação ( $r$ ) entre duas variáveis quaisquer, que será o ponto de partida para a definição de outras medidas usadas. No caso específico deste trabalho, as variáveis consistem na verdade em séries temporais, mas a abordagem é a mesma. O coeficiente de correlação é dado pela equação:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2]}} , \text{ onde}$$

$x_i$  = observações da série X

$\bar{x}$  = média da série X

$y_i$  = observações da série Y

$\bar{y}$  = média da série Y

O coeficiente de correlação está contido no intervalo  $[-1,1]$  e a correlação será tão maior quanto maior for o valor do módulo de  $r$ ; um valor negativo indica correlação negativa e vice-versa. Observando-se a equação acima, pode-se concluir que o coeficiente de

correlação é na verdade a razão entre a covariância de duas séries (variáveis) e o produto de seus desvios-padrão.

No caso das séries temporais, o interesse inicial ao estudar cada série individualmente é identificar se ela é serialmente (temporalmente) autocorrelacionada. Isto significa que existe uma grande correlação entre o valor do evento no tempo  $t$  e valores passados do evento (nos pontos  $t-1, t-2, \dots$ ). O que acontece é que a variável  $Y$  na equação acima será substituída pela própria variável  $X$ , atrasada  $k$  passos (*lags*). Fazendo isso, será possível estudar a correlação entre uma série  $X_t$  com uma série  $X_{t-1}, X_{t-2}$ , etc. Geralmente este estudo é feito através das funções de autocovariância (ACVF) e autocorrelação (ACF).

Assim, definamos as funções de autocovariância (ACVF) e autocorrelação (ACF) para os  $k$  lags:

$$\gamma_k = \text{cov}(X_\tau, X_{\tau+k})$$

$$\rho_k = \text{corr}(X_\tau, X_{\tau+k})$$

Os estimadores destas funções são:

$$\text{ACVF} = c_k = \frac{1}{n} \sum_{t=\max(1,-k)}^{\min(n-k,n)} (x_t - \bar{x})(x_{t+k} - \bar{x})$$

$$\text{ACF} = r_k = \frac{c_k}{c_0}$$

onde  $c_0$  é a variância de  $x_t$ ,  $c_k$  é o estimador de  $\gamma_k$  e  $r_k$  é o estimador de  $\rho_k$ .

O intervalo de confiança deste estimador para testar se  $\rho_k = 0$  é  $\pm 2/\sqrt{n}$ , pois para um  $n$  suficientemente grande cada  $r_k$  tem distribuição Normal com média zero e variância  $1/n$ .

Além de detectar a estrutura de autocorrelação da série, a ACF também é útil para determinar a ordem de um processo de Médias Móveis, pois a ACF de uma  $\text{MA}_{(q)}$  sofre um

*cut off* no *lag*  $q$ . Assim, a ordem deste processo será a correspondente ao último coeficiente significativamente diferente de zero.

Além disso, como mencionado anteriormente, é necessária a estimativa da função de autocorrelação parcial (PACF), que é apenas a aplicação das equações de Yule-Walker na amostra. Assim:

$$r_k = \sum_{j=1}^p \hat{\alpha}_j r_{k-j}, \quad k = 1, 2, \dots, p$$

Novamente o intervalo de confiança para testar se  $\rho_k = 0$  é  $\pm 2/\sqrt{n}$ .

A importância desta função é estimar a ordem de uma  $AR(p)$ , que sofre um *cut off* no *lag*  $p$ . Assim, a ordem deste processo será a correspondente ao último coeficiente parcial significativamente diferente de zero.

Várias características destas funções são discutidas em diversas referências (Hamilton & Watts 1978; Pukkila et al. 1990; Box et al. 1994). Antes da disponibilidade computacional de hoje em dia, estas características eram muito importantes não só para rodar-se o menor número possível de modelos (o que costumava levar muito tempo), como também calcular valores iniciais para que o algoritmo convergisse mais depressa. Hoje um modelo desses rodado num software como S-Plus® demora em média menos de 1 segundo, e portanto tornou-se mais importante a tentativa e erro de diferentes modelos do que tentar acertar o modelo descrito por essas funções, embora elas de fato sirvam para orientar bastante o modelo a ser usado.

Para citar duas características clássicas, o processo  $AR(1)$ , por exemplo apresenta uma queda exponencial na ACF acompanhada de uma correlação significativa na PACF apenas no *lag* 1. Já no processo  $MA(1)$  ocorre o contrário: a PACF apresenta um decaimento exponencial, enquanto a ACF apresenta correlações positivas nos *lags* 0 (claro) e 1.

- **Séries Múltiplas**

Quando o objetivo é estudar a influência de uma série temporal sobre uma outra série temporal de interesse, é também necessário o diagnóstico da relação entre estas séries. Nem

sempre uma relação (quando existe) entre as séries se dá no mesmo ponto do tempo: muitas vezes um fenômeno (uma série) antecede outro (outra série) no tempo. Por esse motivo, a correlação entre as séries não deve ser feita apenas pelo coeficiente de correlação usual (Pearson), e deve ser utilizada a função de correlação cruzada (CCF), que parte do mesmo raciocínio da ACF, mas com a diferença de que agora o objetivo é estudar a correlação entre uma série  $X_t$  e uma outra série deslocada no tempo  $Y_{t+1}, Y_{t+2} \dots$ . Assim, por analogia:

$$c_{xy}(k) = \frac{1}{n} \sum_{t=\max(1,-k)}^{\min(n-k,n)} (x_t - \bar{x})(y_{t+k} - \bar{y})$$

$$r_{xy}(k) = \frac{c_{xy}(k)}{\sqrt{c_{xx}(0)c_{yy}(0)}}$$

onde  $c_{xx}(0)$  e  $c_{yy}(0)$  são as variâncias de  $x_t$  e  $y_t$ , respectivamente.

Porém, no caso da CCF, os dados brutos não devem ser utilizados diretamente, porque as estruturas presentes em ambas as séries como tendências, sazonalidades e correlação serial podem levar ao diagnóstico de correlações cruzadas espúrias entre as séries estudadas. Um efeito comum é a presença de correlações cruzadas significativas em vários *lags* diferentes, o que dificulta a interpretação do correlograma. A correlação cruzada estaria se dando neste caso entre as estruturas seriais das duas séries, ocasionando uma falsa correlação.

Então, como será comentado adiante, as séries devem ser tratadas antes da inspeção da CCF, de modo que as estruturas presentes não atrapalhem os diagnósticos.

- **Correlogramas**

A representação gráfica da ACF juntamente com o seu intervalo de confiança é o correlograma. Ele consiste em um gráfico tipo densidade, onde cada ponto do eixo  $x$  corresponde a um *lag* e o seu valor no eixo  $y$  é o coeficiente de correlação naquele *lag*. A mesma apresentação é feita para a PACF e ainda para a CCF. Desse modo, as características comentadas anteriormente sobre o comportamento dessas funções para os diferentes processos podem ser visualizadas muito mais facilmente

A expressão do atraso no correlograma pode ser feita pelo número efetivo de *lags*, atrasados, onde por exemplo neste trabalho cada unidade corresponde a uma semana de atraso. Esse tipo de apresentação é interessante para a visualização do atraso quando é necessário identificar exatamente qual o *lag* que é significativo, como no caso das correlações cruzadas. Uma outra apresentação é a expressão do atraso em relação ao período da série. Neste caso, o *lag* 1 na verdade corresponde a um atraso de 1 período (que nesse trabalho é um ano ou 52 semanas). Cada ponto do correlograma neste caso vale 1/52 (i.e. a frequência da série). Essa apresentação foi utilizada para os correlogramas para a identificação das séries.

Os correlogramas são elementos muito importantes não só no diagnóstico dos modelos de séries temporais a serem empregados como também consistem importante ferramenta de diagnóstico de ajuste dos modelos escolhidos, como será visto adiante.

## **Estimação dos Parâmetros dos Modelos de Séries Temporais**

Uma vez estudados os correlogramas, o próximo passo é iniciar a modelagem das séries individualmente. Para isso, deve-se estimar os parâmetros do modelo que foi sugerido pelos métodos diagnósticos. Geralmente, mais de um modelo é ajustado e depois são comparados através dos métodos abordados na próxima seção.

A estimação é feita por maximização numérica da função de verossimilhança ( $L$ ). No caso do programa S-Plus® (MathSoft Data Analysis Products Division 1999), a maximização é feita na verdade em relação à aproximação da verossimilhança condicional nas primeiras  $m$  observações, onde

$$m = p + d + Saz(P + D),$$

sendo  $p$  a ordem do componente autorregressivo não sazonal,  $d$  o grau de diferenciação não sazonal,  $Saz$  o período da sazonalidade,  $P$  a ordem do componente autorregressivo sazonal e  $D$  o grau de diferenciação sazonal.

Assim, por exemplo para um modelo SARIMA (1,1,2)x(0,1,1) neste trabalho, onde as séries são semanais e o período é 52,  $m = 54$ .

Já no programa SAS® (SAS Institute Inc 1999) aparentemente este problema não acontece, e a estimação é feita a partir de um algoritmo de mínimos quadrados gerais.

De maneira geral, o processo visa minimizar a quantidade:

$$-2 \ln L(x_{p+1}, K, x_n | x_1, K, x_p)$$

que envolve  $p + q + 1$  parâmetros, correspondendo aos  $\alpha_j = (\alpha_1, K, \alpha_p)$  coeficientes da

autorregressão, mais os  $\beta_i = (\beta_1, K, \beta_q)$  coeficientes das médias móveis e ainda

$\sigma^2 = \text{var}(Z_t)$ . Dessa forma, valores de  $\hat{\alpha}_j$ ,  $\hat{\beta}_i$  e  $\hat{\sigma}^2$  que minimizem

$$-2 \ln L(x_{p+1}, K, x_n | x_1, K, x_p)$$

serão os estimadores dos parâmetros.

Uma outra questão que deve ser colocada é quanto à média da série a ser modelada. No S-Plus® todos os modelos são ajustados assumindo média zero para as séries, e sempre que o modelo tem média significativamente diferente de zero, ela deve ser retirada da série antes de rodar o programa, e acrescentá-la depois. No caso de séries diferenciadas, esse procedimento não é empregado geralmente, pois a média do processo diferenciado é em geral muito perto de zero. Já o SAS® permite incluir no modelo um intercepto, correspondente à média da série. Assim, quando esse programa foi usado, um intercepto foi testado e mantido no modelo sempre que parecesse significativamente diferente de zero (comparando o valor com o seu erro-padrão).

## Métodos para Avaliação dos Modelos de Séries Temporais

Após a estimação dos parâmetros dos diversos modelos, parte-se para a avaliação a fim de saber qual deles é o que melhor se ajusta aos dados em questão. Vários métodos são usados para isto, tanto estatísticas de ajuste dos modelos quanto métodos gráficos.

Um dos preceitos básicos da modelagem é o da parcimônia (quanto menos parâmetros, melhor o modelo). Baseado nisso, os chamados Critérios de Informação de Akaike (AIC), que penalizam a quantidade  $-2 \ln L(x_{p+1}, K, x_n | x_1, K, x_p)$ , somando a esta duas vezes o número de parâmetros do modelo são usados para avaliar a adequação. Portanto, de maneira geral,

quando modelos são comparados, o que possuir o menor valor para os AIC será o de melhor ajuste.

Aqui cabe uma importante observação em relação aos modelos ajustados pelo S-Plus®: como a máxima verossimilhança nesse caso é condicional a  $m$  parâmetros (ver acima), os AIC dos modelos de ordens diferentes não podem ser comparados diretamente. O programa permite, porém que se condicione arbitrariamente a estimação a  $k$  parâmetros, desde que  $k \geq p$ . Assim, após o ajuste de  $i$  modelos a serem comparados, deve-se proceder um novo ajuste, utilizando-se para todos os modelos uma opção que permite fixar  $k = \max(m_i)$ .

Outra estimativa que também foi usada neste trabalho é a variância total do modelo, que deve ser a menor possível e é apresentada juntamente com os AIC.

Os métodos gráficos são baseados na avaliação dos resíduos dos modelos. O objetivo final é obter resíduos totalmente livres de estrutura. Isto significa que eles devem ser compatíveis com um ruído branco, com média zero e variância constante, e evidentemente livres de autocorrelação.

A primeira observação deve ser a distribuição temporal dos resíduos, para a primeira verificação de estacionariedade em primeira e segunda ordens. No caso do S-Plus® 2000 os resíduos são padronizados antes de serem apresentados de forma gráfica. Também são apresentados os correlogramas, tanto da ACF quanto da PACF, para procurar por estruturas ainda presentes nos resíduos.

Além disso, a expressão gráfica de algumas estatísticas também podem ajudar na identificação de um ruído branco. O S-Plus® calcula uma estatística chamada teste de *portmanteau* (Box & Pierce 1970), que consiste em testar se os  $r_k$  são iguais a zero (o que corresponde a um ruído branco) ou não, baseados na demonstração de que, para um  $n$  suficientemente grande e  $m \ll n$ ,

$$Q_m = n \sum_{k=1}^m r_k^2 \sim \chi_{m-u}^2$$

onde  $n$  é o tamanho da série,  $u$  é o número de parâmetros do modelo e  $m$  é escolhido de maneira arbitrária, porém geralmente corresponde a  $10 + p + q$ . O problema é que a saída do

programa indica que na verdade é calculado uma variante desse teste, conhecido como teste de Ljung-Box (Ljung & Box 1978), mas na verdade não é esse o resultado observado (Venables & Ripley 1999).

Um gráfico mostra então os p-valores associados a esta distribuição  $\chi^2_{m-u}$ . Neste caso, evidentemente o objetivo é aceitar a hipótese nula de que  $\rho_k = 0$  e portanto esperam-se valores maiores que 0,05 para tal.

Um último método gráfico utilizado, o periodograma acumulado, pertence à família dos modelos do domínio da frequência, e sua explicação minuciosa exigiria uma grande abordagem desses modelos. Sendo assim, uma explicação simplificada é que este teste corresponde à visualização de uma função acumulada do estimador do espectro de uma série, para cada uma das frequências de Fourier ( $\omega$ ), que é chamado de periodograma. Esta função, quando a série se trata de um ruído branco, deve se apresentar como uma linha reta, formando 45° com o eixo  $x$ . A principal virtude desse método é detectar estruturas cíclicas ainda presentes nos resíduos.

## Previsão Utilizando Modelos de Séries Temporais

Uma vez estimados os parâmetros do modelo e escolhido aquele com o melhor ajuste, é possível fazer-se previsões de pontos futuros no tempo com respectivos intervalos de confiança. A técnica utilizada para fazer as previsões é a aplicação do filtro de Kalman.

Os intervalos de confiança 95% são construídos para as estimativas dos pontos preditos pelo modelo são calculados assumindo normalidade dos erros-padrão das estimativas, sendo que neste caso, para uma série temporal  $Y_i$ :

$$IC\ 95\% = \hat{y}_i \pm 1.96 \times SE(\hat{y}_i),$$

onde  $SE(\hat{y}_i)$  é o erro-padrão associado ao estimador  $\hat{y}_i$ .

## Correlação entre as Séries e Identificação das Funções de Transferência

A identificação das funções de transferência é feita através do estudo das correlações cruzadas entre as séries de entrada no sistema e a série de saída.

Este estudo não deve, entretanto, ser feito com as séries originais, pois a presença de qualquer tipo de estrutura, especialmente na série de entrada, pode levar a uma identificação imprecisa do modelo adequado. O problema é que a presença de estruturas (e.g. tendências, sazonalidades/ciclicidades e até mesmo estrutura de autocorrelação) nas séries pode levar à identificação de correlações cruzadas espúrias, como comentado anteriormente. Portanto, antes de proceder a correlação cruzada, deve-se lançar mão de métodos que permitam a retirada dessas estruturas.

O primeiro método que pode ser empregado é a diferenciação das séries de entrada e saída, e visualização da CCF com os resíduos da diferenciação de ambas as séries. Esse método só será capaz de retirar totalmente a estrutura das séries se ambas forem um passeio aleatório (veja Modelos de Séries Temporais, acima).

Então, a diferenciação, apesar de bastante eficiente para retirar tendências e sazonalidades, não consegue dar conta muitas vezes das estruturas de autocorrelação das séries. Por isso, pode ser empregado ainda o método conhecido como “pré-branqueamento” (Box et al. 1994), que consiste em modelar a série de entrada por um modelo ARIMA ou SARIMA conveniente, e aplicar o mesmo modelo à série de saída. Em geral esse procedimento é bastante eficiente em eliminar as possíveis estruturas, que atrapalhem o estudo das correlações cruzadas. Ele garante que pelo menos uma das séries (a de entrada no sistema) será compatível com ruído branco.

Haugh e Box (1972) propuseram um método algo diferente de obtenção dos resíduos, onde cada série deveria ser ajustada de maneira univariada por um modelo ARIMA ou SARIMA próprios para cada uma delas e os resíduos de cada um dos modelos usados para o estudo das correlações cruzadas. Essa técnica, é também chamada de “pré-branqueamento duplo”. Esse método é mais radical e visa tornar ambas as séries ruídos brancos. Além de tornar mais difícil a interpretação da estrutura em termos da função de transferência (Box et al. 1994), existe uma questão de quanto de estrutura deve-se retirar das séries, afinal, se ambas forem processos puramente aleatórios, elas só serão correlacionadas ao acaso.

As correlações cruzadas podem então ser utilizadas para a identificação da função de transferência mais adequada para o modelo a ser ajustado. Box et al. (1994) dá uma descrição bastante detalhada sobre os diversos formatos que a CCF assume para cada tipo de função, dependendo dos seus parâmetros:  $r$ ,  $s$  e  $b$ , onde  $r$  é a ordem da função  $\delta(B)$ , também chamado de ordem do denominador, já que  $\delta(B)$  se encontra no denominador da função de transferência;  $s$  é a ordem da função  $\omega(B)$ , também chamado de ordem do numerador, já que  $\omega(B)$  se encontra no numerador da função de transferência e  $b$  o atraso da entrada em relação à saída.

Como exemplo dos comportamentos das funções de transferência, eu vou citar alguns modelos, especificamente os que foram utilizados neste trabalho, e uma descrição dos demais modelos pode ser encontrada em Box et al. (1994).

- Modelo (0,1,1)

Neste modelo o impulso começaria a ter seu efeito notado a partir do primeiro *lag* ( $b = 1$ ) e manter-se-ia por 2 *lags* (o efeito se estende por  $s + 1$  *lags*), i.e. o efeito seria notado nos *lags* 1 e 2

- Modelo (1,0,1)

Já neste tipo de modelo, o impulso teria um decaimento exponencial, sendo notado inicialmente no *lag* 1 ( $b = 1$ ). Não é claro, porém, quantos *lags* são necessários para que o efeito não seja mais notado.

- Modelo (2,0,0)

Por fim, este modelo, que a princípio não seria de grande valia, visto que a sua flexibilidade é limitada pelo fato do efeito no primeiro *lag*  $b$  ser fixado em zero, também foi tentado no trabalho, pois com  $b = 0$ , ele corresponde a um modelo que também tem um decaimento algo exponencial, porém mais abrupto no *lag*  $b + 2$ .

Uma vez identificado o modelo de função de transferência a ser utilizado, o passo seguinte seria modelar a série sem o componente  $N_t$  (i.e sem o modelo ARIMA ou SARIMA

presentes) e proceder à identificação novamente, com os resíduos do modelo somente com a função de transferência. A abordagem feita aqui pode parecer que foi o contrário: primeiro foi identificado o componente  $N_t$  e depois foi identificada a função de transferência. Na verdade a identificação foi feita das duas maneiras (que por sinal não parece fazer muita diferença, já que os componentes são aditivos), mas como o objetivo é a comparação dos modelos sem função de transferência com os mesmos modelos com as respectivas funções de transferência, os resultados são apresentados nessa ordem.

## **Ajuste dos Modelos com Regressores Dinâmicos e Modelos de Intervenção**

Apesar dos métodos descritos para os modelos sem regressores ou com intervenções serem basicamente os mesmos a serem utilizados aqui, esta parte da análise foi feita no módulo de Séries Temporais do programa SAS® e como existem algumas diferenças entre os programas, cabem alguns comentários adicionais.

Em primeiro lugar, o algoritmo para a estimação dos parâmetros é diferente do utilizado pelo programa S-Plus® 2000 e apesar dos valores serem bastante aproximados tanto para as estimativas dos parâmetros quanto das medidas de ajuste, algumas diferenças podem ser notadas (geralmente na ordem da segunda casa decimal).

Outra diferença importante é a utilização, pelo SAS® de uma variante dos AIC, que consiste na quantidade  $n \ln(MSE) + 2k$ , onde  $n$  é o número de pontos efetivamente utilizados para a estimação do modelo e  $MSE$  é o Erro Padrão Médio do modelo (SAS Institute Inc 1999). Neste caso, como nos AIC clássicos, o objetivo deve ser a obtenção do menor valor possível, mas neste caso é comum que o valor seja negativo, dependendo do fato do MSE ser menor que 1, ao passo que o primeiro tipo é em geral uma quantia positiva.

Um outro detalhe importante, que na verdade diz respeito à previsão de pontos futuros de casos com esses modelos é que devem ser usados pontos futuros também previstos por modelos univariados das séries meteorológicas. Assim, por exemplo, quando pontos futuros são obtidos para um modelo de casos com pluviosidade como entrada no sistema, as observações futuras da série de pluviosidade são obtidas através de um modelo ARIMA/SARIMA ajustado para esta série, e não com os pontos efetivamente observados.

## **Métodos para Avaliação dos Modelos com Regressores Dinâmicos e Modelos de Intervenção**

Como nos modelos somente com o componente  $N_t$ , os resíduos dos modelos com ambos os componentes deve ser também avaliados, com respeito aos dois componentes. Assim, os resíduos dos modelos devem ser avaliados quanto às suas ACF e PACF, além dos testes de autocorrelação (Portmanteau e Periodograma Acumulado) – esta é a parte de avaliação do componente  $N_t$ . Para avaliar a adequação da função de transferência utilizada, deve-se proceder à CCF entre os resíduos do modelo e os resíduos do modelo que foi utilizado para o pré-branqueamento da série de entrada no sistema. O resultado esperado para a CCF é o da inexistência de correlações cruzadas significativas entre as séries.

## **Resultados**

### **Descrição das Séries**

- **Gráficos Temporais das Séries**

O primeiro passo para uma análise de séries temporais é a observação dos gráficos temporais das séries, para identificar os seus comportamentos durante o período de estudo, destacando a presença de tendências, sazonalidades e ciclicidades.

A Figura 2 mostra as séries de casos, pluviosidade e umidade, na forma de séries-calendário. É interessante notar inicialmente que as séries de casos e de pluviosidade apresentam um comportamento bastante semelhante, com uma relação direta entre a quantidade de pluviosidade e o número de casos de Leptospirose. Embora esta apresentação não permita uma distinção muito clara do tempo em relação às duas séries, essa relação parece ter um atraso, com a chuva precedendo um ou dois pontos o aparecimento de casos. Esse fenômeno será melhor estudado adiante, quando as séries forem cruzadas para a comparação das séries duas a duas.

As épocas que apresentam o maior número de casos e também a maior quantidade de chuva são os meses de março e abril. Não se pode definir claramente uma sazonalidade em ambas as séries e parece haver um número mais elevado de casos na primeira metade do período. A variabilidade de ambas as séries parece ser bastante acentuada, com os casos variando de zero até cerca de 30 casos e a pluviosidade de zero até mais de 300 mm (embora este constitua um ponto claramente extremo).

A umidade se mantém estável e alta durante todo o período, em torno de 80-85%, e mesmo sendo visualizada em escala ótima, é difícil distinguir uma sazonalidade claramente. A variabilidade não parece constante, mas certamente é mais estável que as duas séries anteriores.

A Figura 3 apresenta as séries de temperaturas máxima, média e mínima, todas em mesma escala para permitir comparação. As três séries apresentam clara sazonalidade, com temperaturas mais altas no verão e mais baixas no inverno, como esperado. Parecem ser as séries mais estáveis em termos de variabilidade e não se nota tendência clara em nenhuma

delas. É bom salientar que em média a temperatura se apresenta elevada o ano todo em Salvador.

- **Transformação das Séries**

Existem duas características importantes nas séries temporais, uma obrigatória que é a estabilidade em primeira e segunda ordens e outra, desejável que é a normalidade dos resíduos da série, que é assumida, por exemplo quando se calcula o intervalo de confiança para a ACF e a PACF.

A estabilidade da média das séries pode ser obtida após uma diferenciação conveniente da série, mas a variabilidade nem sempre é controlada com este procedimento. Uma boa maneira de se observar o comportamento da variância da série é a observação dos gráficos temporais da série diferenciada. A distribuição da série pode ser observada por histogramas. O que se constatou em relação às 6 séries em questão, é que as séries de casos e pluviosidade apresentam tanto uma instabilidade da variância quanto uma distribuição bastante desviada para a direita.

Assim, algumas transformações foram tentadas para estas duas séries. As mais recomendadas neste caso seriam a aplicação de raiz quadrada ou de logaritmo natural. A radiciação das séries pode ser feita diretamente, mas apesar de melhorar o problema da variação, não ajuda muito na questão da assimetria, como pode ser verificado nas Figuras 4 para os casos e 5 para a pluviosidade. Já a aplicação do logaritmo natural sofre com o problema da presença de zeros, que não podem ser transformados e deveriam ser recodificados para *missing*. Para contornar este problema, ambas as séries foram deslocadas em uma unidade, para manter os pontos com valor zero. Os resultados podem ser observados também nas Figuras 4 e 5. Essa última transformação é capaz de combinar as duas características, i.e. estabilizar a variabilidade e normalizar a distribuição, e portanto será adotada. Sendo assim, deste ponto adiante, quando as séries de casos e pluviosidade forem citadas, estas serão as séries transformadas, exceto quando assinalado o contrário.

Na Figura 6 pode ser observado que as séries de umidade e de temperaturas não padecem destes problemas, com variabilidade algo estável e distribuições razoavelmente normais.

- **Sazonalidade das Séries**

Um importante componente que deve ser identificado nas séries temporais é a sua sazonalidade e também possíveis ciclicidades. Em séries como as de temperaturas apresentadas, a simples observação dos gráficos temporais é suficiente para a sua identificação. No caso de séries como casos, pluviosidade e umidade esse padrão não é tão claro assim.

Para a identificação várias técnicas podem ser empregadas. Uma delas é uma decomposição sazonal, feita com o uso de regressão local (LOESS), que consiste em tentar identificar um padrão de repetição (sazonal) na série e separá-lo da série original. O resíduo deste procedimento pode ser usado para a obtenção de modelos livres de sazonalidade, quando o objetivo é trabalhar com as séries sem o efeito da sazonalidade e não modelá-la.

O resultado dessa decomposição, que foi feita neste trabalho apenas a título de ilustração, pois os resíduos destes modelos não serão utilizados, é apresentado na Figura 7. As curvas foram convenientemente deslocadas para poder-se visualizar o comportamento dos componentes sazonal e não sazonal (que seriam o resíduo), comparado com a série original. A única curva que está na escala correta é a do componente sazonal, que é sempre apresentada com média zero. O fato de se observar um componente sazonal nos gráficos não significa, necessariamente, que estes componentes estejam realmente presentes e que valha a pena modelá-los, já que o método extrai um componente sazonal de qualquer jeito e mesmo séries simuladas sem componente sazonal podem ter um componente identificado por esse procedimento.

Além da já constatada e clara sazonalidade das séries de temperaturas, o que é denotado não só pelo fato de serem as curvas mais “evidentes” de sazonalidade, mas ainda pelo fato dos resíduos terem ficado praticamente linearizados (i.e. o método de fato decompôs a série em dois componentes bem distintos), o que se observa é que também é possível se identificar um componente sazonal tanto na série de casos quanto na de pluviosidade. Na série de umidade, é possível que uma ciclicidade esteja presente também, pois se nota uma onda dentro de cada período. Este fato poderá ser levado em conta no momento de se proceder a modelagem desta série, principalmente se o objetivo for a tentativa de modelar este comportamento.

Uma outra técnica de visualização, esta tanto para a sazonalidade quanto para a instabilidade da variância, pode ser obtida pela construção de *Box-plots* juntando-se todos os pontos da série correspondentes à mesma semana do ano. A Figura 8 apresenta estes gráficos para as séries de casos, pluviosidade, umidade e temperaturas.

Pode-se observar que a sazonalidade na série de casos acompanha o mesmo comportamento da série de pluviosidade, com número de casos elevados no final do verão, o mesmo período onde são observadas as maiores precipitações. É importante destacar que apesar das séries temporais se iniciarem em 15 de março, o primeiro ponto dos *Box-plots* corresponde na verdade à primeira semana do ano. A variabilidade ainda é um problema nestas duas séries, sendo que em ambas a variabilidade é maior nos meses iniciais e finais do ano, e menor nos meses intermediários.

A umidade apresenta uma característica interessante que é a aparente presença de uma ciclicidade dupla no ano, com 2 picos (em torno da 20<sup>a</sup> e da 46<sup>a</sup> semanas) e dois vales (em torno da 1<sup>a</sup> e 38<sup>a</sup> semanas). A instabilidade da variabilidade está presente e parece que esta é menor em torno do primeiro pico comparando com o restante da série. É importante notar no entanto que este diferencial não seria tão acentuado se observássemos o gráfico em uma outra escala, pois o intervalo de variação é pequeno.

As temperaturas, além de confirmadas as suas sazonalidades marcantes, mostram-se como tendo as variabilidades mais estáveis das séries, apenas um pouco maior no segundo mês do ano.

A decisão de utilizar ou não um componente sazonal nas séries aqui analisadas será feito usando-se os critérios de ajuste dos modelos com e sem os respectivos componentes sazonais, com a ajuda dos correlogramas. Técnicas como o uso de *dummies* sazonais não serão utilizadas aqui.

## **Identificação e Ajuste dos Modelos para as Séries Temporais**

Nesta seção as séries serão estudadas individualmente, para a identificação e ajuste dos modelos ARIMA mais apropriados para cada uma delas. Apesar do objetivo principal não ser a predição destas séries individualmente, o ajuste de cada uma das séries individualmente será importante quando as séries meteorológicas forem contempladas num sistema de

entrada/saída, mais adiante, primeiro para a identificação das funções de transferência e também porque os pontos futuros estimados de cada uma das séries será usado como entrada na previsão de pontos futuros de casos (saída).

Uma importante questão a ser lembrada é que como as séries de casos e pluviosidade estudadas agora são as séries deslocadas de uma unidade e transformadas pela função logarítmica, elas terão que ser transformadas novamente com a equação inversa a esta, para a construção de gráficos temporais da série de casos dos modelos ajustados e previsões.

Para a identificação dos modelos apropriados, serão estudados os correlogramas das funções de autocorrelação (ACF) e de autocorrelação parcial (PACF), tanto para as séries propriamente ditas como para as séries diferenciadas com o objetivo de estabilizar possíveis tendências presentes nas séries e também para a identificação dos componentes sazonais das mesmas. As diferenciações feitas em todas as séries foram de ordens 1 (não-sazonal), 52 (sazonal) e 1 e 52 (sazonal e não-sazonal).

Os correlogramas com diferenciação não sazonal apresentados têm sempre 23 pontos, para permitir uma melhor visualização dos pontos próximos ao *lag* zero. Já os com diferenciação sazonal, o máximo de pontos possível, são apresentados: 141 para a sazonal apenas e 140 para a dupla diferenciação. Isto ocorre porque as séries são relativamente curtas para a modelagem do componente sazonal. Como o procedimento de diferenciação implica na perda de pontos (52 e 53 para as séries diferenciadas sazonalmente e duplamente diferenciadas, respectivamente), e as séries só possuem 194 pontos, são plotados apenas o número de pontos menos um ao todo nos correlogramas.

Com o objetivo de tornar o trabalho mais fluido, as figuras referentes a todos os estágios de identificação, ajuste e verificação dos modelos serão mostradas apenas para a série de casos. Para as demais séries, serão mostrados apenas os correlogramas, para a fase de identificação dos modelos. As demais figuras serão mostradas, se contiverem alguma informação relevante para tal. As informações gerais sobre todos os modelos serão apresentadas sob a forma de tabelas.

- **A Série de Casos**

A Figura 9 apresenta os correlogramas para a série de casos e a sua diferenciação não sazonal. A inspeção da ACF revela duas características importantes. A primeira é a presença de não estacionariedade, denotada pela lenta queda da ACF nos primeiros *lags* da série. Isto significa que a série deve ser diferenciada, e que um modelo ARIMA ou SARIMA deve ser aplicado a esta série.

A segunda é o comportamento da ACF e da PACF da série com diferenciação não sazonal, que pode sugerir um processo de médias móveis de primeira ordem puro, pela queda algo exponencial da PACF e apenas um ponto significativamente diferente de zero na ACF. Na verdade a diferenciação não sazonal apenas, foi capaz de retirar toda a estrutura da série (não mostrado).

Os correlogramas da série diferenciada no seu *lag* sazonal – Figura 10 – revelam que de fato existe um batimento significativo no *lag* 52 (correspondente ao *lag* 1 no correlograma) e que um modelo sazonal pode ser ajustado nesse caso. Tanto a diferenciação sazonal apenas quanto a dupla diferenciação revelam ainda que deve haver alguma estrutura sazonal na série, denotado pela presença de um ponto significativo no *lag* 51 na ACF e o ponto significativo no *lag* 52 da PACF da série duplamente diferenciada e pela presença deste último também na série diferenciada sazonalmente.

Outra observação é a presença de um batimento significativo no *lag* 30 na PACF em ambas as diferenciações (Figura 10), que também é observado com a diferenciação não sazonal, tanto na ACF quanto na PACF.

Baseado nessas informações, iniciou-se o processo de modelagem que é feito por tentativas e com a escolha do melhor modelo, baseado nos critérios citados anteriormente. Três diferentes modelos foram selecionados para os casos: ARIMA(1,1,1), SARIMA(1,0,1) $\times$ (0,1,1) e SARIMA(1,1,1) $\times$ (0,1,1). É interessante notar como esses modelos mais ou menos correspondem às características que foram descritas para o comportamento da ACF e PACF.

As Figuras 11, 12 e 13 apresentam os gráficos de diagnóstico de cada um dos modelos mencionados. Os resíduos dos 3 modelos não parecem apresentar qualquer tendência, nem padrão definido. A ACF e PACF dos resíduos do modelo ARIMA (1,1,1) apresenta aquele mesmo batimento no *lag* 30, embora ele não pareça muito significativo e não esteja presente

nos correlogramas dos resíduos do outros dois modelos (pelo menos não de forma significativa). O teste *portmanteau* mostrado também nas figuras que aqui é apresentado até o *lag* 15 mostra-se não significativo para todos os modelos.

A observação dos periodogramas acumulados na Figura 14 mostra claramente a diferença entre a série original e os resíduos dos 3 modelos. Não parece haver uma diferença significativa em relação à compatibilidade com ruído branco.

A Tabela 1 mostra um resumo de estimativas dos parâmetros dos modelos, suas variâncias totais e os AIC. Em termos de estimativas, os 3 modelos parecem bastante razoáveis, com erros-padrão pequenos em relação aos valores estimados e com nenhuma estimativa muito próxima de 1 (o que poderia indicar problemas no tocante à estacionariedade ou inversibilidade dos processos).

Levando-se em conta apenas os AIC e as variâncias totais dos modelos, o modelo não sazonal apresenta um resultado bem melhor que os sazonais. Observe que nesse caso o número de parâmetros a serem estimados no modelo não sazonal (2) não difere muito do número nos modelos sazonais (3).

- **A Série de Pluviosidade**

Para a série de pluviosidade, a queda da ACF parece ser mais rápida que a de casos, não sendo possível, porém, descartar a possibilidade de um modelo com diferenciação não sazonal. Após a diferenciação, o padrão apresentado é de apenas um e possivelmente dois pontos significativos na ACF e uma queda algo exponencial na PACF, podendo indicar um processo de médias móveis de primeira ou segunda ordem (Figura 15).

Os correlogramas para a série diferenciada na ordem sazonal (Figura 16) revelam um batimento significativo ser observado no *lag* 52, tanto para a ACF quanto para a PACF, o que indica que a pura diferenciação não deve ser suficiente para retirar esta estrutura e um componente talvez de médias móveis sazonal deva ser usado também. Outra observação interessante é que a estrutura do início da série é retirada, e a tentativa de modelo puramente sazonal deve ser aventada. Já os correlogramas da série diferenciada duplamente mostra uma estrutura no início da série que pode corresponder a um processo de MA e apenas um batimento na ACF no *lag* 52.

Após várias tentativas, três modelos foram selecionados para esta série. Dessa vez os modelos sugeridos pelos correlograma não foram tão acurados. Todos os modelos são importantes pois cada um representa um tipo diferente de abordagem da série, no que se refere às diferenciações. Assim, foram tentados modelos com ou sem diferenciação não sazonal e sazonal, devido às características dos correlogramas. Os modelos escolhidos foram: ARIMA (3,0,0) – ou AR (3); ARIMA (0,1,1) e SARIMA (1,0,1)x(0,1,1). Todos os testes diagnósticos se mostraram razoáveis para todos os modelos em questão.

Os diagnósticos gráficos não apresentam diferença significativa entre os modelos e a Figura 17 mostra ainda os periodogramas acumulados, onde o modelo sazonal parece ter um melhor desempenho que os demais.

A Tabela 1 apresenta o resumo dos modelos. O modelo AR (3) apresenta um coeficiente de segunda ordem menor que seu erro-padrão, o que não é obviamente recomendável, apesar dele apresentar a menor variância total. Já em termos de AIC, é o pior dos três apresentados. O modelo IMA (0,1,1) é o que apresenta os menores AIC e o único modelo sazonal selecionado para a série de pluviosidade foi ajustado sem diferenciação não sazonal e apresenta o pior ajuste em termos de AIC.

- **A Série de Umidade**

Observando-se os correlogramas não sazonais para a série de umidade, a queda da ACF parece ser rápida, mas esta estrutura é retirada com a diferenciação não sazonal, o que indica a tentativa de um processo puramente autorregressivo, talvez de ordem 3, que é o último *lag* significativo na PACF. Com a diferenciação, o aspecto se inverte: a PACF parece decair exponencialmente e a ACF tem um ponto significativo no *lag* 1, sugerindo um modelo IMA (0,1,1) – Figura 18.

Não se observa nenhum batimento nos *lags* sazonais com a diferenciação não sazonal (não mostrado), mas com as diferenciações sazonal e sazonal e não sazonal, a ACF mostra um batimento importante no *lag* 52, o que indica a tentativa de um modelo sazonal também, ambas com estrutura ainda no início da série. Quanto às PACF, não há batimentos no *lag* 52 e é improvável que um processo AR sazonal esteja presente. A PACF da série duplamente

diferenciada apresenta claramente uma estrutura de MA, de ordem 1 no componente não sazonal.

Várias tentativas foram feitas e 3 modelos são apresentados a seguir. Dentre os modelos sem diferenciação alguma, i.e. assumindo estabilidade da série, o modelo ARIMA(1,0,1) – ou ARMA(1,1) – se mostrou superior ao ARIMA(3,0,0) – AR(3) – e será apresentado. Dentre os modelos com diferenciação não sazonal apenas, IMA(0,1,1) foi selecionado e por fim um modelo SARIMA(0,1,1)x(0,1,1) foi selecionado para representar os modelos sazonais tentados, já que os modelos com diferenciação sazonal apenas tiveram resultados bastante inferiores.

Novamente, os diagnósticos gráficos destes modelos se mostram razoáveis, e a Tabela 1 apresenta as estimativas para os parâmetros dos mesmos, variâncias e AIC. É interessante notar que os menores AIC são obtidos pelo modelo ARMA(1,1), que apesar de não ser diferenciado na sua ordem não-sazonal apresenta um  $\hat{\alpha}_1$  elevado, mas não preocupante em termos de instabilidade.

- **As Séries de Temperatura Mínima, Média e Máxima**

Apesar das séries de temperaturas terem cada uma um ajuste próprio, os seus gráficos diagnósticos, como dever-se-ia esperar são muito parecidos. A série de temperatura média foi escolhida para representar o grupo das temperaturas na observação dos correlogramas.

O que se nota logo de saída é uma estrutura de autocorrelação bem marcante na ACF. Surpreendentemente a diferenciação não sazonal apenas é capaz de retirar toda a estrutura presente na ACF (Figura 20), inclusive sem batimento algum nas *lags* sazonais (não mostrado), embora a estrutura presente na PACF seja algo confusa, sugerindo que um modelo não sazonal para as temperaturas seria possível, mas provavelmente de ordem algo elevada.

Por outro lado, o correlograma da série diferenciada apenas no *lag* sazonal não é capaz de retirar a estrutura de autocorrelação, mas apenas a altera, de modo que um modelo com diferenciação apenas sazonal não deve ter um ajuste bom (Figura 21). Com ambas as diferenciações a estrutura é de fato retirada, mas um modelo sazonal com algum grau de MA deve ser tentado, pelos batimentos em ambas a ACF e PACF no *lag* 52.

As mesmas características descritas para a temperatura média foram reveladas pelos correlogramas das temperaturas máxima e mínima, não sendo necessária a apresentação dos mesmos.

Assim, vários modelos foram tentados para as 3 séries separadamente e como esperado, modelos sem diferenciação ou com diferenciação apenas sazonal tiveram péssimos resultados para todas elas. Novamente 3 modelos foram escolhidos para cada uma das 3 séries e as suas estimativas, variâncias e AIC podem ser conferidos na Tabela 2. Os modelos apresentados são: ARIMA(2,1,3); ARIMA(5,1,1) e SARIMA(1,1,2)x(0,1,1) para a temperatura máxima, ARIMA(2,1,3); ARIMA(5,1,1) e SARIMA(1,1,1)x(0,1,1) para a temperatura mínima e ARIMA(2,1,3); SARIMA(1,1,2)x(0,1,1) e SARIMA(2,1,2)x(0,1,1) para a temperatura média.

A primeira característica que deve ser notada é a semelhança dos modelos (como esperado pela observação das ACFs e PACFs). O melhor modelo não sazonal para todas as séries é o mesmo – ARIMA (2,1,3). Outro modelo não sazonal foi tentado para as séries de temperatura máxima e mínima. A julgar-se pelos AIC e pela variância total do modelo, seriam até razoáveis, mas as estimativas dos seus parâmetros são muito próximas dos seus erros padrão (especialmente as estimativas dos parâmetros de autocorrelação), o que torna estes modelos mais fracos.

Quanto aos modelos sazonais, cada série foi melhor ajustada por uma ordem, mas todas próximas. Apesar dos AIC e variância serem maiores do que os outros modelos, estas séries obrigatoriamente têm que ser modeladas por um modelo sazonal, dada a flagrante sazonalidade presente nas séries.

### **Correlação entre as séries**

Para o estudo das correlações entre as séries meteorológicas e a série de casos, não basta apenas a aplicação da função de correlação cruzada, porque, como já foi comentado, uma correlação espúria pode ser resultado apenas da presença de estruturas nas séries envolvidas.

Uma primeira tentativa de se obter resíduos compatíveis com ruído branco seria tentar simplesmente diferenciar as séries e aplicar a CCF. Este procedimento, porém não se

demonstrou ser muito eficaz, como pode ser observado na Figura 22, já que a estrutura de correlação entre as séries parece seguir um comportamento um tanto confuso, com muitos *lags* significativos e ainda em diferentes sentidos.

Um passo seguinte neste processo é a aplicação do “pré-branqueamento” propriamente dito, que consiste em modelar a cada uma das séries de entrada com o melhor modelo possível e aplicar o mesmo modelo à série de saída (casos).

Como pode ser observado na Figura 23, este procedimento resulta num padrão bem mais adequado, com apenas dois *lags* significativos e sempre na mesma direção para todas as séries cruzadas com a série de casos. Para as de pluviosidade e umidade, ambos os *lags* apresentam correlações positivas com os casos – i.e. quanto mais pluviosidade ou mais umidade, maior o número de casos como consequência, e isso ocorre com um atraso no tempo de uma semana e o efeito se mantém por duas semanas. A diferença entre elas está primeiro na intensidade, que é maior para a umidade e também quanto ao segundo *lag*, que se apresenta quase com a mesma intensidade do primeiro no caso da pluviosidade, e que decai um pouco para a umidade.

Já no caso das temperaturas, representadas na Figura 23 pelas máxima e média, as correlações são negativas e temperaturas mais altas diminuiriam o consequente número de casos presentes, sendo as características de atraso e duração do efeito iguais para todas as séries. Quanto à intensidade, não há diferença entre as temperaturas, ficando em torno de  $-0.3$  e o segundo *lag* apresenta uma queda, no caso da temperatura máxima, mas não para a média. A temperatura mínima, não mostrada, tem o mesmo padrão da temperatura máxima.

A técnica, de “pré-branqueamento duplo” também foi tentada, apenas a título de curiosidade e apresentou resultados bastante parecidos com o do “pré-branqueamento” clássico, sempre que os modelos de entrada e saída eram ambos não sazonais ou ambos sazonais. Quando havia uma discordância quanto à sazonalidade dos modelos, os resultados pareceram bastante confusos, lembrando os obtidos com diferenciação apenas.

Baseado na observação das correlações cruzadas entre as séries, é lícito supor que o período de incubação da leptospirose nessa população foi de 1 a 2 semanas, correspondendo ao *lags* significativos entre a série de casos e as demais séries meteorológicas.

Ainda dentro do estudo de correlações cruzadas entre as séries, é importante também estabelecer que relações as séries meteorológicas guardam entre si. Como neste caso o objetivo não é a identificação de funções de transferência entre as séries, as correlações cruzadas entre as séries meteorológicas foram estabelecidas com a CCF entre as séries apenas diferenciadas.

A Figura 24 mostra a correlação positiva e sem atraso entre as séries de pluviosidade e umidade. A correlação negativa no *lag* 1 além de muito menor relativamente, provavelmente se deve à presença de autocorrelação que ambas as séries mantêm após a diferenciação. Para as correlações com as temperaturas, são mostradas apenas em relação à temperatura média, já que as outras duas apresentam um padrão bastante parecido. Assim, ainda na Figura 24 pode-se observar a correlação negativa entre a pluviosidade e temperatura média e umidade e temperatura média. Para exemplificar a correlação entre as temperaturas, a CCF entre temperatura máxima e mínima também é apresentada, e existe uma altíssima correlação positiva sem atraso, com coeficientes perto de 0.7. As correlações entre as temperaturas máxima e mínima em relação à temperatura média apresentam coeficientes próximos a 0.9.

## **Identificação e Ajuste dos Modelos com Funções de Transferência**

Uma vez estudado o padrão de correlação cruzada entre as séries, é possível agora a identificação de funções de transferência que melhor se ajustem a cada um dos sistemas. Pelo menos 3 tipos diferentes de funções de transferência podem ser tentadas neste caso, que são os mesmos para todas as séries, visto que o padrão de correlação cruzada é o mesmo entre a série de casos e as demais séries, diferindo apenas em termos de sentido e magnitude do padrão de correlação (Figura 23).

O primeiro seria o caso de assumir que na verdade o impulso de entrada nos *lags* 1 e 2 são de igual magnitude, o que representaria uma função do tipo  $(0,1,1)$ , onde o efeito se inicia no segundo *lag* e se mantém por 2 *lags* consecutivos. Uma outra hipótese é que o decaimento deste impulso se dê de maneira exponencial, sugerindo uma função do tipo  $(1,0,1)$ . Por fim, uma função do tipo  $(2,0,0)$  poderia ser tentada, mesmo que ela corresponda, em termos, à função anterior, pois o primeiro ponto do impulso nesse caso é fixado em zero (i.e. o efeito só começa a ser observado no segundo *lag*, apesar desta função não ser atrasada).

Vários modelos com regressores dinâmicos foram tentados, utilizando-se os três tipos de funções de transferência citados, além dos 3 modelos SARIMA/ARIMA escolhidos anteriormente para a série de casos (Tabelas 1 e 2). Vale lembrar que as séries meteorológicas entram no modelo apenas diferenciadas e não como os resíduos decorrentes do “pré-branqueamento”.

Como já foi comentado, a seqüência que deve se seguir para encontrar um modelo com ajuste adequado é modelar a série apenas com a função de transferência e estudar os resíduos deste modelo para ajustar o componente ARIMA/SARIMA. Mas como o que se quer comparar aqui são modelos com e sem funções de transferência e como na verdade os resíduos dos modelos apenas com as funções de transferência são muito parecidos com os da série somente, a apresentação é feita ao contrário: para cada modelo SARIMA/ARIMA da série de casos os modelos de função de transferência são comparados.

Como diagnóstico de ajuste desses modelos, os resíduos dos mesmos foram analisados em termos de autocorrelação e ainda de correlação cruzada com os resíduos dos modelos usados para o diagnóstico das funções de transferência. Todos os 45 modelos que serão apresentados mostraram resultados satisfatórios, mas a título de exemplo, 3 deles serão apresentados. A Figura 25 mostra os diagnósticos para o modelo SARIMA(1,1,1)x(0,1,1) com a pluviosidade como regressor dinâmico, de ordem (2,0,0). O gráfico superior é o correlograma da ACF do modelo completo, mostrando ausência de estruturas de autocorrelação; o segundo correlograma é o da série simples do modelo usado para a pluviosidade – nesse caso um AR(3) – que mostra a ausência de estrutura para os seus resíduos; e o gráfico inferior, o correlograma da CCF, mostrando que a estrutura de correlação cruzada entre as séries foi também retirada pelo modelo.

A Figura 26 mostra os mesmos gráficos, mas para o modelo SARIMA(1,0,1)x(0,1,1) com umidade modelada por um ARMA(1,1) e com função de transferência de ordem (2,0,0) e a Figura 27, o modelo SARIMA (1,0,1)x(0,1,1) com temperatura média modelada por um ARIMA(2,1,3), com função de transferência de ordem (0,1,1). Como no caso do modelo anterior, tanto a estrutura de autocorrelação quanto a de correlação cruzada em ambos os casos foram satisfatoriamente retiradas.

Agora vamos passar a analisar separadamente os modelos com regressores dinâmicos para cada um dos modelos de casos. Só para lembrar novamente, os modelos são comparados

em termos das suas variâncias totais, de seus AIC (que aqui são na verdade a versão modificada e baseados no MSE), além é claro das estimativas dos coeficientes e seus erros-padrão. É bom lembrar que como se trata da versão modificada dos AIC, eles devem ser usados para comparar modelos que tenham o mesmo componente ARIMA/SARIMA, já que a dependência deste no número de observações efetivamente utilizadas  $n$  pode levar a valores muito diferentes, ao se comparar modelos de ordens diferentes.

- **Regressores Dinâmicos e o Modelo ARIMA(1,1,1)**

A Tabela 3 apresenta os resultados para as 3 funções de transferência usadas, tendo como ruído o modelo ARIMA(1,1,1) para a série de casos. A primeira observação é que as estimativas dos coeficientes de autorregressão e médias móveis dos modelos são bastante parecidas com as do modelo sem a função de transferência, seja qual for a série que entra como regressor dinâmico. A avaliação dos modelos não deve ser feita somente pela observação da variância e AIC, pois como pode ser notado muitas das estimativas dos coeficientes das funções de transferência apresentam erros-padrão inaceitáveis (muitas vezes maiores que as próprias estimativas).

De maneira geral, combinando essas avaliações, o modelo com pluviosidade e o modelo com umidade apresentam um melhor ajuste com uma função de transferência de ordem (2,0,0). Os modelos são os de menor variância e AIC em relação aos demais, e apesar de o erro-padrão das estimativas não serem muito pequenos, (como é o caso dos outros modelos), o valor das estimativas dos modelos (2,0,0) para estas séries é bem maior que zero, o que não acontece com os demais.

Já os modelos com as temperaturas como regressores dinâmicos ajustaram melhor com a função de ordem (0,1,1), que nesse caso têm variâncias parecidas com o de ordem (2,0,0), mas estes apresentam erros-padrão inaceitavelmente elevados.

- **Regressores Dinâmicos e o Modelo SARIMA(1,0,1)x(0,1,1)**

A Tabela 4 apresenta os resultados para as 3 funções de transferência usadas, tendo como ruído o modelo SARIMA(1,0,1)x(0,1,1) para a série de casos. Comparando-se com as Tabelas 1 e 2, as estimativas dos coeficientes de autorregressão e médias móveis, ao contrário dos modelos anteriores, são algo diferentes do modelo sem as funções de transferência, embora sejam bastante parecidos entre si, seja qual for a série que entra como regressor

dinâmico. O motivo desta diferença provavelmente está no fato das séries aqui serem modeladas com a média presente no modelo (já que o programa SAS® aceita que o modelo tenha um intercepto e modele esta média, o que não acontece no S-Plus® 2000).

Como no modelo anterior, combinando as avaliações, o modelo com pluviosidade e o modelo com umidade apresentam um melhor ajuste com uma função de transferência de ordem (2,0,0). Neste caso, as variâncias e AIC destes modelos são bem menores que os demais e o mesmo comentário do erro padrão das estimativas feito para o grupo anterior de modelos vale para este grupo também.

Já os modelos com as temperaturas como regressores dinâmicos ajustaram melhor novamente com a função de ordem (0,1,1), avaliando-se não pelas variâncias que realmente são menores para o modelo (2,0,0), mas pelos seus erros-padrão inaceitavelmente elevados.

- **Regressores Dinâmicos e o Modelo SARIMA(1,1,1) x(0,1,1)**

A Tabela 5 apresenta os resultados para as 3 funções de transferência usadas, tendo como ruído o modelo SARIMA(1,1,1)x(0,1,1) para a série de casos. Comparando-se com as Tabelas 1 e 2, as estimativas dos coeficientes de autorregressão e médias móveis, são bastante parecidas com as dos modelos sem as funções de transferência, exceção feita ao modelo com umidade, ordem (2,0,0), que não só apresenta uma estimativa para o coeficiente de autocorrelação menor que os demais modelos, mas também um erro-padrão bastante elevado.

Assim, para a pluviosidade, uma função de transferência de ordem (2,0,0) foi selecionada como a melhor dos três, ao passo que para a umidade o de ordem (1,0,1) parece combinar o maior número de qualidades de ajuste.

Os modelos com as temperaturas como regressores dinâmicos não apresentam nenhuma surpresa e novamente ajustaram melhor com a função de ordem (0,1,1). Nesse caso, a diferença na variância não foi tão acentuada como no caso dos modelos SARIMA (1,0,1)x(0,1,1).

Um próximo passo que poderia ser tentado seria a construção de um modelo que utilizasse mais de um regressor dinâmico (i.e. mais de uma série como entrada). Este procedimento, porém foi desencorajado pelo fato de todas as séries meteorológicas serem

altamente correlacionadas entre si no *lag* 0, como já foi comentado, e pode ser observado na Figura 24. Neste caso a presença de mais de um regressor dinâmico no modelo acarretaria uma menor precisão do mesmo.

- **Comparação dos modelos com e sem Regressores Dinâmicos**

Uma vez identificados os “melhores” modelos, é possível agora estudar as contribuições da presença dos regressores dinâmicos para a melhora da previsão do número de casos, tanto em termos de pontos presentes na série como de pontos futuros. Para isso serão apresentados os gráficos das séries ajustadas com os seus intervalos de confiança 95%, os pontos das séries, as previsões para 12 semanas futuras, também com seus IC 95% e ainda os 4 pontos que foram separados da série original de casos, para comparação. Os resultados dos modelos foram transformados de volta para os seus valores originais.

Como mencionado anteriormente, para a obtenção dos pontos futuros nos modelos com regressores dinâmicos, os valores que entram no modelo como estimativas futuras dos regressores devem ser obtidas através de um modelo ARIMA/SARIMA e não ser uma observação real. Assim, para cada uma das séries foi escolhido um modelo para gerar estes pontos.

Para a série de pluviosidade e também de umidade, foi usado o modelo IMA(1,1), cujos resultados podem ser observados na Tabela 1 e para as temperaturas foram utilizados modelos sazonais – SARIMA(1,1,2)x(0,1,1) para temperaturas máxima e média e SARIMA(1,1,1)x(0,1,1) para a temperatura mínima. A escolha seguiu os mesmos critérios mencionados para os demais modelos, mas no caso das temperaturas, foram escolhidos modelos sazonais, devido à clara sazonalidade presente nessas séries, o que leva a previsões mais precisas.

As Figuras 28, 29 e 30 apresentam as curvas ajustadas e as previsões para os modelos ARIMA(1,1,1) sem regressores e com cada uma das séries como regressor dinâmico. As ordens das funções de transferência utilizadas foram aquelas apontadas na seção anterior como as de melhor ajuste. O que se nota logo de início é que vários pontos da série caem fora do intervalo de confiança. Esse é um efeito na verdade da transformação que os valores tiveram que sofrer, já que com a série em valores não transformados de volta muito menos pontos caem fora do IC 95%.

Outra observação importante é que o intervalo de confiança de todos os modelos apresentam valores inferiores a zero no seu limite inferior. Isso é um efeito de estarmos lidando com modelos criados para dar conta de variáveis contínuas e que podem assumir valores menores ou maiores que zero. Como nesses modelos a sazonalidade não foi levada em conta, observa-se também uma certa “rigidez” dos pontos futuros estimados, principalmente no modelo sem regressores, onde estes são quase uma linha paralela ao eixo do tempo, com o IC alargando consideravelmente com o passar do tempo.

A presença de regressores dinâmicos, especialmente as temperaturas parece de fato contribuir com um movimento mais suave do modelo, mas ainda sem refletir a sazonalidade presente no modelo de casos. Também a variação inicial na série parece ser mais levada em conta na predição com a presença dos regressores.

As Figuras 31, 32 e 33 apresentam os mesmos gráficos para o modelo  $SARIMA(1,0,1) \times (0,1,1)$  e as Figuras 34, 35 e 36 para as séries o modelo  $SARIMA(1,1,1) \times (0,1,1)$ . O que se nota de início é o fato de nesse caso a “perda” de pontos ser bem maior que no modelo  $ARIMA(1,1,1)$  e só existirem predições a partir do segundo ano da série. A maior instabilidade vista no início dos pontos preditos se deve provavelmente à essa região de pontos extremos no início da série, mas que depois é melhor estabilizado, principalmente no caso dos modelos diferenciados não sazonalmente.

Os pontos futuros preditos parecem ser mais precisos com os modelos sazonais, sendo que as flutuações estão presentes e o IC não parece se alargar tanto como no modelo não sazonal. A questão aqui é que a presença das demais séries no modelo não parece ajudar muito nesse resultado, apenas ajudando na estabilidade dos pontos iniciais preditos, pelo menos ao se observar apenas as Figuras 28 – 36.

Para estudar melhor o efeito da entrada dos regressores dinâmicos na previsão de pontos futuros, lançarei mão de uma observação um pouco mais minuciosa dos dois componentes da estimação, que são os valores futuros preditos, que podem ser comparados com as observações reais (neste caso, com os 4 pontos que tinham sido separados) e também estudar o comportamento dos ICs para os pontos futuros.

É claro que uma comparação das estimativas pontuais apenas não faz muito sentido em termos estatísticos, mas o estudo da variabilidade percentual das estimativas, comparadas com o valor observado de casos pode levar a algumas observações interessantes. Isso é mostrado graficamente na Figura 37, que combina a variabilidade dos 18 modelos considerados acima, separadamente para cada modelo ARIMA/SARIMA ajustado para os casos. A comparação foi feita em relação às estimativas já transformadas de volta para a escala inicial e os números não foram arredondados.

A primeira observação interessante é que a variabilidade do primeiro ponto da série sem regressores é bastante alto para os 3 diferentes grupos de modelos considerados e que a presença dos regressores dinâmicos a diminui significativamente. Esse comportamento pode ser explicado pelo fato do último ponto observado usado para a estimativa é bastante alto (7 casos foram observados) e o primeiro ponto tem uma estimativa alta (em torno de 5 casos), já que como todos os modelos têm um componente AR de primeira ordem, o valor imediatamente passado influencia bastante a previsão. Como só foram observados 2 casos na realidade, a variabilidade disparou. Na presença dos regressores, porém, o efeito foi de fato amenizado e estimativas mais precisas foram obtidas. Nos outros 3 pontos comparados o que ocorre é que os modelos sem componente sazonal para os casos tendem a superestimar os pontos futuros de uma maneira quase linear, não importando a presença ou não de regressores. Já os modelos com componente sazonal parecem fornecer de fato estimativas mais estáveis, embora o modelo com diferenciação sazonal apenas apresente uma tendência a subestimar os casos, ao passo que o modelo com dupla diferenciação parece superestimá-los. A presença dos regressores não parece mesmo ter muita influência nesse caso.

Para o estudo do comportamento dos ICs a melhor opção é estudar o comportamento do erro-padrão associado a cada estimativa considerada, já que o IC é apenas uma função linear dessa medida (estimativa  $\pm 1.96 \times$  erro-padrão). A opção mais uma vez foi estudar o erro-padrão já transformado para a unidade inicial.

O gráfico superior da Figura 38 mostra o comportamento do erro-padrão para os 3 conjuntos de modelos, para os 12 pontos que foram previstos. O conjunto de modelos sem componente sazonal para os casos apresentam um aumento inicialmente parecendo que estaria estabilizando, mas logo aumentando de forma linear, sendo que os modelos com as temperaturas parecem ter um resultado algo melhor. O modelo com dupla diferenciação apresenta também uma tendência linear de aumento do valor do erro-padrão, mas com um

ângulo muito menor que aquele do modelo não sazonal. Neste caso o modelo com pluviosidade é o que tem monotonamente o menor erro. Por fim, o melhor conjunto de modelos parece ser o sazonal com diferenciação não sazonal, que parece mesmo atingir um *plateau* e manter um erro quase constante para os pontos a partir do sétimo erro-padrão. Novamente o modelo com pluviosidade é o que tem monotonamente o menor erro.

O gráfico inferior da Figura 38 é um detalhe dos erros-padrão de todos os modelos considerados sem regressores dinâmicas e com pluviosidade como regressor, já que de maneira geral os modelos que tiveram os menores erros foram os com pluviosidade. O que se nota é que apenas no primeiro ponto o modelo não sazonal com pluviosidade tem um erro um pouco menor que o modelo sazonal duplamente diferenciado com pluviosidade. Ambos os modelos sazonais com pluviosidade mantêm o erro-padrão abaixo de todos os outros modelos considerados a partir do segundo ponto predito, mas o modelo com diferenciação não sazonal apenas de fato aumenta com o tempo ao passo que o de dupla diferenciação parece muito mais estável.

## **Construção de um Modelo de Intervenção**

A seguir, baseado nos modelos obtidos, um estudo de intervenção foi tentado primeiramente averiguar se o aparente declínio que ocorreu na série de casos na segunda metade da série, i.e. a partir de 1998 é significativa ou não, na presença ou não das séries meteorológicas.

O tipo de intervenção utilizada neste caso foi a do tipo “degrau”, iniciando na primeira semana de Janeiro de 1998. Se a diminuição estiver presente no modelo sem os regressores dinâmicos, mas desaparecer na presença destes, isto significa que na verdade não houve influência de outros fatores no declínio do número de casos além da própria modificação nas variáveis meteorológicas.

A intervenção foi tentada em todos os 18 modelos apresentados na seção anterior, mas em nenhum deles o erro-padrão da estimativa do coeficiente foi suficientemente pequena para sugerir a presença do efeito, nem mesmo nos modelos sem os regressores dinâmicos, o que nos leva a sugerir que na verdade não houve uma diminuição significativa do número de casos na segunda metade da série.

## **Discussão**

O objetivo geral deste trabalho é estudar as relações entre fenômenos meteorológicos e a ocorrência de casos de Leptospirose, usando-se para tal, modelos matemáticos adequados aos dados disponíveis, i.e. modelos de séries temporais.

A identificação de casos de leptospirose não é uma tarefa fácil, devido ao largo espectro que a doença pode apresentar, variando desde um quadro gripal, até formas muito graves, que podem levar à morte do paciente como a Síndrome de Weil (Martins & Castiñeiras 1994; Dupont et al. 1997; Trevejo et al. 1998). Os casos menos graves da doença em geral não são identificados, tanto pelo fato de muitas vezes o paciente sequer procurar o serviço de saúde, quanto por um não diagnóstico por parte do profissional de saúde, que muitas vezes pode classificar o quadro como gripe. A saída para esse problema seria o uso de um método diagnóstico laboratorial confiável que pudesse ser usado sempre que um caso da doença fosse aventado. O problema é que o método largamente utilizado e de escolha para a confirmação laboratorial da leptospirose – MAT – (Tappero et al. 2000) deve ser pareado no intervalo de 2 semanas, o que implica no retorno do paciente ao local de atendimento para a coleta de uma segunda amostra de soro. Além disso, a utilização de um critério amplo para a indicação da confirmação laboratorial poderia causar problemas logísticos importantes no sistema laboratorial local, visto que a presença de sinais e sintomas compatíveis com quadro gripal é comum em pacientes que procuram atendimento médico.

Assim, uma maneira de se identificar casos da doença para estudos epidemiológicos em uma população é restringir o critério de definição de casos, o que terá a vantagem de possibilitar um aumento no valor preditivo do teste positivo, diminuindo assim a proporção de testes falsos negativos, mas terá a desvantagem de não diagnosticar casos que seriam efetivamente de leptospirose. Essa abordagem na verdade é utilizada de maneira geral e o teste laboratorial é pedido sempre que se suspeita da doença. Claro que critérios epidemiológicos são também utilizados, como por exemplo contato definido com urina de ratos.

Para este estudo, os casos foram restringidos àqueles que apresentavam pelo menos um critério de gravidade da doença. Apesar desse critério não abarcar todo o conjunto de casos verdadeiramente presentes na população, é razoável supor que as tendências temporais

do número desses casos acompanhe as dos casos totais, apenas numa escala menor que a real. Mais uma vez, seria desejável que todos os casos incluídos no estudo fossem confirmados laboratorialmente, para não estarmos analisando casos que na verdade são doenças outras que não leptospirose. O problema do pareamento, porém, existe mesmo em se tratando de casos com alguma gravidade, e dos casos incluídos no estudo, apenas 50% tiveram confirmação sorológica.

A decisão de incluir no estudo todos os casos definidos pelos critérios clínicos descritos apenas, sem levar-se em conta a confirmação sorológica decorreu então do fato da perda de casos ser significativa nessa situação e também seletiva no tempo, pois a maioria dos resultados laboratoriais foram obtidos para os casos mais antigos da doença (na primeira metade da série). Outro fato importante foi que um estudo realizado com os casos ocorridos até novembro de 1996 (Ko et al. 1999) indicou que o valor preditivo positivo do diagnóstico clínico de leptospirose, usando-se os critérios mencionados, foi de 91%, o que parece bastante razoável e contrabalança o fato de não sermos obrigados a descartar praticamente metade da série de casos.

Um fator que introduziu imprecisão nesse estudo foi a definição utilizada para a data de início dos sintomas. A única data precisa que pode ser obtida do paciente é a de diagnóstico da doença, que é a data em que o serviço médico foi procurado e que portanto tem um registro objetivo. Essa data não poderia ser utilizada como o ponto inicial para a doença, pois cada paciente pode procurar atendimento médico em distintas fases da doença. Para contornar esse problema, é perguntado ao paciente o número de dias que ele apresenta os sinais e sintomas em questão. A imprecisão então é decorrente de possíveis vieses de memória do paciente que pode não se lembrar com precisão quantos dias vem apresentando os sinais e sintomas ou ainda não assinalar um determinado sintoma, por considerá-lo de menor importância.

A questão do local de residência dos casos é outro fator a ser considerado. Como o Hospital Couto Maia é um centro de referência para doenças infecciosas, é esperado que pacientes que residam em áreas afastadas de Salvador sejam referenciados ou até procurem o Hospital por iniciativa própria. Esse fato deve ser levado em consideração porque como o objetivo aqui é o estudo das relações com as séries meteorológicas, a exposição do paciente a esses fenômenos deve corresponder aos dados meteorológicos disponíveis, ou seja, o paciente deve ter sido exposto à pluviosidade, umidade e às temperaturas que foram efetivamente

aferidas e usadas no estudo. A primeira questão é que se o paciente mora em um local muito afastado de Salvador, é muito pouco provável que a sua exposição corresponda ao que foi efetivamente aferido. Por esse motivo, os casos foram restringidos a pacientes residentes na região Metropolitana de Salvador, composta pelas 8 cidade citadas anteriormente. Uma segunda questão é o uso do local de residência para ser o marcador da exposição. Na verdade não há nenhuma garantia de que o paciente foi exposto próximo ao seu local de residência. Ele pode ter sido exposto por exemplo próximo ao seu local de trabalho, que pode ser bem distante do seu local de residência, até mesmo em outra cidade.

As variáveis meteorológicas, por sua vez, foram coletadas em um único centro de aferição, localizado em um bairro de Salvador. É claro que a aferição neste local não traduz de maneira precisa o que ocorre na cidade e nem tão pouco na Região Metropolitana como um todo e deve ser considerada como um comportamento médio dessas variáveis nessa região. A variabilidade nesse caso, porém, não deve ser considerada como um fator de total imprecisão, porque não é razoável supor que se, por exemplo, a temperatura aferida é de 30°C, o paciente esteja tão distante que sua temperatura de exposição seja 10°C.

Combinando todas essas imprecisões presentes com respeito aos casos e aos fenômenos meteorológicos, é importante, então, ter-se em mente que os resultados analisados aqui devem ser considerados como comportamentos médios de todos os fenômenos.

As séries estudadas tiveram que sofrer algumas transformações para tornar possível a utilização dos modelos empregados neste trabalho. A primeira modificação que foi introduzida em todas as séries foi a agregação em semanas. A necessidade dessa agregação deveu-se principalmente ao fato das séries de casos e pluviosidade apresentarem características não desejáveis em relação aos modelos utilizados. O número de casos além de ser uma variável discreta de contagem, apresenta um número extremamente elevado de dias onde nenhum caso foi detectado. A pluviosidade por sua vez, apesar de ser uma variável contínua e aferida em milímetros de chuva, apresenta o mesmo problema de muitos dias por ano sem chuva (o que é característico de Salvador). Para se ter uma idéia do problema, a série diária de casos tem um total de 755 dias (54.43%) com zero casos registrados e a série semanal apenas 12 semanas (6.12%) com zero casos. Já a pluviosidade diária apresenta 622 dias (44.84%) sem chuva, ao passo que a série semanal apresenta 11 semanas (5.60%) sem chuva alguma.

Por esse motivo, todas as séries foram agregadas em semanas o que permitiu que os modelos fossem aplicados com muito mais facilidade. Este processo, porém não é livre de conseqüências e a perda de precisão na estimativa dos parâmetros e dos pontos futuros foram descritas após a agregação de séries temporais (Tiao & Wei 1976).

Esse procedimento apesar de ter melhorado muito a distribuição destas duas séries, não foi suficiente para resolver totalmente seus problemas. Numa primeira análise das séries semanais de casos e pluviosidade, foi constatado que além de permanecerem extremamente assimétricas à direita, apresentavam uma instabilidade nas suas variabilidades, como pode ser observado nas Figuras 4 e 5. O problema da distribuição não é impeditivo para o uso dos modelos clássicos usados no estudo, mas a normalidade do resíduo dos modelos é desejável, principalmente quando levarmos em conta os intervalos de confiança para os testes de ruído branco e ainda a construção dos intervalos de confiança para as estimativas das séries. Já o problema da variabilidade é mais grave e a estacionariedade da série deve ser obtida tanto em primeira como em segunda ordens.

A aplicação da raiz quadrada melhora de fato a questão da variabilidade, mas não chega a melhorar muito o problema da assimetria da distribuição em ambas as séries. Já o uso da transformação logarítmica parece dar conta melhor dos dois aspectos. O problema da transformação logarítmica é que como ambas as séries apresentam pontos com valor zero, estes devem ser convertidos para *missing* se essa transformação for empregada. Para contornar este problema, a solução foi deslocar ambas as séries de uma unidade e assim levar em conta todos os pontos da série. A crítica a esse procedimento em relação aos casos pode ser que esse deslocamento provoca uma mudança significativa na característica da série, já que as contagens de casos não são muito altas e uma unidade pode não ser considerada como suficientemente próxima de zero. O problema é que se fosse somada a esta série um valor mais “próximo” de zero, como por exemplo 0.01 ou 0.1, a distribuição das séries seria muito distorcida, como pode ser observado na Figura 39. Por esse motivo, a transformação adotada foi deslocar ambas as séries em uma unidade e aplicar-lhes o logaritmo natural. A conseqüência desta transformação é que a série deve ser transformada “de volta” ao seu valor original após o ajuste dos modelos, o que pode distorcer os valores preditos.

Um outro problema é o fato dessa transformação poder introduzir distorções nas relações entre as séries estudadas. É claro que algum grau será introduzido de qualquer forma, mas a questão é o quanto. Para exemplificar este problema, a Figura 40 apresenta quatro

correlogramas de algumas CCFs que já tinham sido estudadas e observadas nas Figuras 23 e 24. Os modelos usados para o pré-branqueamento foram os mesmos para a umidade e temperatura média (como na Figura 23), e a comparação entre a pluviosidade e umidade foi feita apenas com a diferenciação de ambas as séries (como na Figura 24). A única diferença é na comparação da pluviosidade original e a série de casos original, onde a série de pluviosidade foi modelada novamente – neste caso eu usei um método automático de ajuste por autorregressão apenas. Comparando-se com os correlogramas onde foram usadas as séries transformadas, percebe-se que o grau de distorção não parece ser um problema grave na relação que as séries guardam entre si. O mais distorcido nesse caso foi exatamente a relação dos casos com a pluviosidade. Mas é bom lembrar que nesse caso a transformação foi aplicada a ambas as séries e além do mais, a diferença mesmo é que existe um *lag* a mais com correlação significativamente diferente de zero, porém em sentido oposto, ou seja é uma correlação negativa. Essa inversão de sentido junto com o valor algo pequeno ( $\sim 0.2$ ) pode indicar uma correlação que na verdade não existe, principalmente pelo fato de ter sido usado um método automático de ajuste (que nem sempre produz resultados satisfatórios em termos de retirada de correlações espúrias). O método automático foi usado porque nenhum modelo ajusta bem a série de pluviosidade não transformada – motivo pelo qual a série foi transformada.

Ainda em relação aos casos, um outro tipo de abordagem que poderia ter sido feita é o uso de taxas de incidência em vez do número de casos. A única característica que mudaria nesse caso é o fato de a variável passar a ser contínua e não mais de contagem. Porém, devido ao curto período das séries em questão (menos de 4 anos de acompanhamento), o resultado seria equivalente a dividir o número de casos por uma constante, mantendo assim as mesmas características quanto a sua distribuição e variabilidade.

O primeiro objetivo específico do estudo é determinar o tempo de incubação médio nessa população estudada. Uma questão interessante pode ser aqui levantada: a determinação desse tempo de incubação foi feito através do estudo das correlações cruzadas entre as séries meteorológicas e a série de casos. Existe nesse processo uma ambigüidade no sentido de que a determinação do tempo de incubação – que é o atraso que existe até a primeira correlação significativa entre as duas séries – é reforçada pelo fato de sabermos *a priori* que existe um tempo de incubação da doença e que fatores meteorológicos influenciam de fato o aparecimento de casos de leptospirose, ou seja, dado que existe uma correlação entre os fatores meteorológicos e os casos, e que essa relação deve sofrer um atraso no tempo, esse

atraso pode ser considerado como o tempo de incubação médio da doença nessa população. Por outro lado, o fato de existir uma correlação significativa com um atraso compatível com o tempo de incubação da doença reforça a conclusão de que os fatores meteorológicos de fato influenciam o aparecimento de casos de leptospirose. Na verdade, tudo é feito ao mesmo tempo: o que se constata é uma correlação significativa com um atraso de uma a duas semanas e isso significa que há uma relação entre os fatores meteorológicos e os casos de leptospirose e que essa correlação se dá com um atraso que é o tempo de incubação da doença – Figura 23.

A associação da leptospirose com fatores meteorológicos, especialmente precipitação pluviométrica é largamente descrita na literatura (Corwin et al. 1990; Everard et al. 1992; Perrocheau & Perolat 1997; Trevejo et al. 1998) e goza de uma forte plausibilidade biológica, dada a penetração ativa da espiroqueta através de micro feridas na pele do paciente, com a água da chuva servindo como um veículo perfeito para a sua locomoção. O tempo de incubação que foi constatado nesse trabalho está de acordo com o tempo de incubação médio descrito também na literatura (Martins & Castiñeiras 1994; Tappero et al. 2000).

Outros fatores que influenciam a transmissão da leptospirose são também as condições de temperatura, umidade e pH do solo, o que determina uma sobrevivência maior da espiroqueta no meio ambiente. O problema aqui é se a umidade e temperaturas atmosféricas – que é o que foi aferido nesse estudo – têm de fato uma influência na sobrevivência da espiroqueta, ou se elas são apenas uma medida indireta da pluviosidade. Como pode ser observado na Figura 24, as séries meteorológicas têm de fato uma alta correlação e sem atraso entre si, e isso é um forte indício de que estamos medindo o mesmo fenômeno de várias maneiras diferentes. Porém, não é por esse motivo que as séries de umidade e de temperatura média devem ser descartadas. Lembrando mais uma vez que estamos lidando com aferições médias, não necessariamente a aferição da pluviosidade em um único local é o melhor a ser feito. Por exemplo, a umidade, que é uma aferição indireta da pluviosidade quando medida em um só local, pode ser mais representativa de uma média geral que a própria pluviosidade. Isso parece inclusive acontecer, já que os coeficientes de correlação da umidade são mais elevados que os da pluviosidade. Por esse motivo, não foram utilizados modelos que incluíssem mais de uma série meteorológica como regressor dinâmico simultaneamente, mas sim diferentes modelos com cada uma das séries individualmente como regressor dinâmico com relação aos casos i.e. não há nenhum modelo com mais de um regressor dinâmico).

O segundo objetivo específico desse trabalho é identificar um modelo preditivo dos casos de leptospirose, usando as informações sobre fenômenos meteorológicos disponíveis. Logo, uma vez identificadas as relações entre as séries temporais, o próximo passo é tentar tirar proveito desse fato, usando de alguma forma essa informação para melhorar não só a predição de pontos presentes dos casos (melhor ajuste do modelo) como também melhorar a capacidade de previsão de pontos futuros com o modelo. A técnica escolhida para levar em conta a informação das séries meteorológicas foi a de considerar um sistema linear aberto onde as séries meteorológicas funcionam como entrada no sistema e a de casos a de saída. Esse tipo de modelo é chamado de modelo com regressores dinâmicos (Box & Jenkins 1968; Box et al. 1974; Helfenstein 1991; Box et al. 1994; Helfenstein 1996). O termo dinâmico nesse caso não significa que os coeficientes destes modelos variem no tempo. Significa que existe uma função que transfere linearmente o efeito de uma série sobre a outra, de uma forma dinâmica, e definida de acordo com a função de transferência utilizada.

O ajuste dos modelos com a previsão e a construção dos intervalos de confiança são até satisfatórios, para todos os modelos, embora a princípio não pareça haver uma melhoria do ajuste na presença dos regressores dinâmicos, pelo menos observando-se as Figuras 28 a 36. O ajuste, porém, é de longe mais adequado do que por exemplo um modelo linear para ajustar uma série desse tipo. Certamente a transformação aplicada às séries de casos e pluviosidade tiveram um papel nesse ajuste não muito bom. Por exemplo, os pontos que caíram fora dos ICs nos modelos não caíram fora antes das séries serem transformadas de volta para a sua escala original.

Um outro inconveniente do uso desses modelos para uma série como a de casos é que os ICs podem ter (e têm) muitas vezes limites inferiores menores que zero, já que as séries são ajustadas como se fossem contínuas e não de contagem. Como o menor número de casos possível é zero, a interpretação de um IC com limite inferior negativo fica bastante prejudicado.

Em relação às estimativas de pontos futuros, de uma maneira geral pode-se dizer que a presença dos regressores dinâmicos melhoram a capacidade de previsão dos modelos. O grupo de modelos sazonais apresentou uma superioridade, comparado com os modelos com componente sazonal, como comentado e avaliado pela observação da variação das estimativas pontuais e pelo comportamento dos seus erros-padrão. Dentre todos os regressores, a pluviosidade parece ser a variável que melhora mais a capacidade preditiva, pelo menos no

caso dos modelos com componente sazonal (Figura 38). Apesar de um dos pontos observados cair fora do IC 95%, o modelo SARIMA(1,0,1)x(0,1,1) com a pluviosidade como regressor dinâmico de ordem (2,0,0) pode ser eleito como o de melhor capacidade de predição entre os modelos estudados nesse trabalho. Este modelo pode ser escrito como:

$$Y_t = -0.14 + \frac{-0.15}{(1-0.17B+0.26B^2)}X_t + \frac{(1-0.32B)(1-0.77B^{52})}{(1-0.68B)(1-B^{52})}Z_t,$$

onde  $Y_t$  é a séries de casos,  $X_t$  é a série de pluviosidade,  $Z_t$  é o erro do modelo (nesse caso um ruído branco) e  $B$  é o *backshift operator*.

Isso reforça o fato da causalidade entre a precipitação pluviométrica e o aparecimento de casos de leptospirose nessa população. Reforça porque de fato a correlação que foi identificada anteriormente é capaz de contribuir para melhorar a capacidade de predição dos modelos, que são dois fenômenos ligados, mas ainda assim diferentes. O fato de existir uma correlação não implica necessariamente que o modelo que leve em conta essa variável terá um desempenho melhor que o modelo sem regressores.

Quanto ao terceiro objetivo original do estudo, os resultados não foram frutíferos, pois na verdade a aparente queda observada no número de casos ocorrido na segunda metade da série não foi estatisticamente significativa, no sentido de que as estimativas tinham um erro-padrão muito próximo ou até mais elevado que seus valores. A idéia nesse caso era tentar descobrir se a queda no número de casos (se houvesse) passasse a ser não significativa na presença dos regressores dinâmicos. Se isso acontecesse, poderíamos concluir que na verdade a queda teria sido apenas consequência de condições meteorológicas diversas durante os anos de 1998 e 1999 e não devido às erráticas medidas de desratização tomadas em Salvador durante esse período, o que foi feito de maneira localizada e esporádica. Nesse caso, nem mesmo sem a influência das variáveis foi possível detectar qualquer diminuição significativa.

De qualquer forma, estes modelos sofrem com todas as limitações apontadas, já que para o estudo de fenômenos biológicos no tempo, os modelos de séries temporais clássicos não se adequam totalmente por vários motivos. Em primeiro lugar, foram desenvolvidos para dar conta de séries contínuas e estacionárias em primeira e segunda ordens (Box et al. 1994). Entretanto, grande parte dos estudos de séries temporais em saúde pública lida com séries de contagem, como a de casos de leptospirose. A transformação dessas séries de contagem de

casos em séries de incidências, que seriam séries contínuas, na verdade não afeta a natureza da sua distribuição, já que em geral o que estará acontecendo é a divisão da série por uma constante, a não ser no caso de séries históricas mais longas e agrupadas por exemplo por ano e para uma unidade geográfica maior. É claro que com transformações e aproximações é possível obter-se resultados razoáveis, mas muitas vezes não totalmente satisfatórios.

Uma classe de modelos que tratam séries temporais como variáveis de contagem foi proposta e pode ser encontrada na literatura (Cardinal et al. 1999). Trata-se de uma variante do modelo autorregressivo, chamado INAR (*Integer – Autorregressive*) e foi desenvolvido para ajustar processos autorregressivos com números inteiros e positivos. Esses modelos devem ser melhor estudados para aplicação em epidemiologia.

O segundo problema é que os modelos temporais clássicos consideram a série inteira tendo um comportamento uniforme, que é tratado como um comportamento médio e constante durante todo o tempo de observação. Em se tratando de doenças infecciosas, as séries não se comportam desta maneira. Em uma situação em que a transmissão se dá pessoa-a-pessoa, por exemplo, certamente o aparecimento de casos de uma doença influenciará o aparecimento de outros casos de uma maneira crescente no tempo, i.e. quanto maior o número de casos maior a influência que estes terão no aparecimento de mais casos.

Com a leptospirose que tem um complexo mecanismo de transmissão e manutenção no meio ambiente o efeito de variáveis externas, como a pluviosidade certamente não pode ser considerado constante no tempo, e assumir que existe um comportamento médio pode levar a imprecisões. Por exemplo, imaginemos uma situação em que existe um período relativamente seco numa localidade e digamos que isso afete a transmissibilidade da espiroqueta entre os roedores locais de modo que a prevalência entre eles caia bastante. Mesmo que se siga um período de chuvas intenso, o aparecimento de casos de leptospirose pode não ser elevado, pela ausência da circulação da espiroqueta no meio ambiente. Por outro lado, num período de chuvas constantes, mas não suficientes para alagamentos (que facilitam muito a transmissão) seguida de uma forte chuva pode representar uma explosão de casos, pois a espiroqueta pode estar circulando na população de reservatórios.

Mais recentemente modelos que levam em conta esta variação do efeito têm sido desenvolvidos, e são conhecidos como modelos lineares dinâmicos (Pole et al. 1994; West & Harrison 1997), que não devem ser confundidos com os modelos usados nesse trabalho com

regressores dinâmicos. Nestes modelos, seus coeficientes são escritos como uma função do tempo, o que pode dar conta desses comportamentos não constantes encontrados em doenças infecciosas por exemplo. Os modelos clássicos podem ser considerados como um caso particular destes modelos, quando a sua variação no tempo é nula.

Estes modelos, porém, ainda foram idealizados para variáveis contínuas, que como já disse não correspondem à realidade de séries epidemiológicas. Recentemente, uma nova classe de modelos têm sido usados em aplicações biológicas, os chamados modelos lineares dinâmicos generalizados (Gamerman 1998; Ferreira & Gamerman 1999). Esses modelos combinam as características de serem dinâmicos no tempo com a possibilidade do uso de distribuições da família exponencial, como a distribuição de Poisson, que sem dúvida descrevem melhor esses processos biológicos. Aliás, olhando-se a distribuição da série original de casos na Figura 4, ela segue claramente uma distribuição de Poisson.

A maior dificuldade para o emprego destes novos modelos, principalmente os modelos lineares dinâmicos generalizados é a falta de rotinas computacionais para a sua aplicação. Isto significa que estas rotinas devem ser desenvolvidas, o que implica em programação em alguma linguagem, como o próprio S ou R.

O uso desse tipo de modelo permitirá inclusive o estudo das séries diárias, o que levará a uma melhor precisão das relações entre as séries, além de dispensar transformações que podem levar a distorções e até mesmo a erros de estimação dos parâmetros.

Até agora a discussão se limitou ao uso de modelos, mas com o uso dos mesmos dados ora disponíveis. Uma outra abordagem que deve ser mencionada é o uso de dados espaço-temporais. O que pode acontecer também é que as condições espaciais dos diversos locais que compõem a Região Metropolitana de Salvador tenham um comportamento bastante diferenciado entre si. Sendo assim, uma análise espaço-temporal de casos de leptospirose seria bastante interessante, embora necessite a coleta extra de dados para este tipo de refinamento.

## **Conclusões**

O estudo das relações entre variáveis meteorológicas e a detecção de casos graves de leptospirose em Salvador – BA foi feito com a utilização de modelos de séries temporais clássicos, os modelos de Box-Jenkins (Box et al. 1994). Estes modelos se mostraram bastante satisfatórios para a determinação das correlações cruzadas entre as séries de pluviosidade, umidade e temperaturas e a série de casos, já que uma estrutura bem definida foi identificada após as séries terem recebido o tratamento apropriado, ou seja, a aplicação de técnicas capazes de retirar as estruturas de tendências, sazonalidade e autocorrelações das séries.

A grande questão porém diz respeito ao tipo de modelo que deve ser usado com o objetivo não mais de retirar a estrutura de correlação serial, mas para servir como um modelo para prever a magnitude dos casos graves de leptospirose e seja capaz de avaliar o efeito de intervenções feitas no sentido de diminuir a incidência da doença em uma cidade como Salvador. Apesar dos modelos serem capazes de prever com alguma precisão pontos futuros de casos e de terem um ajuste obviamente superior a um modelo linear por exemplo, os vários problemas apontados sem dúvida limitam a sua aplicação em séries temporais de interesse biológico. Não obstante, uma importante conclusão é que a presença das variáveis meteorológicas de fato melhoram a capacidade de previsão dos modelos e devem ser usadas, seja qual for o modelo utilizado. Modelos mais condizentes com a realidade biológica da série de casos como os modelos lineares dinâmicos generalizados seriam uma ótima alternativa para uma segunda análise desses dados, desde que rotinas computacionais fossem desenvolvidas para tal. Além disso, com a obtenção extra de dados seria possível realizar um estudo espaço-temporal, o que enriqueceria o conhecimento sobre a dinâmica da doença.

Por fim, os modelos utilizados neste estudo não foram capazes de identificar uma queda significativa no número de casos observados em Salvador – BA na segunda metade da série (i.e. a partir de Janeiro de 1988).

## **Referências Bibliográficas**

- Alberdi, J. C., J. Diaz, J. C. Montero & I. Miron (1998). Daily mortality in Madrid community 1986-1992: relationship with meteorological variables. *European Journal of Epidemiology* **14**(6): 571-8.
- Anderson, D. C., D. S. Folland, M. D. Fox, C. M. Patton & A. F. Kaufmann (1978). Leptospirosis: a common-source outbreak due to leptospire of the grippityphosa serogroup. *American Journal of Epidemiology* **107**(6): 538-44.
- Box, G. E. P. & G. M. Jenkins (1968). Some recent advances in forecasting and control (Part I). *Applied Statistics* **17**(2): 91-109.
- Box, G. E. P., G. M. Jenkins & J. F. MacGregor (1974). Some recent advances in forecasting and control (Part II). *Applied Statistics* **23**(2): 158-79.
- Box, G. E. P., G. M. Jenkins & G. C. Reinsel (1994). *Time Series Analysis: Forecasting and Control*. New Jersey, Prentice-Hall, Inc.
- Box, G. E. P. & D. A. Pierce (1970). Distribution of residual autocorrelations in autorregressive-integrated moving average time series models. *Journal of the American Statistical Association* **65**(332): 1509-26.
- Cardinal, M., R. Roy & J. Lambert (1999). On the application of integer-valued time series models for the analysis of disease incidence. *Statistics in Medicine* **18**(15): 2025-39.
- CDC (1997). Outbreak of leptospirosis among white-water rafters--Costa Rica, 1996. *MMWR Morbidity and Mortality Weekly Report* **46**(25): 577-9.
- Chatfield, C. (1996). *The Analysis of Time Series*. London, Chapman & Hall.
- Collares-Pereira, M., H. Korver, W. J. Terpstra, M. Santos-Reis, M. G. Ramalhinho, M. L. Mathias, M. M. Oom, R. Fons, R. Libois & F. Petrucci-Fonseca (1997). First epidemiological data on pathogenic leptospire isolated on the Azorean islands. *European Journal of Epidemiology* **13**(4): 435-41.
- Corwin, A., A. Ryan, W. Bloys, R. Thomas, B. Deniega & D. Watts (1990). A waterborne outbreak of leptospirosis among United States military personnel in Okinawa, Japan. *International Journal of Epidemiology* **19**(3): 743-8.
- Diggle, P. J. (1992). *Time Series: A Biostatistical Introduction*. New York, Oxford University Press Inc.
- Douglin, C. P., C. Jordan, R. Rock, A. Hurley & P. N. Levett (1997). Risk factors for severe leptospirosis in the parish of St. Andrew, Barbados [letter]. *Emerging Infectious Diseases* **3**(1): 78-80.

- Dupont, H., D. Dupont-Perdrizet, J. L. Perie, S. Zehner-Hansen, B. Jarrige & J. B. Daijardin (1997). Leptospirosis: prognostic factors associated with mortality. *Clinical Infectious Diseases* **25**(3): 720-4.
- Everard, C. O., S. Bennett, C. N. Edwards, G. D. Nicholson, T. A. Hassell, D. G. Carrington & J. D. Everard (1992). An investigation of some risk factors for severe leptospirosis on Barbados. *Journal of Tropical Medicine and Hygiene* **95**(1): 13-22.
- Fernandez-Perez, C., J. Tejada & M. Carrasco (1998). Multivariate time series analysis in nosocomial infection surveillance: a case study. *International Journal of Epidemiology* **27**(2): 282-8.
- Ferreira, M. A. R. & D. Gamerman (1999). Análise Bayesiana de séries temporais de contagem via modelos dinâmicos Bayesianos generalizados. *Cadernos de Saúde Coletiva* **6**: 145-155.
- Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalised linear models. *Biometrika* **85**(1): 215-227.
- Hamilton, D. H. & D. G. Watts (1978). Interpreting partial autocorrelation functions of seasonal time series. *Biometrika* **65**(1): 135-40.
- Helfenstein, U. (1986). Box-Jenkins modelling of some viral infectious diseases. *Statistics in Medicine* **5**(1): 37-47.
- Helfenstein, U. (1991). The use of transfer function models, intervention analysis and related time series methods in epidemiology. *International Journal of Epidemiology* **20**(3): 808-15.
- Helfenstein, U. (1996). Box-Jenkins modelling in medical research. *Statistical Methods in Medical Research* **5**(1): 3-22.
- Helfenstein, U., U. Ackermann-Liebrich, C. Braun-Fahrlander & H. U. Wanner (1991). The environmental accident at 'Schweizerhalle' and respiratory diseases in children: a time series analysis. *Statistics in Medicine* **10**(10): 1481-92.
- Ko, A. I., M. Galvao Reis, C. M. Ribeiro Dourado, W. D. Johnson, Jr. & L. W. Riley (1999). Urban epidemic of severe leptospirosis in Brazil. Salvador Leptospirosis Study Group [see comments]. *Lancet* **354**(9181): 820-5.
- Kuriakose, M., C. K. Eapen & R. Paul (1997). Leptospirosis in Kolenchery, Kerala, India: epidemiology, prevalent local serogroups and serovars and a new serovar. *European Journal of Epidemiology* **13**(6): 691-7.
- Ljung, G. M. & G. E. P. Box (1978). On a measure of lack of fit in time series models. *Biometrika* **65**(2): 297-303.
- Marotto, P. C., M. S. Marotto, D. L. Santos, T. N. Souza & A. C. Seguro (1997). Outcome of leptospirosis in children. *American Journal of Tropical Medicine and Hygiene* **56**(3): 307-10.

- Martinez-Schnell, B. & A. Zaidi (1989). Time series analysis of injuries. *Statistics in Medicine* **8**(12): 1497-508.
- Martins, F. S. V. & T. M. P. P. Castiñeiras (1994). IV. Leptospirose. *Doenças Infecciosas: Conduta Diagnóstica e Terapêutica*. M. Schechter and D. V. Marangoni. Rio de Janeiro, Guanabara-Koogan: 98-104.
- MathSoft Data Analysis Products Division (1999). *S-PLUS 2000 Guide to Statistics, Volume 2*. Seattle, WA, Data Analysis Products Division, MathSoft.
- Milner, P. C. (1988). Forecasting the demand on accident and emergency departments in health districts in the Trent region. *Statistics in Medicine* **7**(10): 1061-72.
- Murhekar, M. V., A. P. Sugunan, P. Vijayachari, S. Sharma & S. C. Sehgal (1998). Risk factors in the transmission of leptospiral infection. *Indian Journal of Medical Research* **107**: 218-23.
- Perrocheau, A. & P. Perolat (1997). Epidemiology of leptospirosis in New Caledonia (South Pacific): a one-year survey. *European Journal of Epidemiology* **13**(2): 161-7.
- Pole, A., M. West & J. Harrison (1994). *Applied Bayesian Forecasting and Time Series Analysis*. London, Chapman & Hall.
- Pukkila, T., S. Koreisha & A. Kallinen (1990). The Identification of ARMA Models. *Biometrika* **77**(3): 537-48.
- Saez, M., J. Sunyer, J. Castellsague, C. Murillo & J. M. Anto (1995). Relationship between weather temperature and mortality: a time series analysis approach in Barcelona. *International Journal of Epidemiology* **24**(3): 576-82.
- SAS Institute Inc (1999). *SAS OnlineDoc. CD-ROM, Version 8*. Cary, NC, SAS Institute Inc.
- Schnell, D., A. Zaidi & G. Reynolds (1989). A time series analysis of gonorrhoea surveillance data. *Statistics in Medicine* **8**(3): 343-52; discussion 363.
- Stroup, D. F., S. B. Thacker & J. L. Herndon (1988). Application of multiple time series analysis to the estimation of pneumonia and influenza mortality by age 1962-1983. *Statistics in Medicine* **7**(10): 1045-59.
- Tappero, J. W., D. A. Ashford & B. A. Perkins (2000). *Leptospira Species (Leptospirosis)*. *Mandell: Principles and Practice of Infectious Diseases*. G. L. Mandell, J. E. Bennett and R. Dolin. Philadelphia, Churchill Livingstone, Inc.: 2495-2501.
- Tiao, G. C. & W. S. Wei (1976). Effect of temporal aggregation on the dynamic relation of two time series variables. *Biometrika* **63**(3): 513-23.
- Trejejo, R. T., J. G. Rigau-Perez, D. A. Ashford, E. M. McClure, C. Jarquin-Gonzalez, J. J. Amador, J. O. de los Reyes, A. Gonzalez, S. R. Zaki, W. J. Shieh, R. G. McLean, R. S. Nasci, R. S. Weyant, C. A. Bolin, S. L. Bragg, B. A. Perkins & R. A. Spiegel (1998). Epidemic

- leptospirosis associated with pulmonary hemorrhage-Nicaragua, 1995. *Journal of Infectious Diseases* **178**(5): 1457-63.
- VanBrackle, L. & G. D. Williamson (1999). A study of the average run length characteristics of the National Notifiable Diseases Surveillance System. *Statistics in Medicine* **18**(23): 3309-19.
- Venables, W. N. & B. D. Ripley (1999). *Modern Applied Statistics with S-Plus*. New York, Springer.
- Venables, W. N., D. M. Smith & t. R. D. C. Team (2001). *An Introduction to R*, <http://lib.stat.cmu.edu/R/CRAN/doc/manuals/R-intro.pdf>.
- Watier, L., S. Richardson & B. Hubert (1991). A time series construction of an alert threshold with application to *S. bovis* in France. *Statistics in Medicine* **10**(10): 1493-509.
- West, M. & J. Harrison (1997). *Bayesian Forecasting and Dynamics Models*. New York, Springer-Verlag.
- Williamson, G. D. & W. H. G. (1999). A monitoring system for detecting aberrations in public health surveillance reports. *Statistics in Medicine* **18**(23): 3283-98.
- Zaidi, A. A., D. J. Schnell & G. H. Reynolds (1989). Time series analysis of syphilis surveillance data. *Statistics in Medicine* **8**(3): 353-62; discussion 363.

## **Anexo 1 – Tabelas**

**Tabela 1 – Modelos Seleccionados para Casos, Pluviosidade e Umidade – Estimativas dos Parâmetros, Variância de AIC dos Modelos. Os números entre parênteses são o erro padrão da estimativa**

Estimador	Casos			Pluviosidade			Umidade		
	ARIMA(1,1,1)	SARIMA (1,0,1)x(0,1,1)	SARIMA (1,1,1)x(0,1,1)	ARIMA(3,0,0)	IMA (0,1,1)	SARIMA (1,0,1)x(0,1,1)	ARMA(1,1)	IMA (0,1,1)	SARIMA (0,1,1)x(0,1,1)
$\hat{\alpha}_1$	0.24 (0.11)	0.71 (0.14)	0.21 (0.10)	0.26 (0.07)	-	-0.45 (0.28)	0.82 (0.08)	-	-
$\hat{\alpha}_2$	-	-	-	0.03 (0.07)	-	-	-	-	-
$\hat{\alpha}_3$	-	-	-	0.21 (0.07)	-	-	-	-	-
$\hat{\beta}_1$	0.78 (0.07)	0.40 (0.18)	0.91 (0.04)	-	0.80 (0.04)	-0.64 (0.25)	0.53 (0.12)	0.70 (0.05)	0.88 (0.04)
$\hat{\beta}_2$	-	-	-	-	-	-	-	-	-
$\hat{\Theta}_1$	-	0.63 (0.07)	0.65 (0.06)	-	-	0.79 (0.05)	-	-	0.42 (0.08)
$\hat{\sigma}^2$	0.43	0.56	0.54	1.86	1.93	2.00	11.33	12.06	16.63
AIC	273.01	351.44	346.56	490.85	489.06	544.66	487.16	490.39	516.10

Tabela 2 – Modelos Selecionados para Temperaturas Máxima, Mínima e Média – Estimativas dos Parâmetros, Variância de AIC dos Modelos. Os números entre parênteses são o erro padrão da estimativa

Estimador	Temperatura Máxima			Temperatura Mínima			Temperatura Média		
	ARIMA(2,1,3)	ARIMA (5,1,1)	SARIMA (1,1,2)X(0,1,1)	ARIMA(2,1,3)	ARIMA (5,1,1)	SARIMA (1,1,1)X(0,1,1)	ARIMA(2,1,3)	SARIMA (1,1,2)X(0,1,1)	SARIMA (2,1,2)X(0,1,1)
$\hat{\alpha}_1$	0.13 (0.11)	-0.03 (0.24)	-0.68 (0.14)	0.10 (0.14)	0.27 (0.22)	0.23 (0.11)	0.14 (0.11)	-0.61 (0.14)	-0.61 (0.14)
$\hat{\alpha}_2$	0.65 (0.11)	-0.07 (0.13)	-	0.66 (0.14)	0.04 (0.10)	-	0.63 (0.11)	-	< 0.01 (0.13)
$\hat{\alpha}_3$	-	0.16 (0.09)	-	-	0.09 (0.08)	-	-	-	-
$\hat{\alpha}_4$	-	-0.01 (0.07)	-	-	< 0.01 (0.07)	-	-	-	-
$\hat{\alpha}_5$	-	0.27 (0.07)	-	-	0.21 (0.07)	-	-	-	-
$\hat{\beta}_1$	0.58 (0.10)	0.43 (0.25)	-0.11 (0.12)	0.51 (0.13)	0.64 (0.22)	0.81 (0.07)	0.52 (0.10)	-0.11 (0.12)	-0.11 (0.11)
$\hat{\beta}_2$	0.76 (0.09)	-	-	0.74 (0.13)	-	-	0.76 (0.08)	0.64 (0.07)	0.64 (0.11)
$\hat{\beta}_3$	-0.58 (0.06)	-	-	-0.48 (0.07)	-	-	-0.55 (0.06)	-	-
$\hat{\Theta}_1$	-	-	0.49 (0.07)	-	-	0.62 (0.07)	-	0.50 (0.07)	0.49 (0.07)
$\hat{\sigma}^2$	0.73	0.72	0.95	0.39	0.38	0.45	0.37	0.47	0.47
AIC	358.03	357.27	415.42	272.37	267.23	317.66	256.96	317.88	317.41

**Tabela 3 – Modelos Seleccionados com Ruído ARIMA(1,1,1) com Pluviosidade, Umidade e Temperaturas Máxima, Mínima e Média como Regressores Dinâmicos e 3 Diferentes Funções de Transferência – Estimativas dos Parâmetros, Variância de AIC dos Modelos. Os modelos entre parênteses representam a ordem das funções de transferência  $(r,s,b)$ . Os números entre parênteses no corpo da Tabela são o erro padrão das estimativas. Os modelos seleccionados estão destacados em negro.**

Modelo	Pluviosidade			Umidade			Temperatura Máxima			Temperatura Mínima			Temperatura Média		
	(0,1,1)	(1,0,1)	<b>(2,0,0)</b>	(0,1,1)	(1,0,1)	<b>(2,0,0)</b>	<b>(0,1,1)</b>	(1,0,1)	(2,0,0)	<b>(0,1,1)</b>	(1,0,1)	(2,0,0)	<b>(0,1,1)</b>	(1,0,1)	(2,0,0)
<b>Estimador</b>															
$\hat{\alpha}_1$	0.25 (0.11)	0.22 (0.11)	<b>0.18</b> <b>(0.11)</b>	0.23 (0.11)	0.18 (0.11)	<b>0.20</b> <b>(0.11)</b>	<b>0.26</b> <b>(0.11)</b>	0.21 (0.10)	0.19 (0.12)	<b>0.26</b> <b>(0.11)</b>	0.22 (0.11)	0.21 (0.11)	<b>0.26</b> <b>(0.10)</b>	0.20 (0.10)	0.19 (0.12)
$\hat{\beta}_1$	0.77 (0.07)	0.77 (0.07)	<b>0.73</b> <b>(0.08)</b>	0.77 (0.07)	0.75 (0.08)	<b>0.74</b> <b>(0.08)</b>	<b>0.80</b> <b>(0.07)</b>	0.79 (0.07)	0.73 (0.08)	<b>0.80</b> <b>(0.07)</b>	0.79 (0.07)	0.74 (0.08)	<b>0.08</b> <b>(0.07)</b>	0.79 (0.07)	0.73 (0.08)
$\hat{\omega}_0$	0.05 (0.03)	-0.03 (0.02)	<b>-0.11</b> <b>(0.02)</b>	0.03 (0.01)	0.04 (0.01)	<b>-0.05</b> <b>(0.01)</b>	<b>-0.15</b> <b>(0.04)</b>	-0.17 (0.05)	0.14 (0.04)	<b>-0.19</b> <b>(0.07)</b>	-0.25 (0.07)	0.21 (0.06)	<b>-0.18</b> <b>(0.06)</b>	-0.25 (0.07)	0.21 (0.05)
$\hat{\omega}_1$	-0.1 (0.03)	-	-	-0.04 (0.01)	-	-	<b>0.22</b> <b>(0.04)</b>	-	-	<b>0.25</b> <b>(0.07)</b>	-	-	<b>0.31</b> <b>(0.06)</b>	-	-
$\hat{\delta}_1$	-	-0.72 (0.24)	<b>0.28</b> <b>(0.16)</b>	-	0.71 (0.07)	<b>0.25</b> <b>(0.18)</b>	-	0.69 (0.07)	0.10 (0.08)	-	0.69 (0.08)	0.07 (0.15)	-	0.70 (0.07)	0.09 (0.11)
$\hat{\delta}_2$	-	-	<b>-0.4</b> <b>(0.18)</b>	-	-	<b>-0.34</b> <b>(0.18)</b>	-	-	-0.72 (0.1)	-	-	-0.53 (0.17)	-	-	-0.6 (0.14)
$\hat{\sigma}^2$	0.40	0.43	<b>0.39</b>	0.41	0.41	<b>0.39</b>	<b>0.39</b>	0.41	0.39	<b>0.40</b>	0.41	0.40	<b>0.39</b>	0.41	0.39
AIC	-167.43	-155.05	<b>-174.47</b>	-165.83	-163.86	<b>-174.35</b>	<b>-175.54</b>	-165.14	-175.41	<b>-167.26</b>	-165.38	-168.56	<b>-174.49</b>	-165.39	-172.90

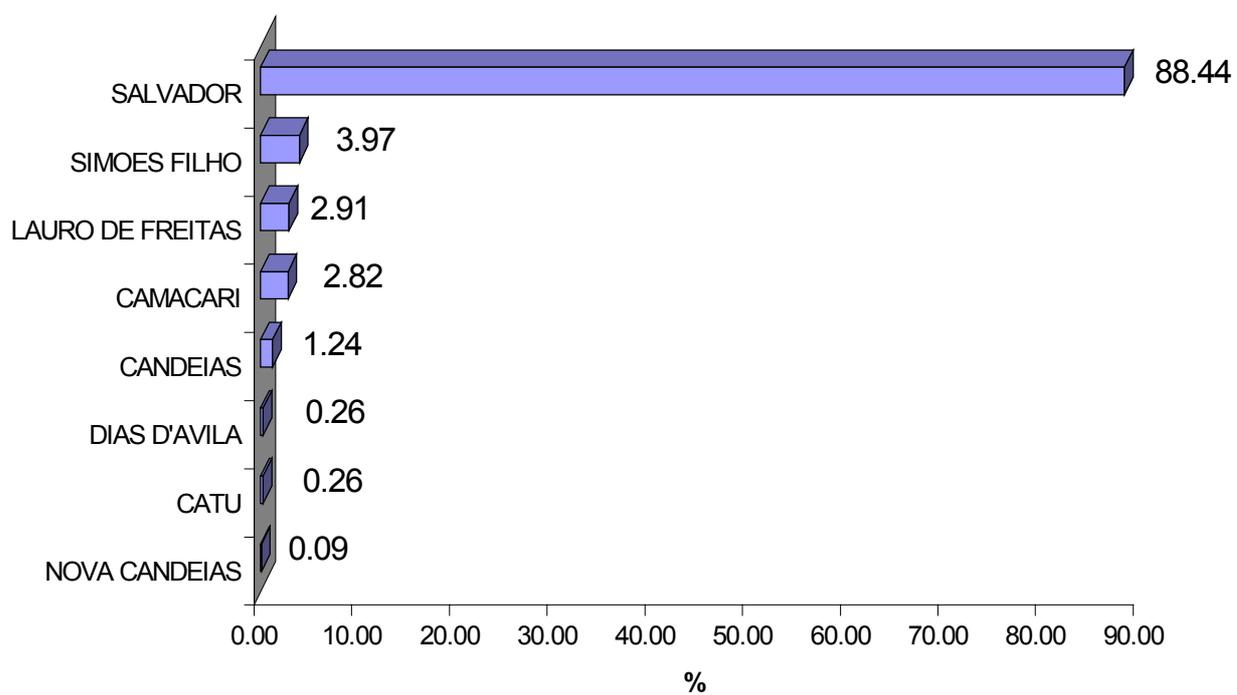
**Tabela 4 – Modelos Seleccionados com Ruído SARIMA (1,0,1)x(0,1,1) com Pluviosidade, Umidade e Temperaturas Máxima, Mínima e Média como Regressores Dinâmicos e 3 Diferentes Funções de Transferência – Estimativas dos Parâmetros, Variância de AIC dos Modelos. Os modelos entre parênteses representam a ordem das funções de transferência  $(r,s,b)$ . Os números entre parênteses no corpo da Tabela são o erro padrão das estimativas. Os modelos seleccionados estão destacados em negro.**

Estimador	Pluviosidade			Umidade			Temperatura Máxima			Temperatura Mínima			Temperatura Média		
	(0,1,1)	(1,0,1)	<b>(2,0,0)</b>	(0,1,1)	(1,0,1)	<b>(2,0,0)</b>	<b>(0,1,1)</b>	(1,0,1)	(2,0,0)	<b>(0,1,1)</b>	(1,0,1)	(2,0,0)	<b>(0,1,1)</b>	(1,0,1)	(2,0,0)
Média	-0.14 (0.08)	-0.18 (0.09)	<b>-0.14</b> <b>(0.07)</b>	-0.14 (0.08)	-0.16 (0.19)	<b>-0.14</b> <b>(0.09)</b>	<b>-0.13</b> <b>(0.07)</b>	-0.18 (0.09)	-0.13 (0.07)	<b>-0.14</b> <b>(0.07)</b>	-0.28 (0.08)	-0.14 (0.07)	<b>-0.13</b> <b>(0.07)</b>	-0.2 (0.08)	-0.14 (0.07)
$\hat{\alpha}_1$	0.60 (0.17)	0.65 (0.16)	<b>0.68</b> <b>(0.14)</b>	0.60 (0.18)	0.95 (0.05)	<b>0.72</b> <b>(0.14)</b>	<b>0.60</b> <b>(0.15)</b>	0.63 (0.17)	0.69 (0.12)	<b>0.57</b> <b>(0.17)</b>	0.3 (0.340)	0.67 (0.14)	<b>0.59</b> <b>(0.16)</b>	0.59 (0.19)	0.68 (0.13)
$\hat{\beta}_1$	0.25 (0.20)	0.32 (0.20)	<b>0.32</b> <b>(0.17)</b>	0.28 (0.22)	0.81 (0.09)	<b>0.40</b> <b>(0.18)</b>	<b>0.26</b> <b>(0.19)</b>	0.32 (0.21)	0.34 (0.15)	<b>0.22</b> <b>(0.20)</b>	0.08 (0.36)	0.33 (0.17)	<b>0.24</b> <b>(0.19)</b>	0.28 (0.23)	0.34 (0.16)
$\hat{\Theta}_1$	0.67 (0.22)	0.62 (0.19)	<b>0.77</b> <b>(0.34)</b>	0.60 (0.19)	0.46 (0.14)	<b>0.50</b> <b>(0.15)</b>	<b>0.83</b> <b>(0.46)</b>	0.59 (0.17)	0.98 (5.64)	<b>0.73</b> <b>(0.27)</b>	0.56 (0.17)	0.80 (0.40)	<b>0.81</b> <b>(0.39)</b>	0.58 (0.17)	0.89 (0.74)
$\hat{\omega}_0$	0.08 (0.03)	0.13 (0.04)	<b>-0.15</b> <b>(0.03)</b>	0.04 (0.01)	0.07 (0.01)	<b>-0.06</b> <b>(0.01)</b>	<b>-0.17</b> <b>(0.04)</b>	-0.24 (0.06)	0.22 (0.04)	<b>-0.24</b> <b>(0.07)</b>	-0.51 (0.08)	0.33 (0.06)	<b>-0.21</b> <b>(0.07)</b>	-0.37 (0.08)	0.34 (0.06)
$\hat{\omega}_1$	-0.11 (0.03)	-	-	-0.04 (0.01)	-	-	<b>0.21</b> <b>(0.04)</b>	-	-	<b>0.24</b> <b>(0.07)</b>	-	-	<b>0.28</b> <b>(0.06)</b>	-	-
$\hat{\delta}_1$	-	0.82 (0.09)	<b>0.17</b> <b>(0.14)</b>	-	0.82 (0.06)	<b>0.18</b> <b>(0.16)</b>	-	0.74 (0.08)	0.12 (0.12)	-	0.94 (0.02)	0.02 (0.14)	-	0.82 (0.05)	0.12 (0.13)
$\hat{\delta}_2$	-	-	<b>-0.26</b> <b>(0.15)</b>	-	-	<b>-0.27</b> <b>(0.16)</b>	-	-	-0.41 (0.13)	-	-	-0.35 (0.14)	-	-	-0.31 (0.14)
$\hat{\sigma}^2$	0.51	0.53	<b>0.42</b>	0.54	0.55	<b>0.51</b>	<b>0.44</b>	0.53	0.35	<b>0.50</b>	0.48	0.42	<b>0.46</b>	0.53	0.39
AIC	-57.15	-57.42	<b>-75.34</b>	-54.99	-63.32	<b>-72.07</b>	<b>-61.47</b>	-60.63	-70.65	<b>-55.20</b>	-76.56	-68.54	<b>-57.69</b>	-62.92	-71.53

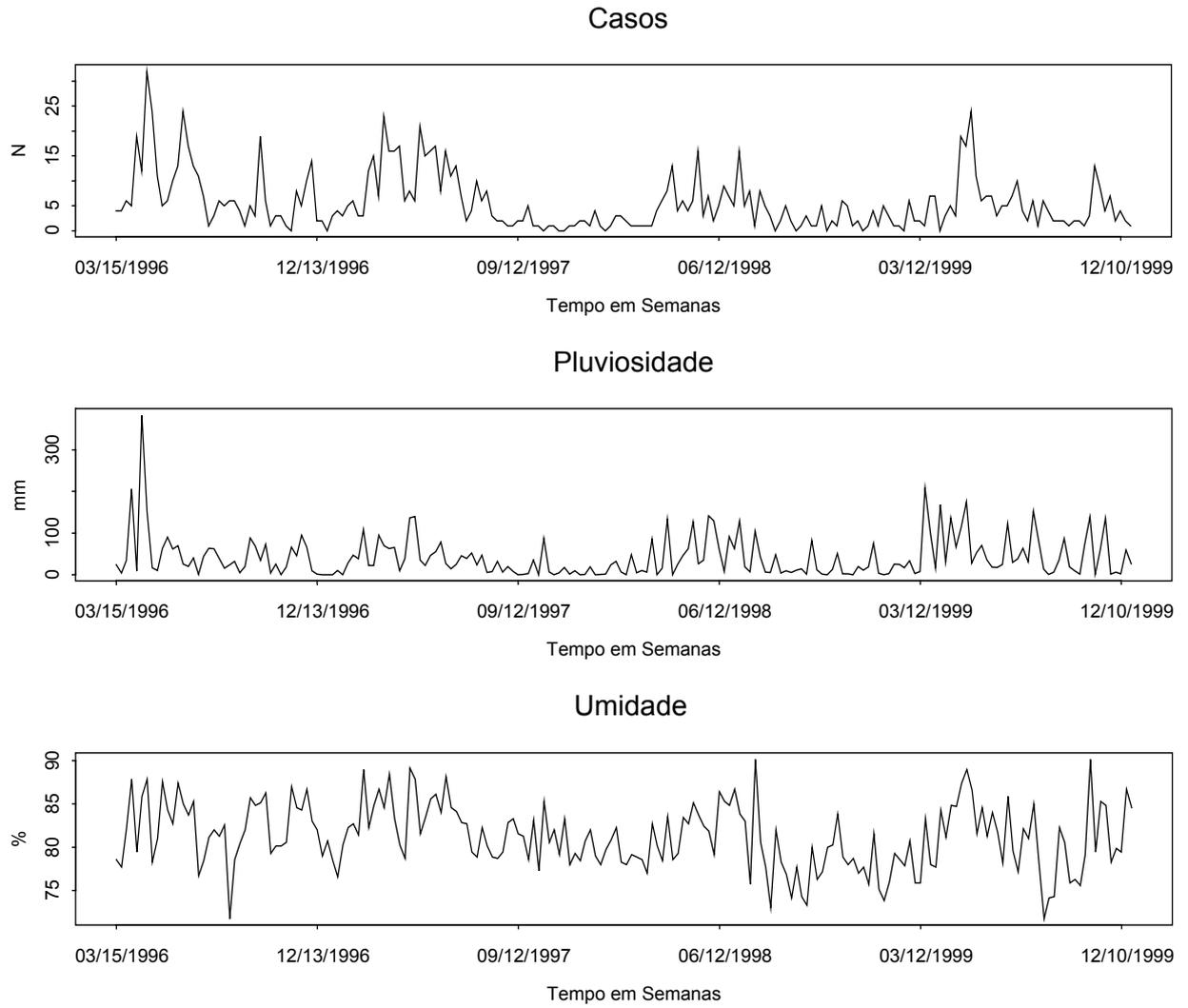
**Tabela 5 – Modelos Seleccionados com Ruído SARIMA (1,1,1)x(0,1,1) com Pluviosidade, Umidade e Temperaturas Máxima, Mínima e Média como Regressores Dinâmicos e 3 Diferentes Funções de Transferência – Estimativas dos Parâmetros, Variância de AIC dos Modelos. Os modelos entre parênteses representam a ordem das funções de transferência  $(r,s,b)$ . Os números entre parênteses no corpo da Tabela são o erro padrão das estimativas. Os modelos seleccionados estão destacados em negrito.**

Estimador	Pluviosidade			Umidade			Temperatura Máxima			Temperatura Mínima			Temperatura Média		
	(0,1,1)	(1,0,1)	<b>(2,0,0)</b>	(0,1,1)	<b>(1,0,1)</b>	(2,0,0)	<b>(0,1,1)</b>	(1,0,1)	(2,0,0)	<b>(0,1,1)</b>	(1,0,1)	(2,0,0)	<b>(0,1,1)</b>	(1,0,1)	(2,0,0)
$\hat{\alpha}_1$	0.26 (0.10)	0.24 (0.10)	<b>0.26</b> <b>(0.09)</b>	0.22 (0.10)	<b>0.22</b> <b>(0.10)</b>	0.10 (0.09)	<b>0.26</b> <b>(0.09)</b>	0.24 (0.10)	0.24 (0.09)	<b>0.26</b> <b>(0.10)</b>	0.23 (0.10)	0.24 (0.09)	<b>0.26</b> <b>(0.09)</b>	0.24 (0.10)	0.24 (0.09)
$\hat{\beta}_1$	0.91 (0.05)	0.91 (0.05)	<b>0.90</b> <b>(0.05)</b>	0.91 (0.05)	<b>0.90</b> <b>(0.05)</b>	0.89 (0.05)	<b>0.91</b> <b>(0.05)</b>	0.91 (0.05)	0.90 (0.05)	<b>0.92</b> <b>(0.05)</b>	0.92 (0.05)	0.90 (0.05)	<b>0.91</b> <b>(0.05)</b>	0.91 (0.05)	0.90 (0.05)
$\hat{\Theta}_1$	0.64 (0.02)	0.59 (0.18)	<b>0.72</b> <b>(0.27)</b>	0.58 (0.17)	<b>0.50</b> <b>(0.15)</b>	0.58 (0.18)	<b>0.75</b> <b>(0.29)</b>	0.57 (0.17)	0.80 (0.39)	<b>0.67</b> <b>(0.22)</b>	0.54 (0.16)	0.72 (0.28)	<b>0.74</b> <b>(0.27)</b>	0.56 (0.17)	0.78 (0.34)
$\hat{\omega}_0$	0.08 (0.03)	0.10 (0.03)	<b>-0.15</b> <b>(0.03)</b>	0.04 (0.01)	<b>0.05</b> <b>(0.01)</b>	0.08 (0.01)	<b>-0.17</b> <b>(0.05)</b>	-0.18 (0.05)	0.23 (0.04)	<b>-0.24</b> <b>(0.07)</b>	-0.34 (0.08)	0.33 (0.06)	<b>-0.21</b> <b>(0.07)</b>	-0.26 (0.08)	0.35 (0.06)
$\hat{\omega}_1$	-0.11 (0.03)	-	-	-0.04 (0.01)	-	-	<b>0.21</b> <b>(0.05)</b>	-	-	<b>0.23</b> <b>(0.07)</b>	-	-	<b>0.27</b> <b>(0.07)</b>	-	-
$\hat{\delta}_1$	-	0.61 (0.09)	<b>0.19</b> <b>(0.15)</b>	-	<b>0.57</b> <b>(0.07)</b>	0.72 (0.13)	-	0.53 (0.08)	0.14 (0.14)	-	0.70 (0.06)	0.03 (0.15)	-	0.58 (0.08)	0.15 (0.15)
$\hat{\delta}_2$	-	-	<b>-0.25</b> <b>(0.15)</b>	-	-	-0.35 (0.12)	-	-	-0.36 (0.15)	-	-	-0.33 (0.15)	-	-	-0.28 (0.15)
$\hat{\sigma}^2$	0.52	0.55	<b>0.43</b>	0.55	<b>0.56</b>	0.44	<b>0.47</b>	0.55	0.42	<b>0.51</b>	0.54	0.45	<b>0.48</b>	0.56	0.42
AIC	-56.60	-54.17	<b>-75.39</b>	-54.98	<b>-57.58</b>	-85.88	<b>-60.99</b>	-56.78	-69.81	<b>-54.84</b>	-60.96	-69.11	<b>-57.05</b>	-56.43	-72.18

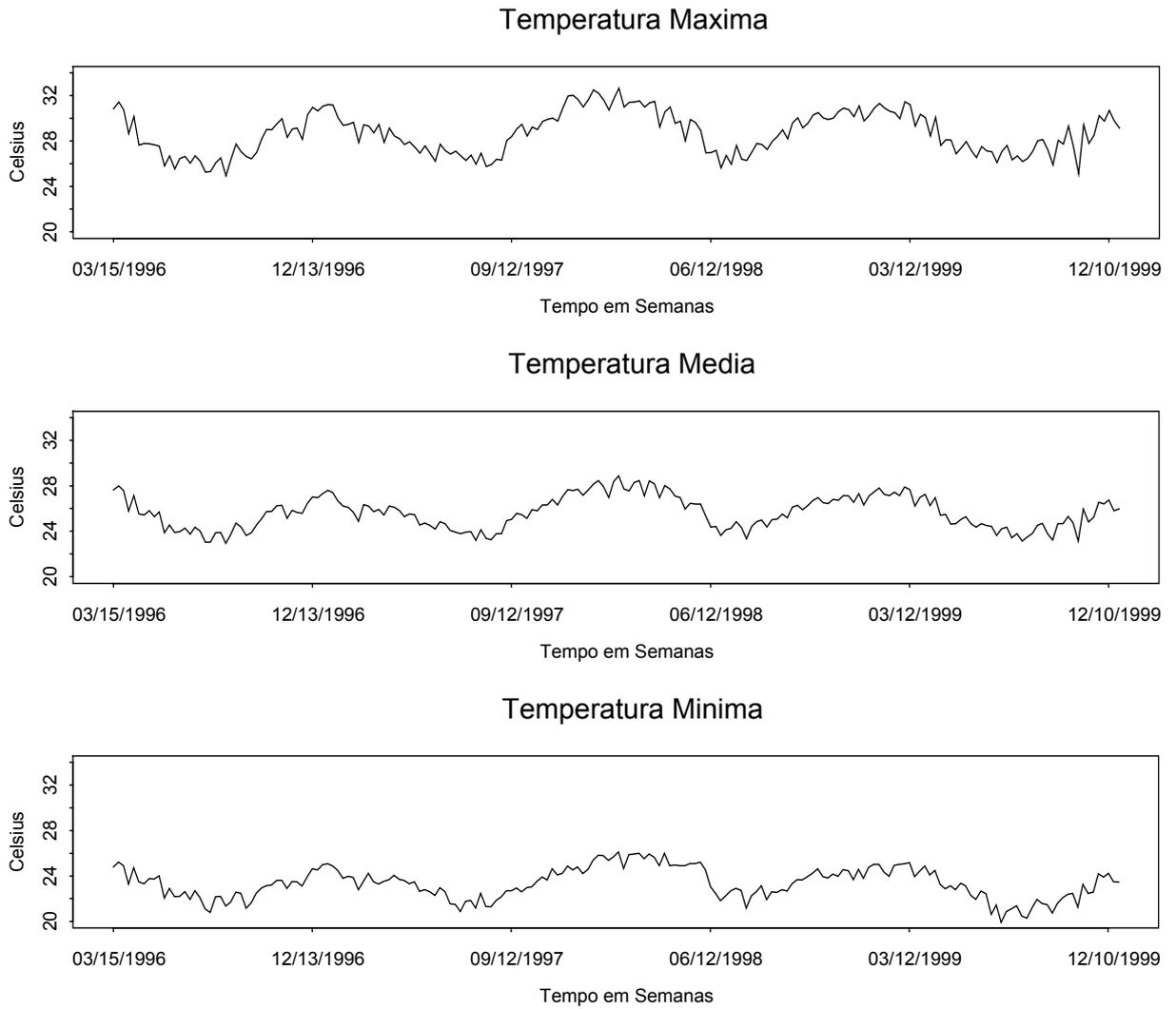
## **Anexo 2 – Figuras**



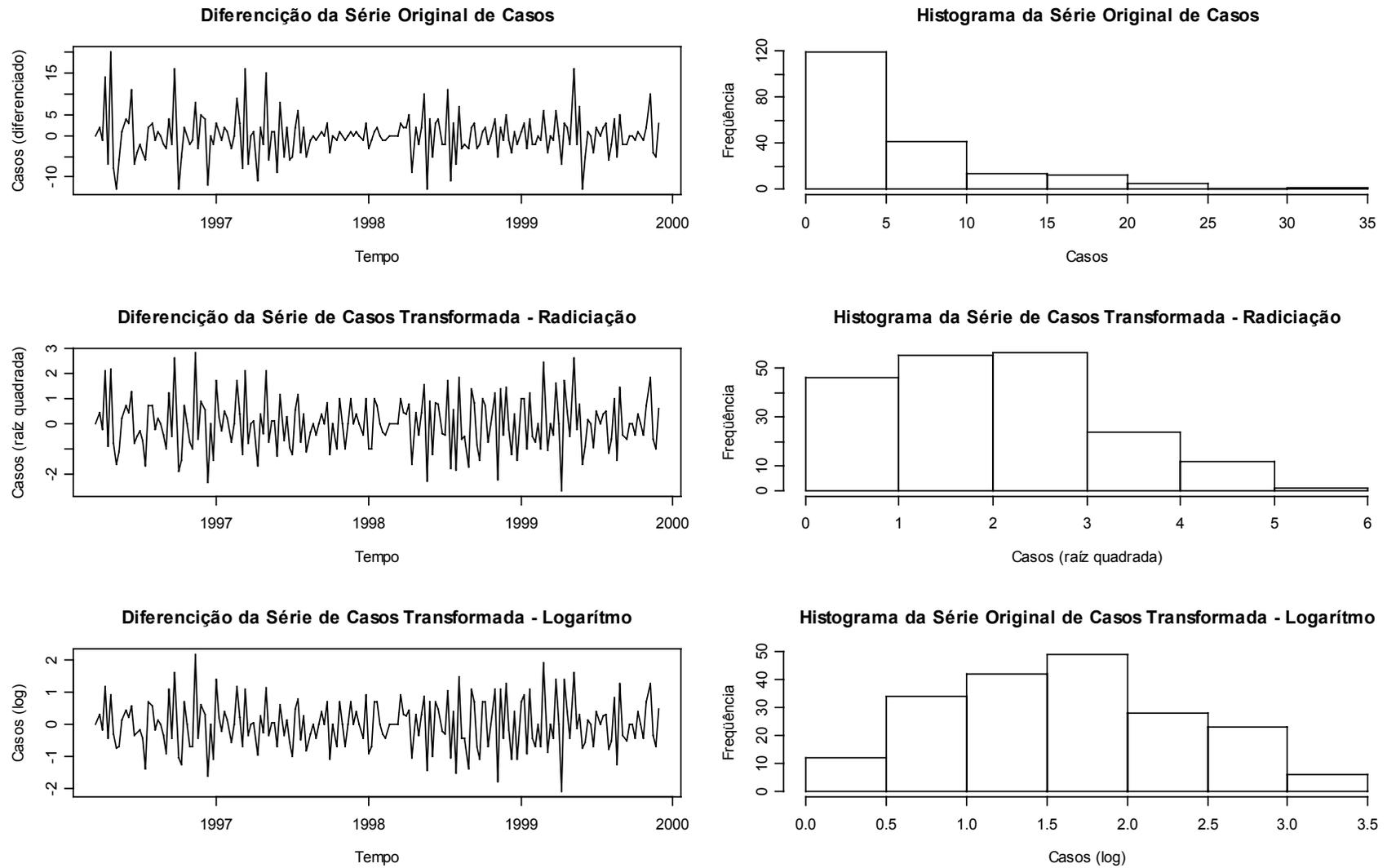
**Figura 1 – Percentual da Procedência dos Casos de Leptospirose**



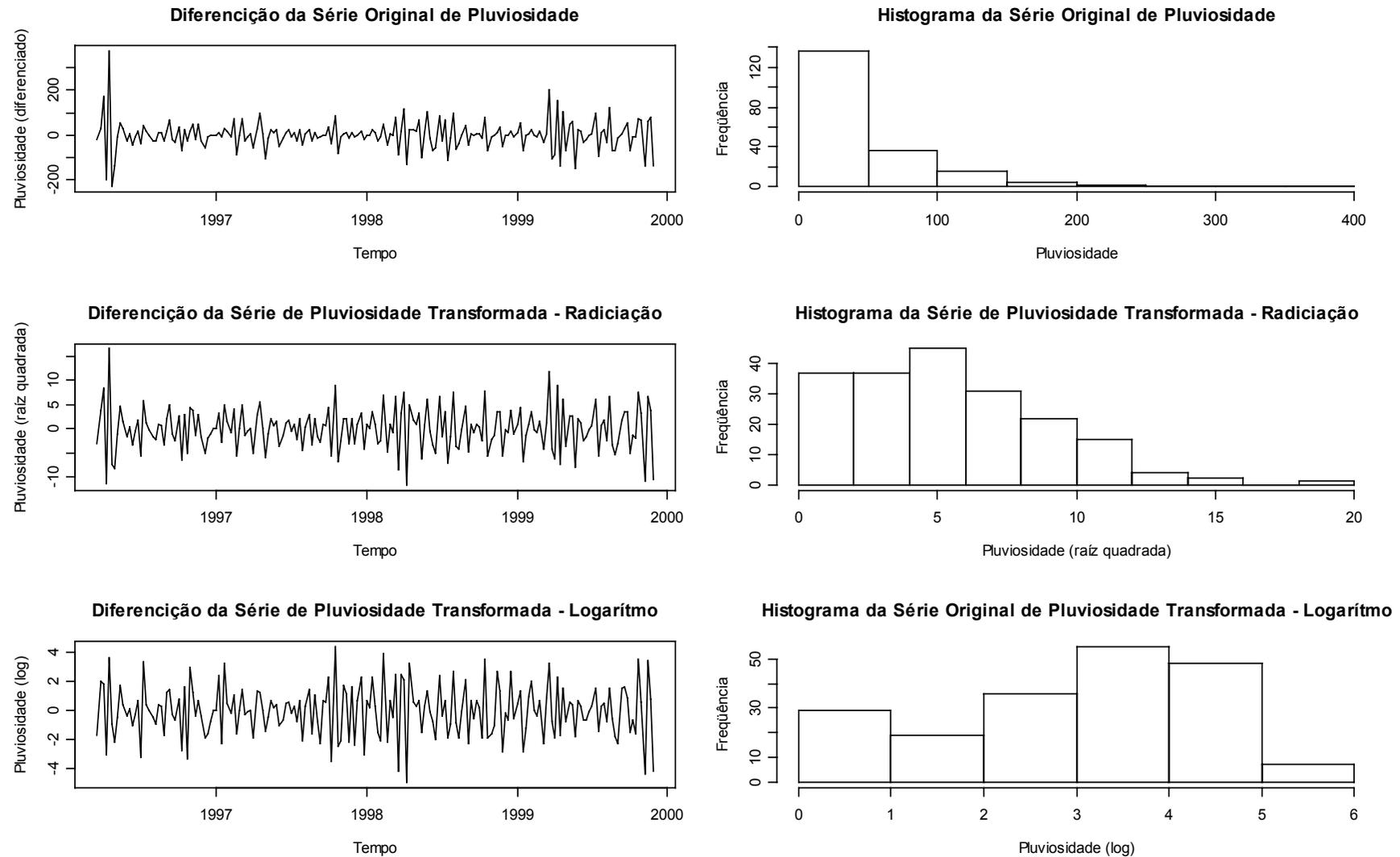
**Figura 2 - Gráficos Temporais das Séries Calendário de Casos, Pluviosidade e Umidade**



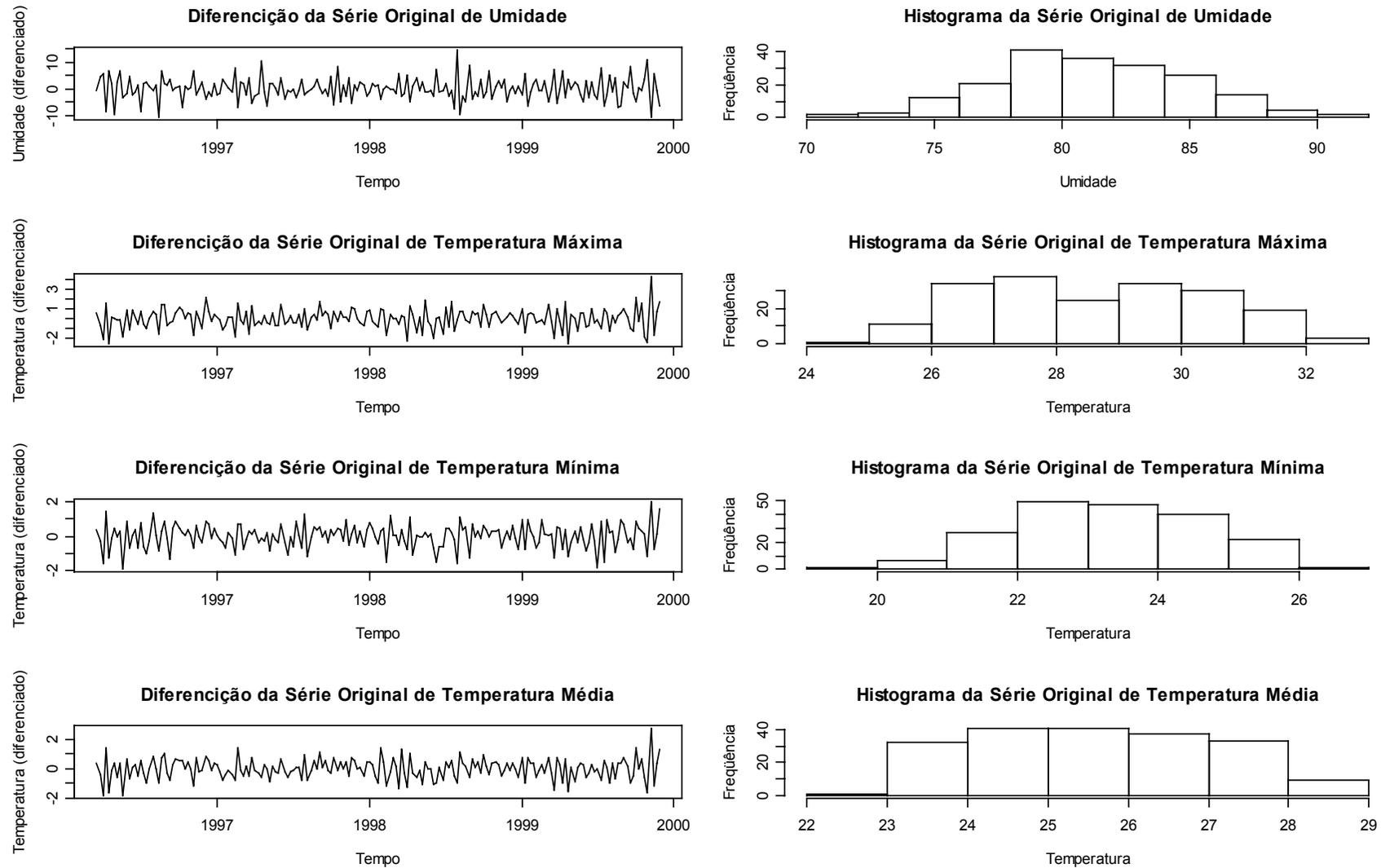
**Figura 3 - Gráficos Temporais das Séries calendário de Temperaturas Máxima, Média e Mínima**



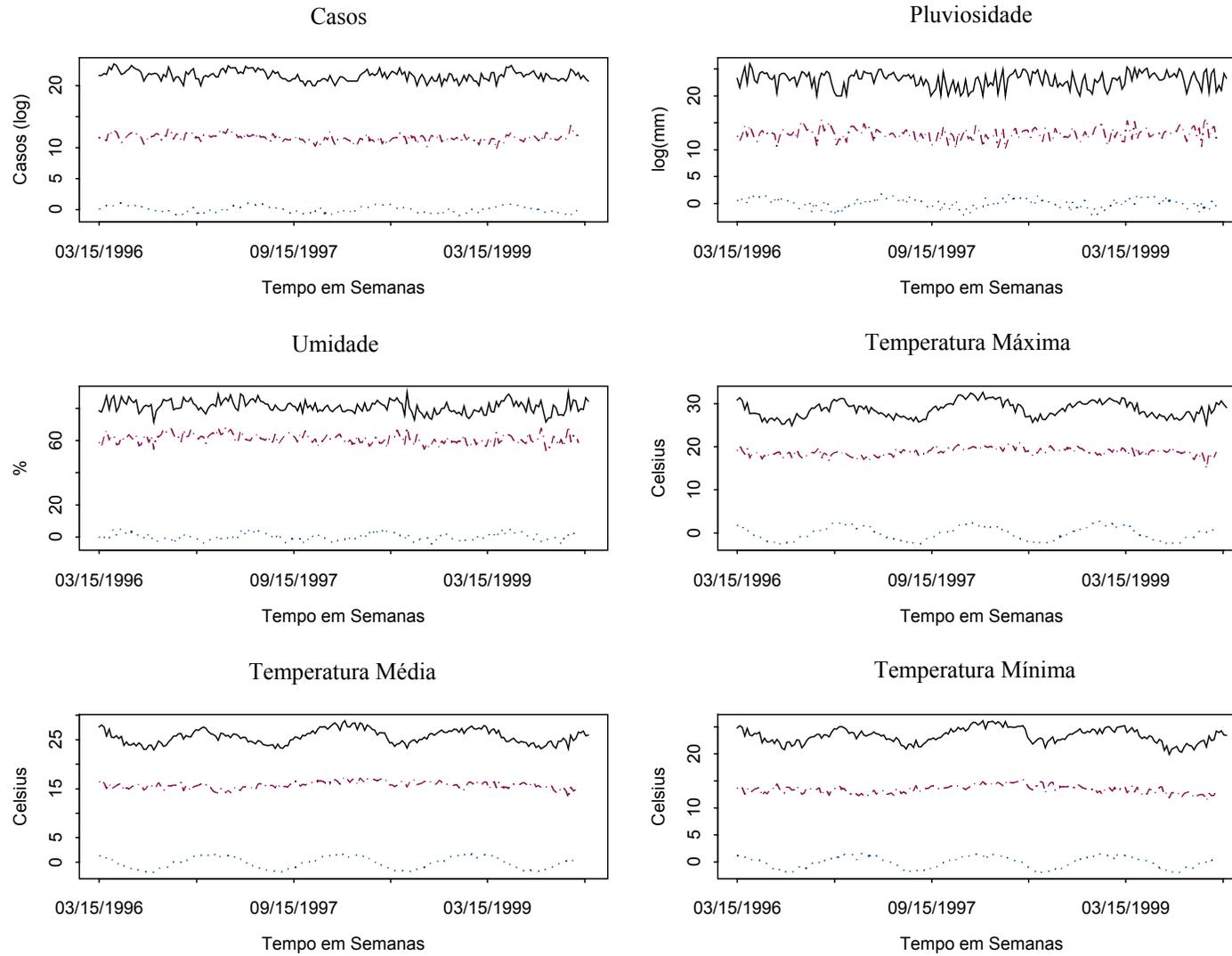
**Figura 4 – Série de Casos – Gráficos Temporais da Série Diferenciada e Histogramas Comparando a Série Original e as Transformações Raiz Quadrada e Logarítmica Deslocada de 1 Unidade.**



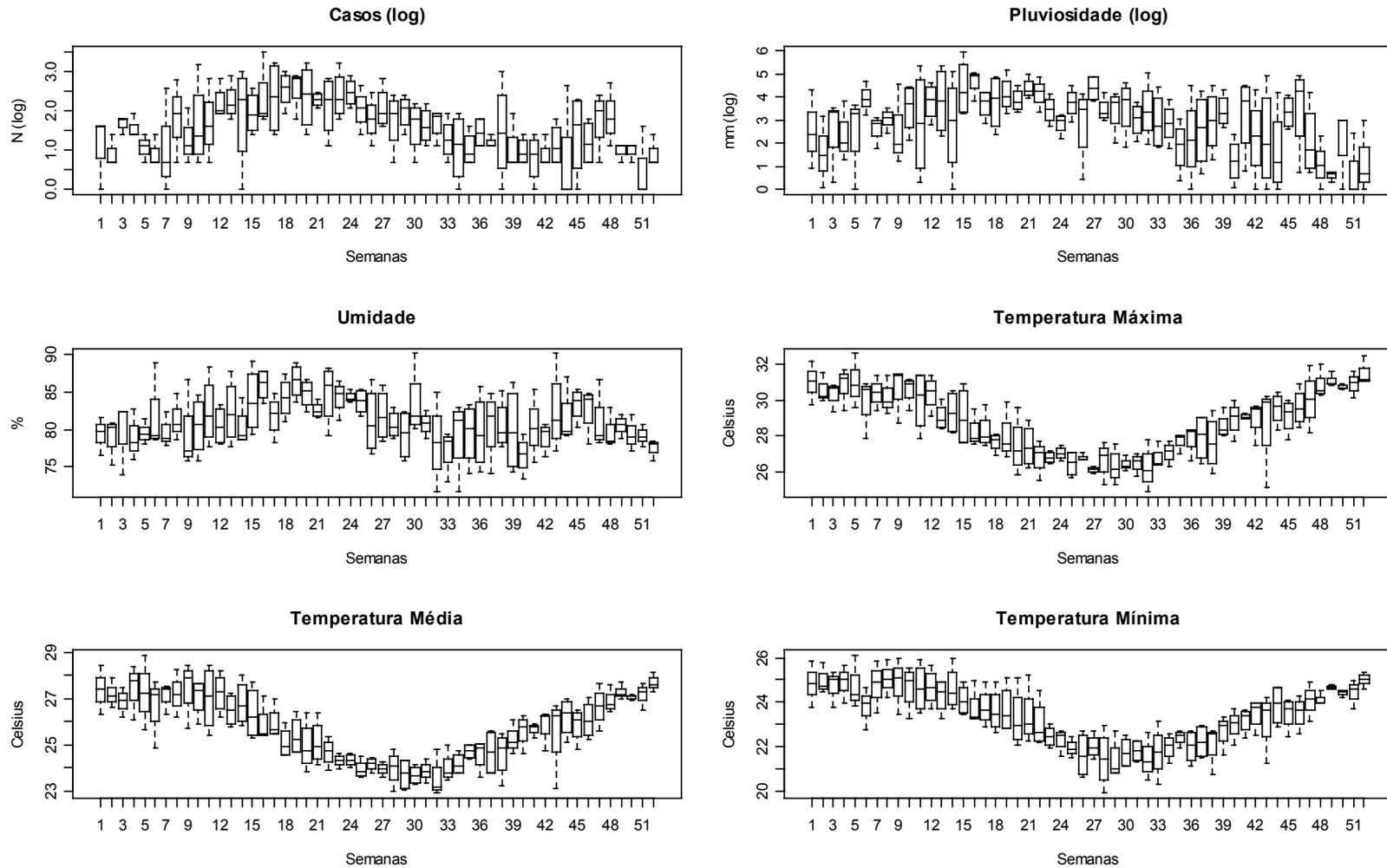
**Figura 5 – Série de Pluviosidade – Gráficos Temporais da Série Diferenciada e Histogramas Comparando a Série Original e as Transformações Raiz Quadrada e Logarítmica com Deslocamento de 1 Unidade.**



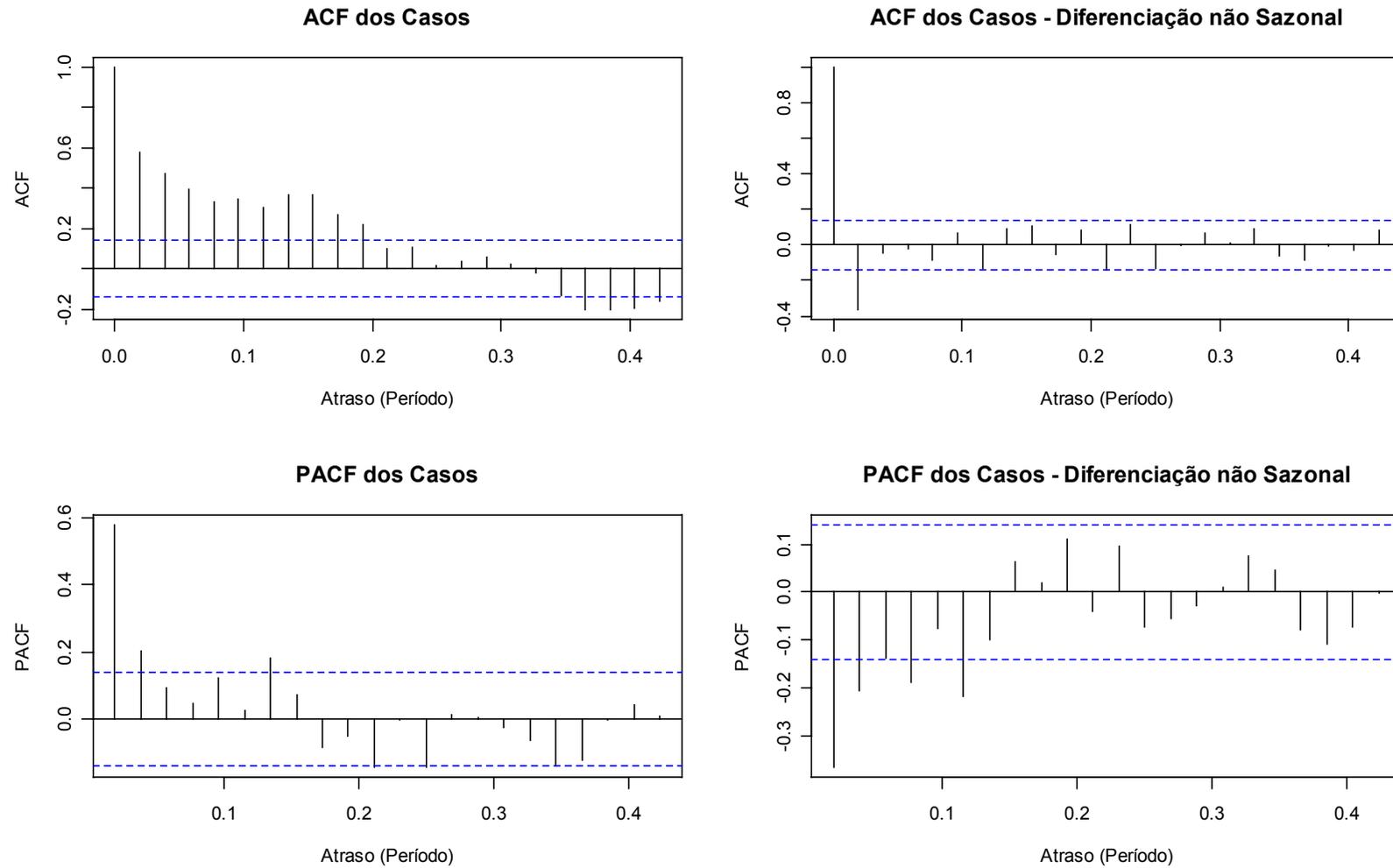
**Figura 6 – Gráficos Temporais das Séries Diferenciadas e Histogramas das Séries de Umidade e Temperaturas Máxima, Mínima e Média**



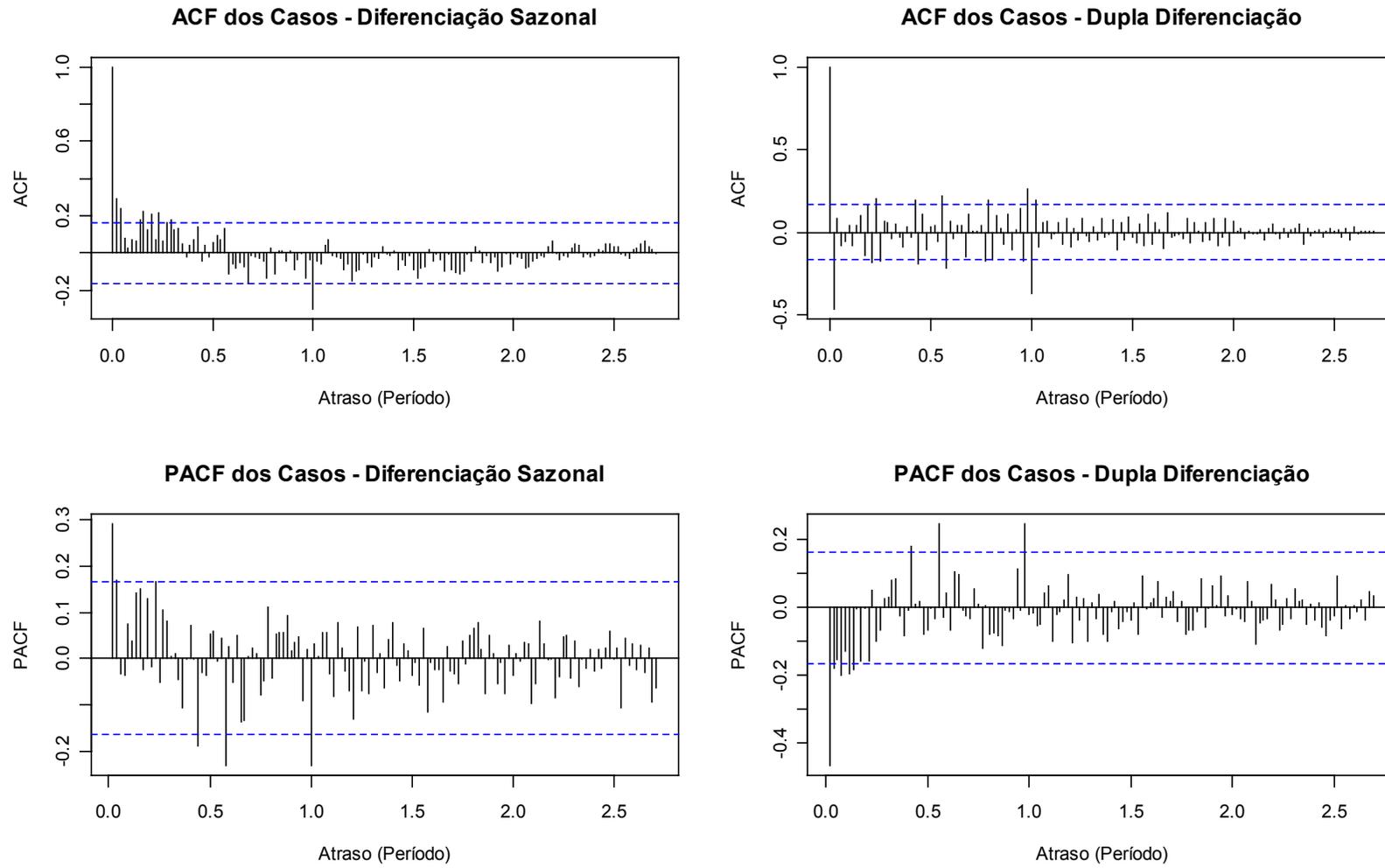
**Figura 7 – Gráficos da Decomposição Sazonal via LOESS das Séries Temporais de Casos, Pluviosidade, Umidade e Temperaturas Máxima, Média e Mínima. A curva superior é a série original e foi deslocada +20 unidades para casos e pluviosidade. A curva do meio são os resíduos e foram deslocadas +10 para casos e pluviosidade, -20 para Umidade e -10 para as 3 temperaturas. Curva inferior é a sazonalidade não deslocada.**



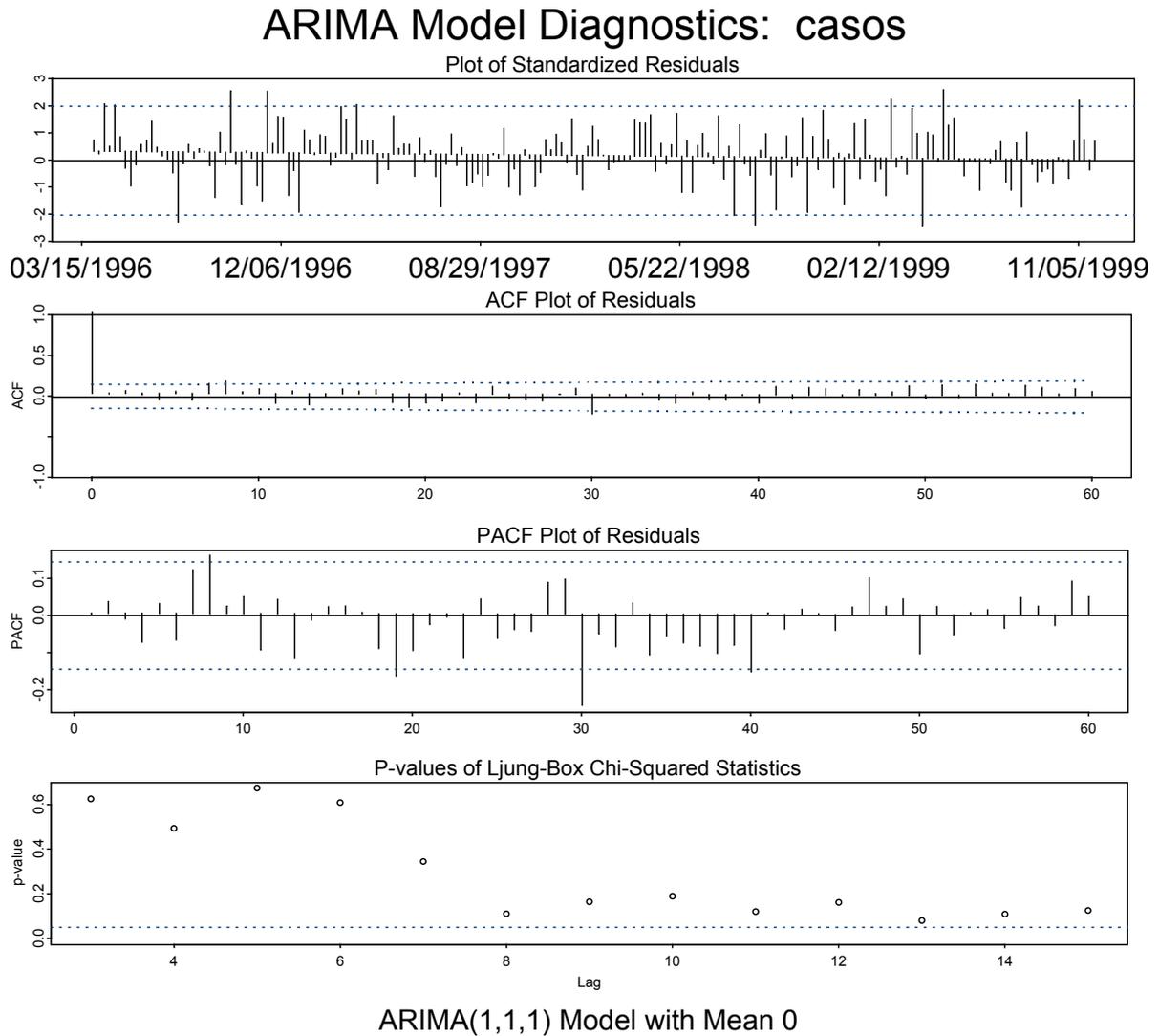
**Figura 8 – Box-Plots Sazonais das Séries de Casos, Pluviosidade, Umidade e Temperaturas Máxima, Média e Mínima. Apesar das séries começarem em Março, a semana número 1 neste gráfico corresponde à primeira semana do ano.**



**Figura 9 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Casos e sua Diferenciação não Sazonal.**

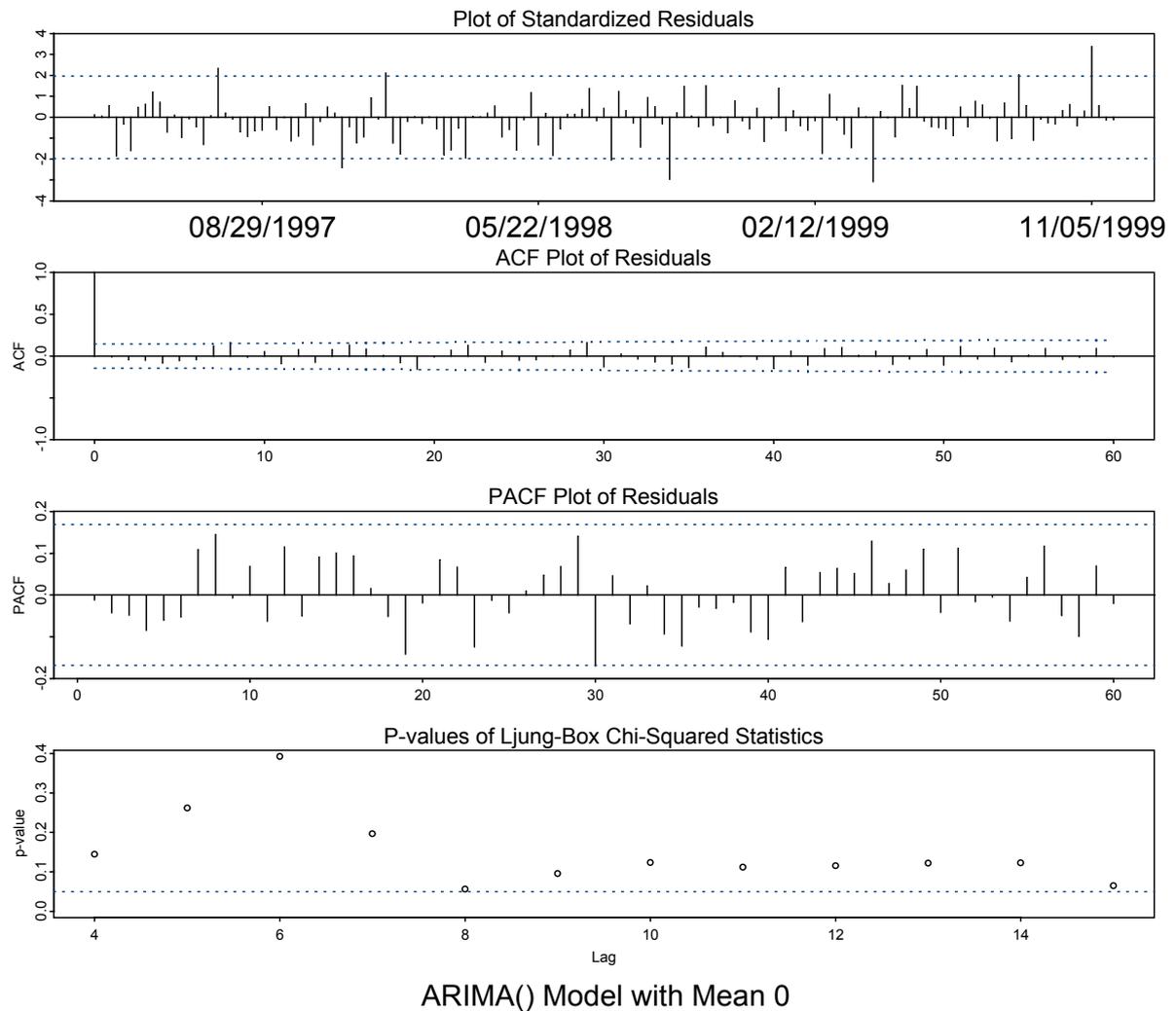


**Figura 10 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Casos com Diferenciação Sazonal e com Diferenciação Sazonal e não sazonal.**



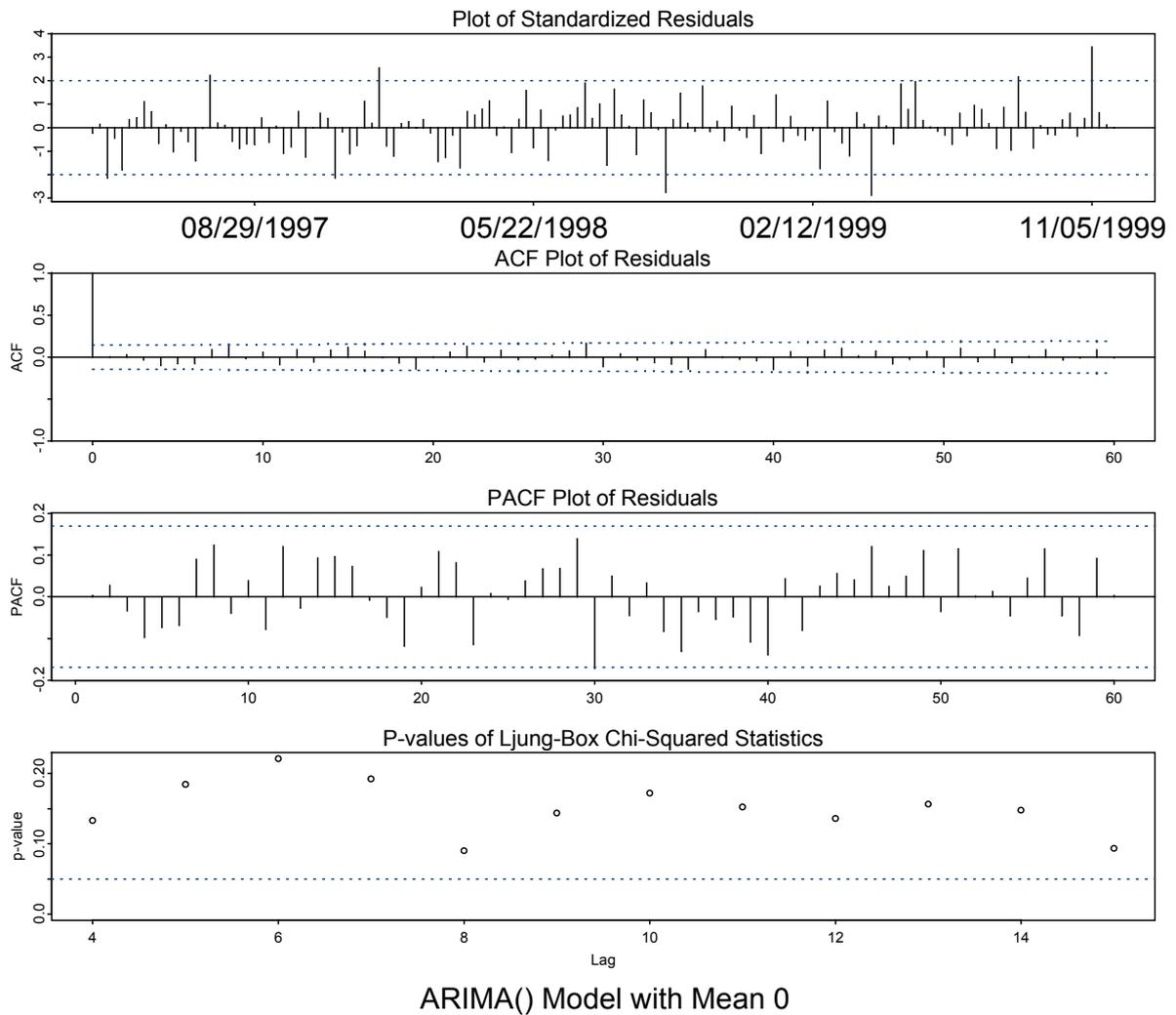
**Figura 11 – Saída do Programa S-Plus Mostrando os Gráficos de Diagnóstico de Ajuste para a Série de Casos – Modelo ARIMA (1,1,1) – (a) Gráfico Temporal dos Resíduos Padronizados; (b) Correlograma da ACF dos Resíduos; (c) Correlograma da PACF dos Resíduos; (d) Teste de Portmanteau para Autocorrelação dos Resíduos.**

## ARIMA Model Diagnostics: casos - mean(casos)

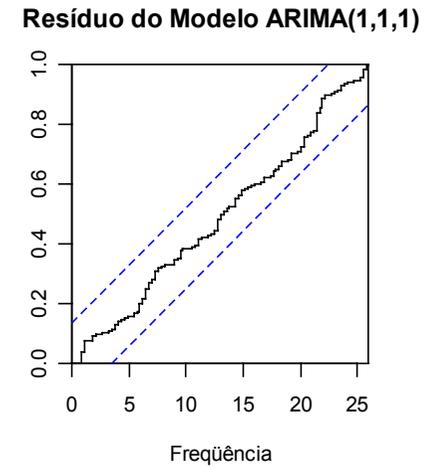
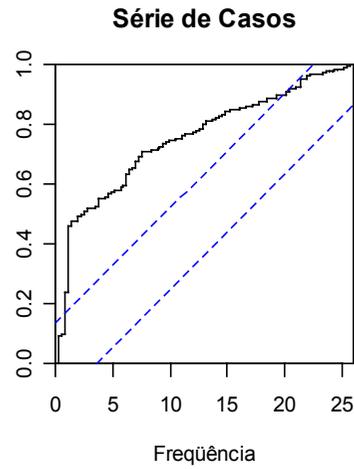


**Figura 12 – Saída do Programa S-Plus Mostrando os Gráficos de Diagnóstico de Ajuste para a Série de Casos – Modelo SARIMA (1,0,1)x(0,1,1) – (a) Gráfico Temporal dos Resíduos Padronizados; (b) Correlograma da ACF dos Resíduos; (c) Correlograma da PACF dos Resíduos; (d) Teste de Portmanteau para Autocorrelação dos Resíduos.**

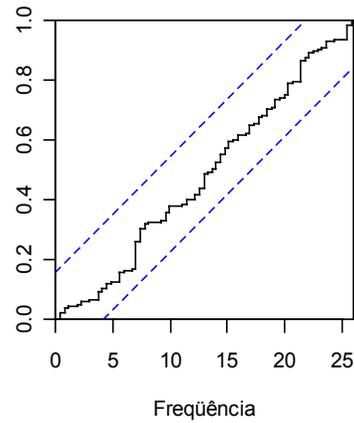
## ARIMA Model Diagnostics: casos



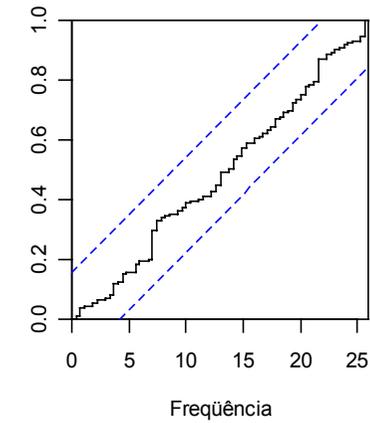
**Figura 13 – Saída do programa S-Plus Mostrando os Gráficos de Diagnóstico de Ajuste para a Série de Casos – Modelo SARIMA (1,1,1)x(0,1,1) – (a) Gráfico Temporal dos Resíduos Padronizados; (b) Correlograma da ACF dos Resíduos; (c) Correlograma da PACF dos Resíduos; (d) Teste de Portmanteau para Autocorrelação dos Resíduos.**



**Resíduo do Modelo SARIMA(1,0,1)x(0,1,1)**



**Resíduo do Modelo SARIMA(1,1,1)x(0,1,1)**



**Figura 14 –Periodogramas Acumulados e IC 95% para a Série de Casos e Resíduos dos modelos ARIMA (1,1,1), SARIMA (1,0,1)x(0,1,1) e SARIMA (1,1,1)x(0,1,1).**

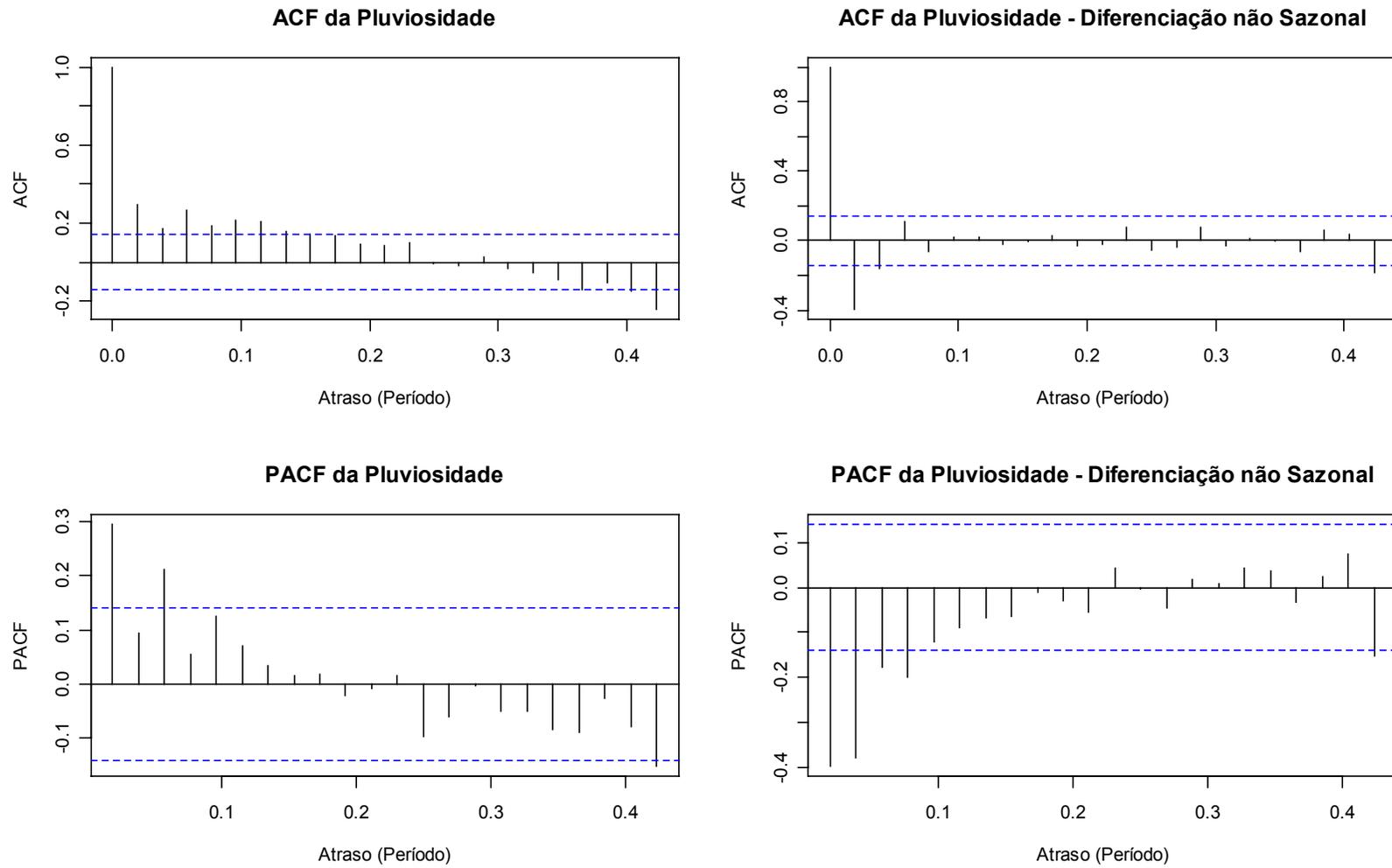
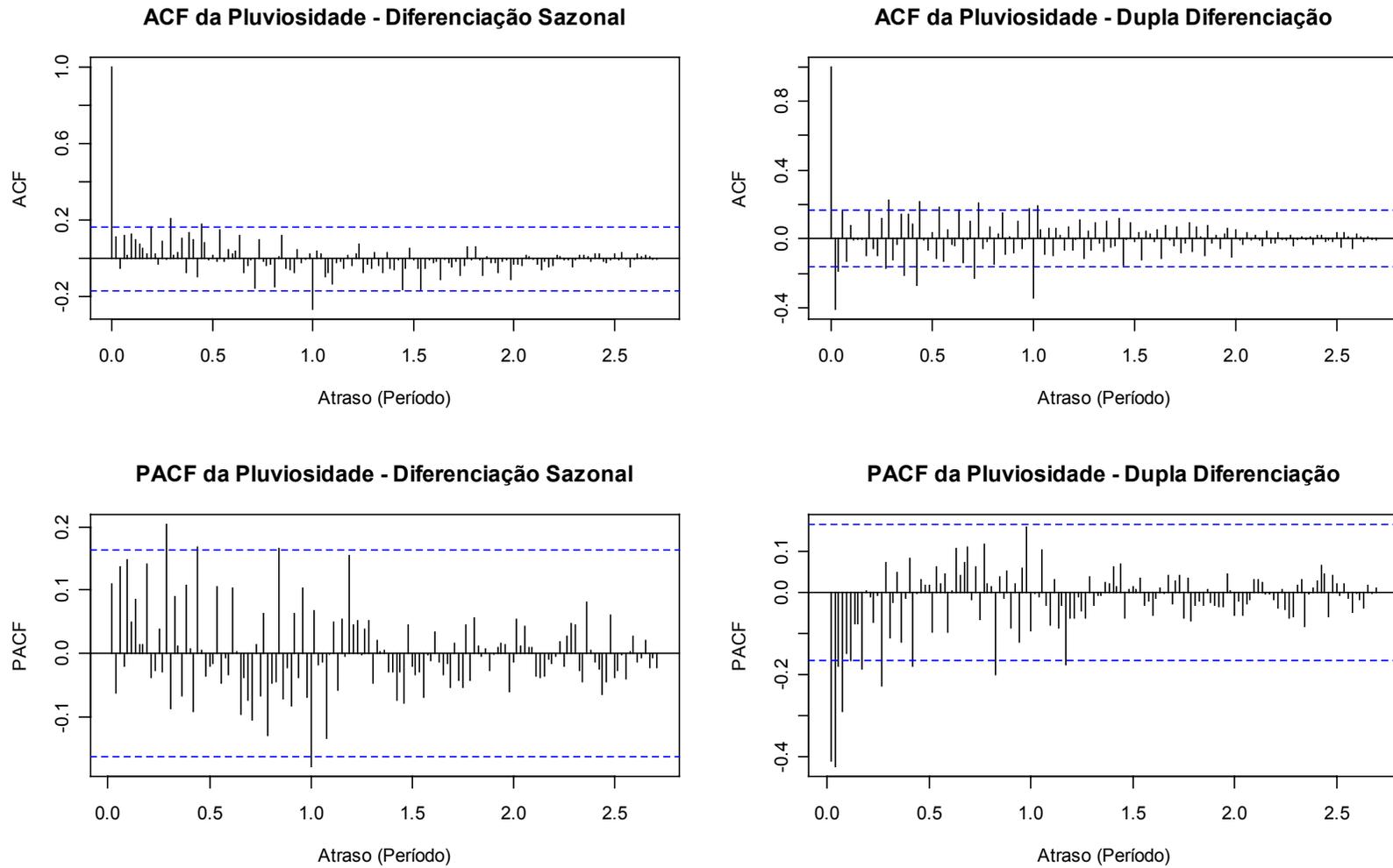


Figura 15 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Pluviosidade e sua Diferenciação não Sazonal.



**Figura 16 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Pluviosidade com Diferenciação Sazonal e com Diferenciação Sazonal e não sazonal.**

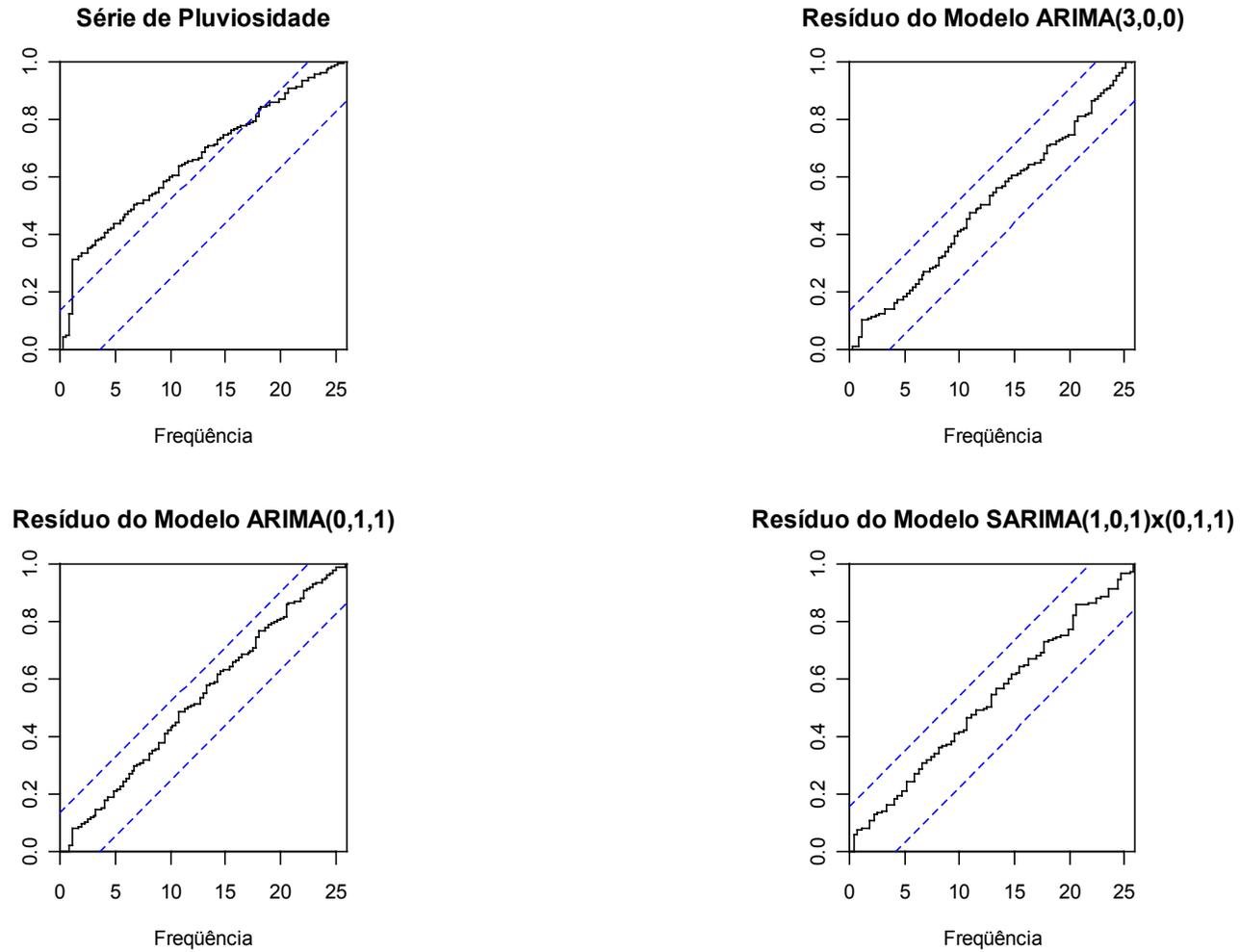


Figura 17 –Periodogramas Acumulados e IC 95% para a Série de Pluviosidade e Resíduos dos modelos ARIMA (3,0,1), ARIMA (0,1,1) e SARIMA (1,0,1)x(0,1,1).

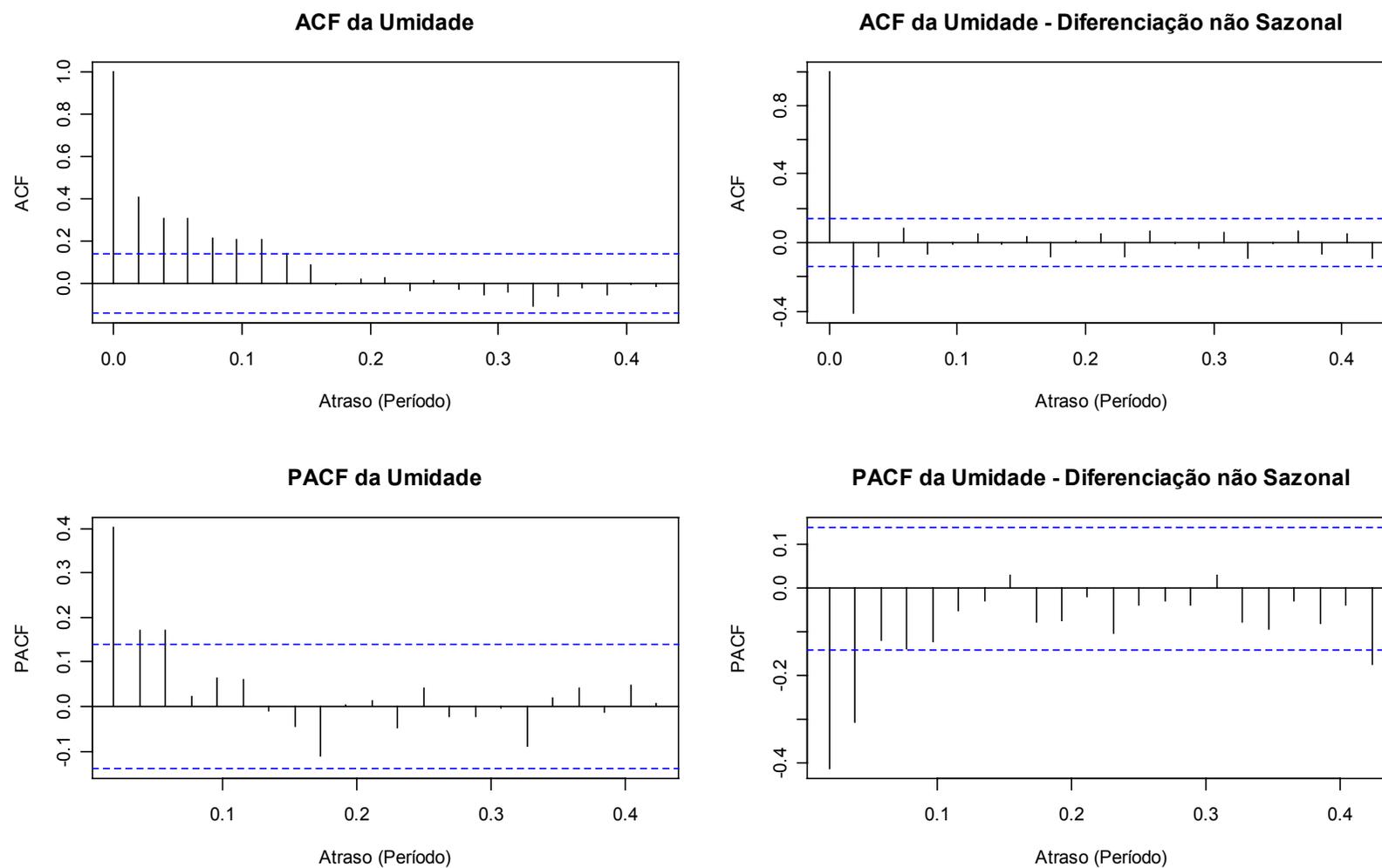
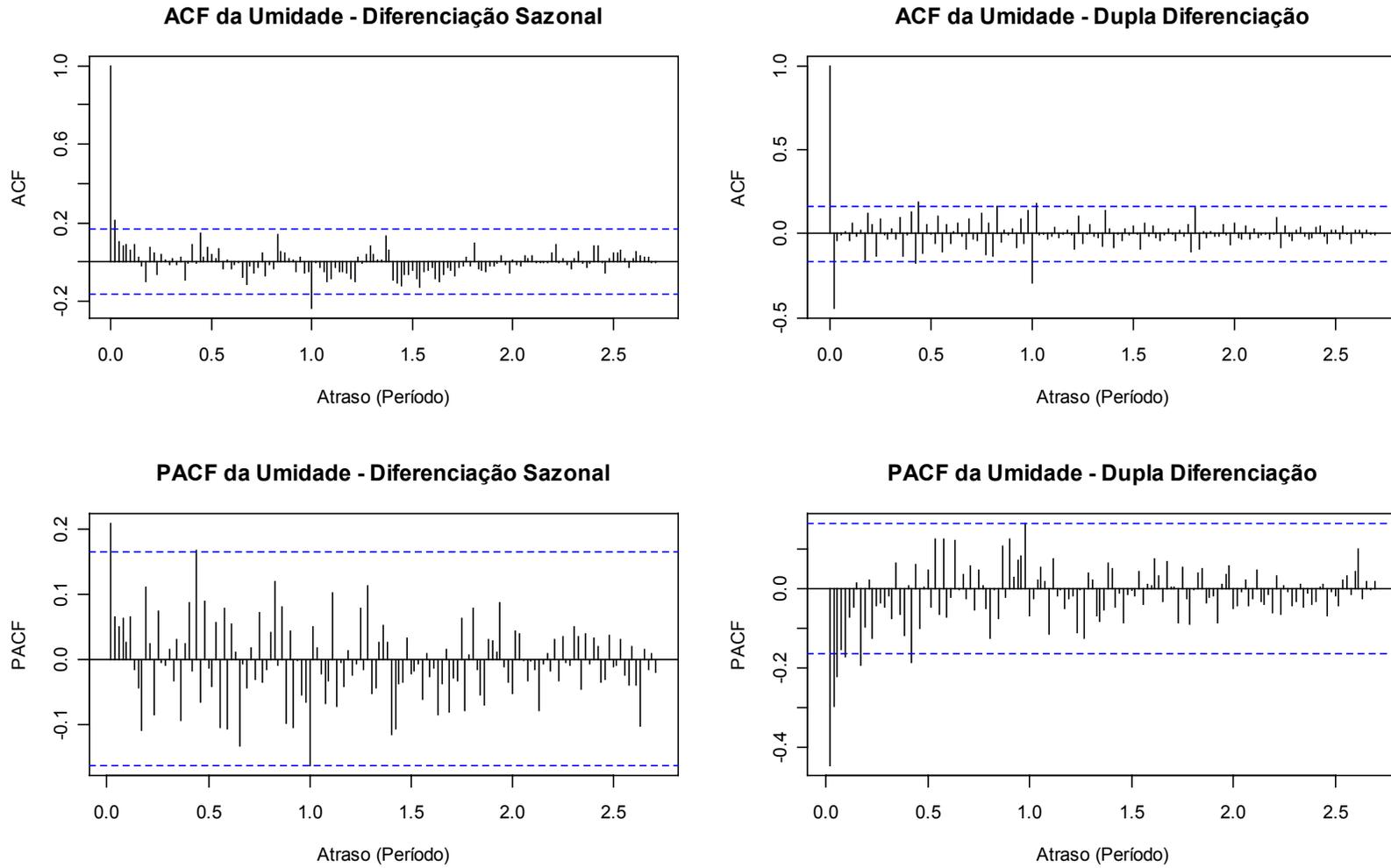
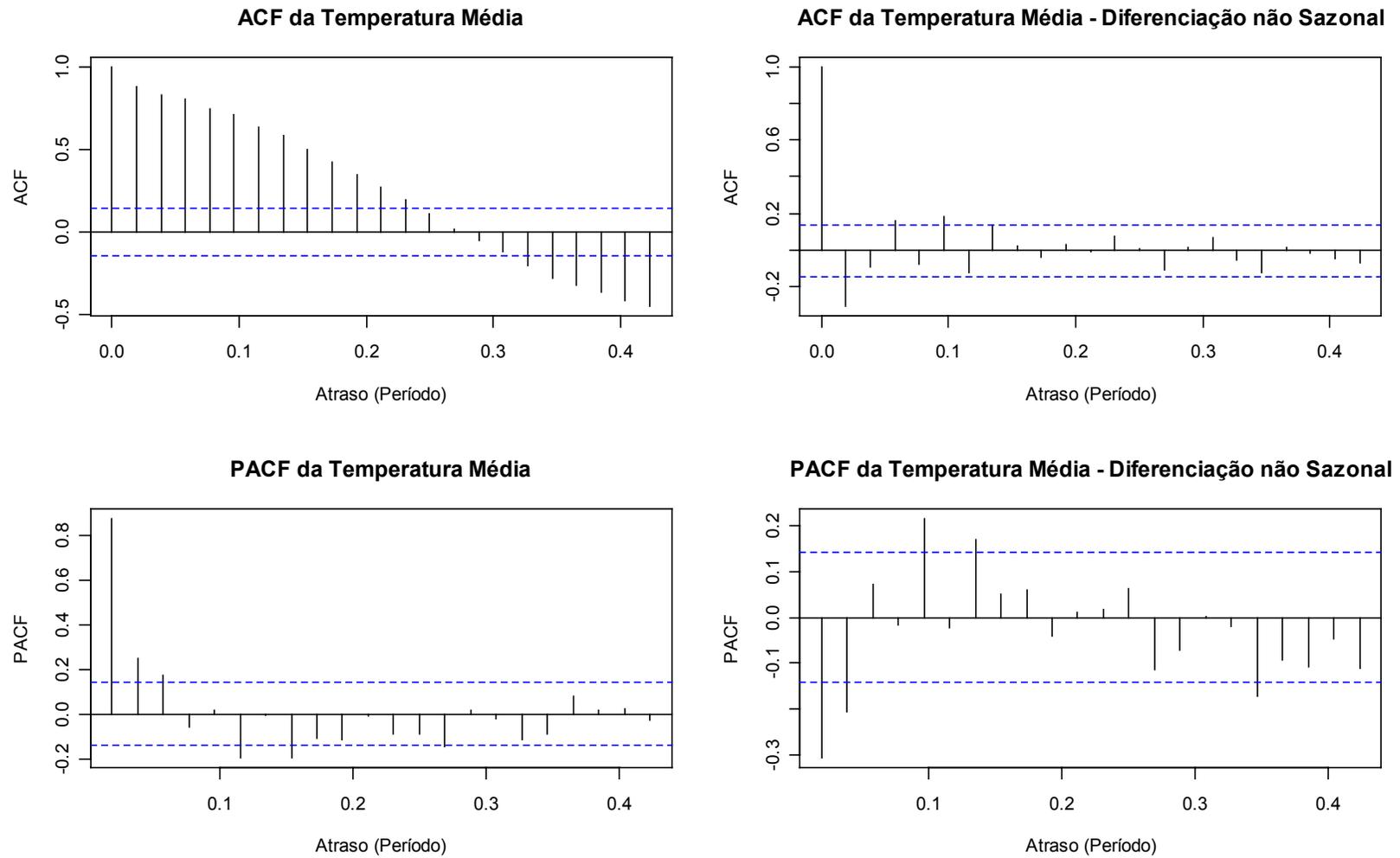


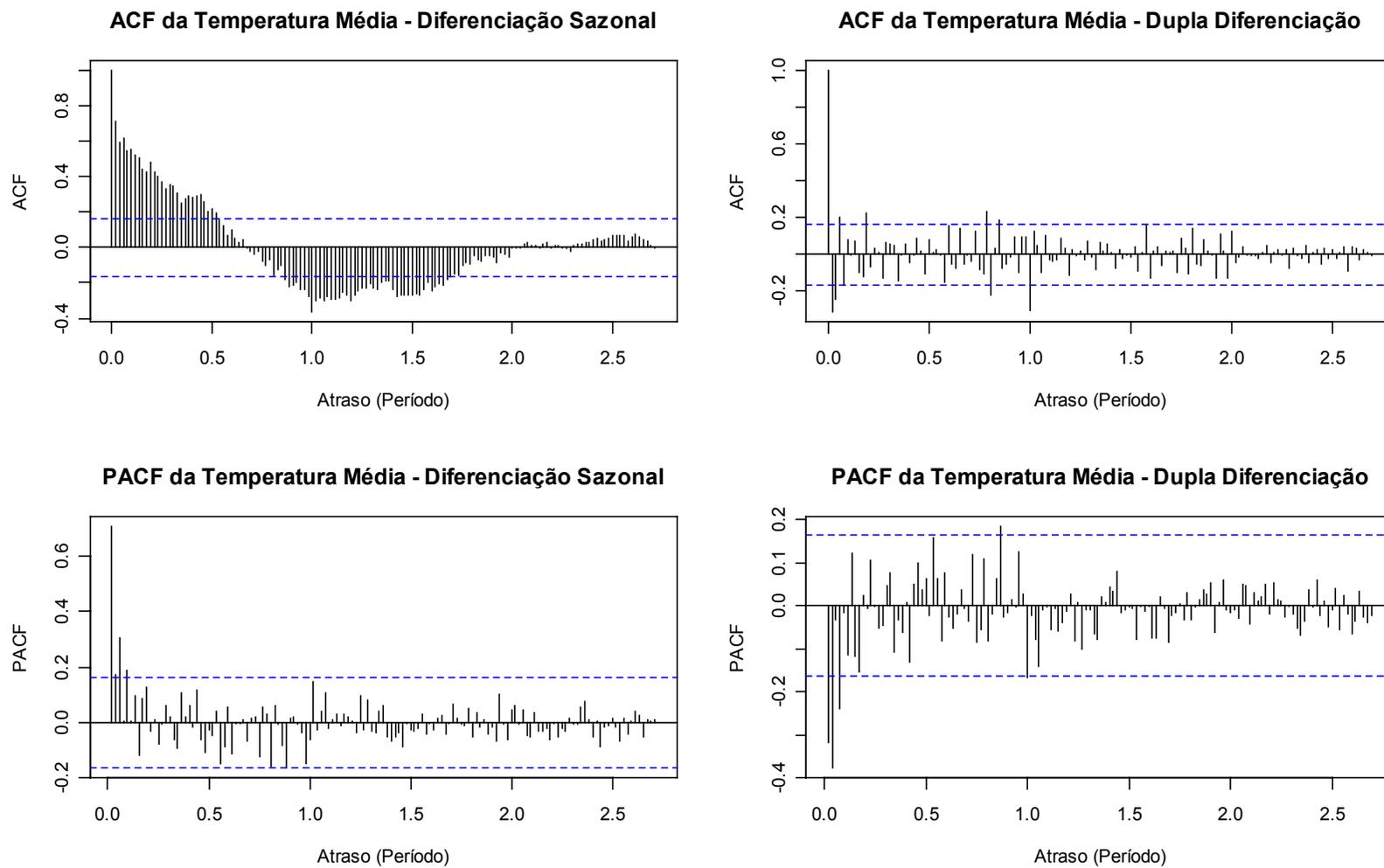
Figura 18 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Umidade e sua Diferenciação não Sazonal.



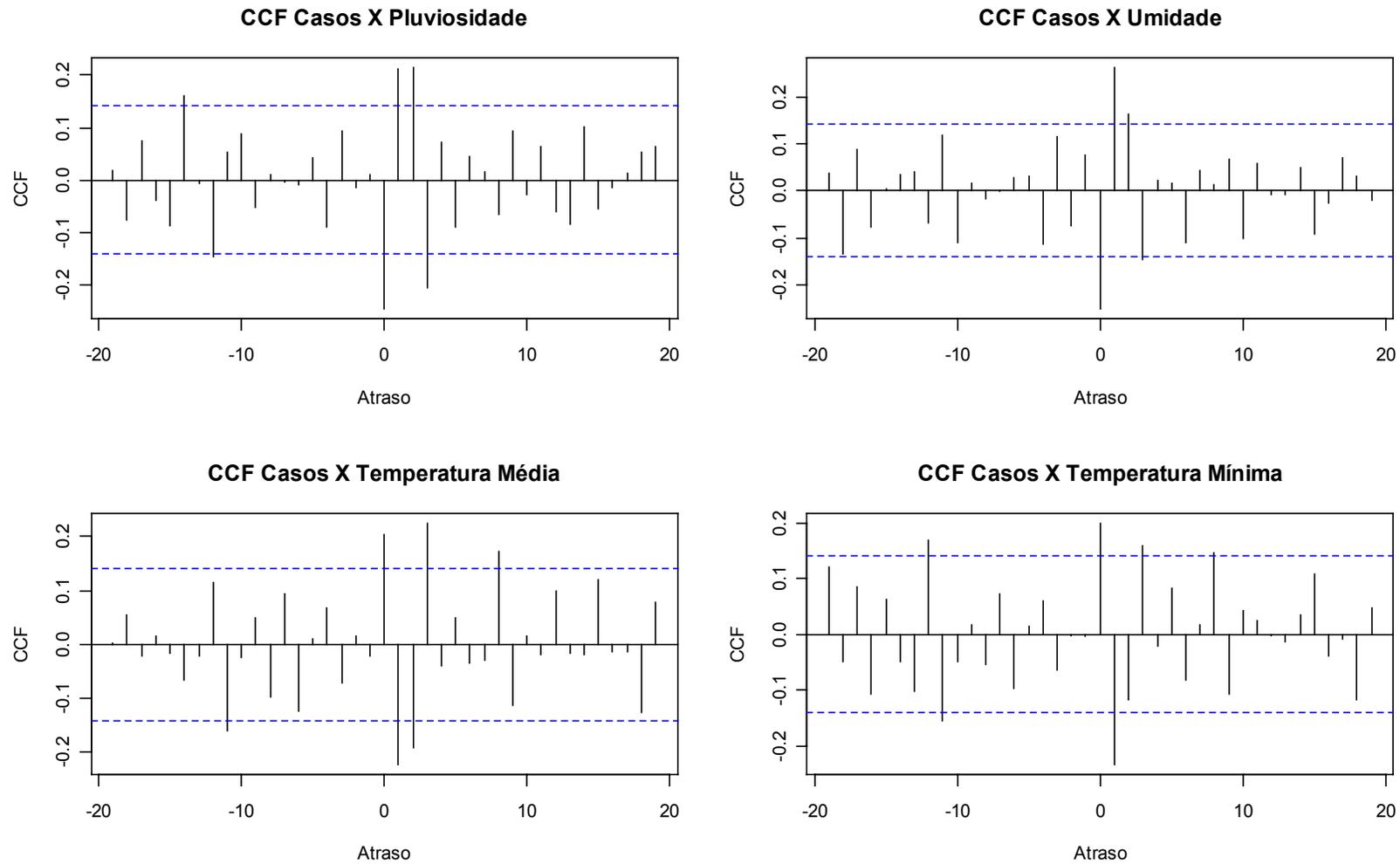
**Figura 19 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Umidade com Diferenciação Sazonal e com Diferenciação Sazonal e não sazonal.**



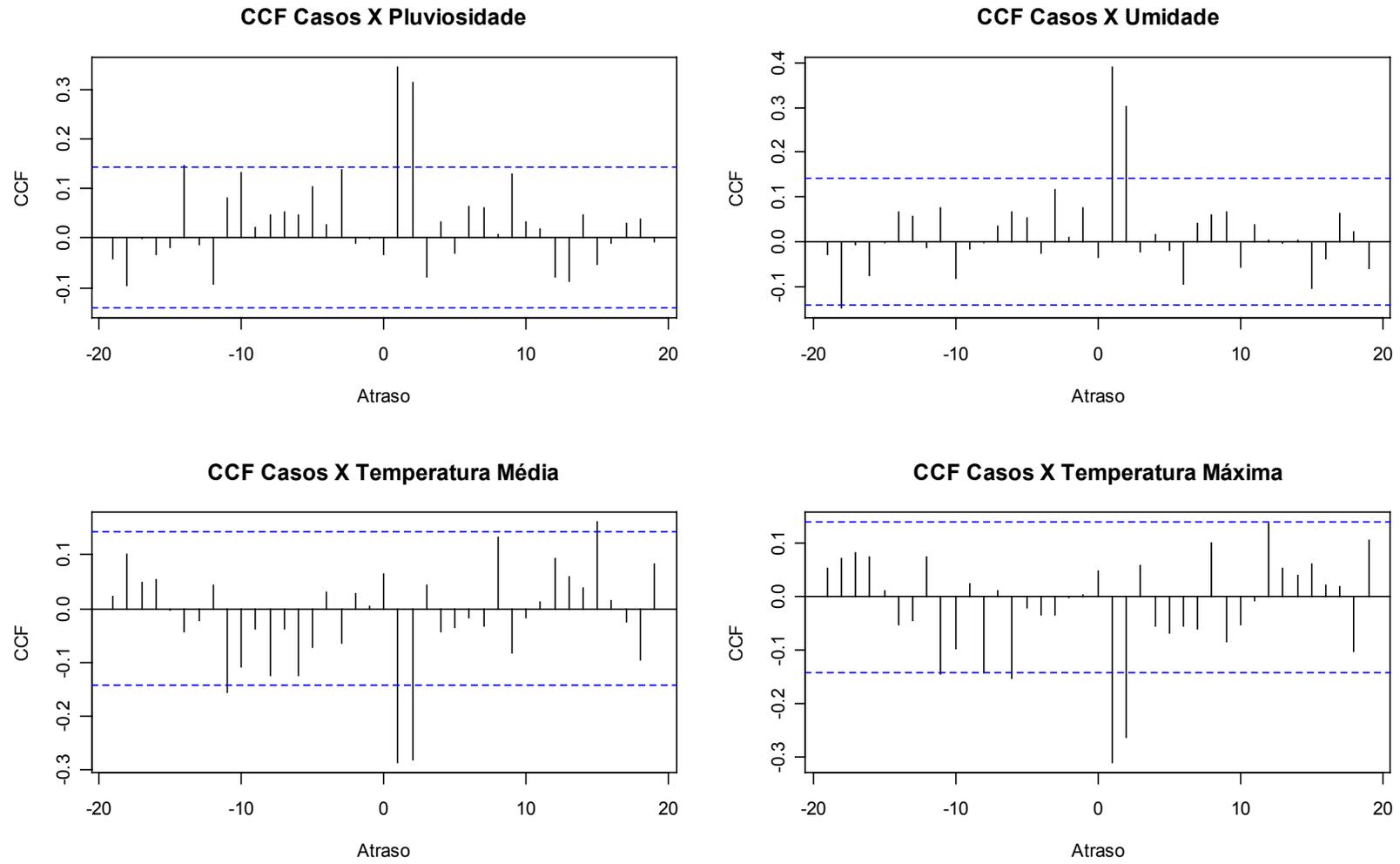
**Figura 20 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Temperatura Média e sua Diferenciação não Sazonal.**



**Figura 21 – Correlogramas das Funções de Autocorrelação e Autocorrelação Parcial e IC 95% para a Série de Temperatura Média com Diferenciação Sazonal e com Diferenciação Sazonal e não sazonal.**



**Figura 22 –Correlogramas das CCFs e IC 95% para a Série de Casos e: (a) Pluviosidade; (b) Umidade; (c) Temperatura Média; (d) Temperatura Mínima – Ambas as séries foram diferenciadas apenas.**



**Figura 23 – Correlogramas das CCFs e IC 95% para a Série de Casos e: (a) Pluviosidade – AR (3); (b) Umidade – ARMA (1,1); (c) Temperatura Média – ARIMA (2,1,3); (d) Temperatura Máxima ARIMA (2,1,3) – Ambas as séries foram modeladas por modelos ARIMA de mesma ordem – Pré-branqueamento**

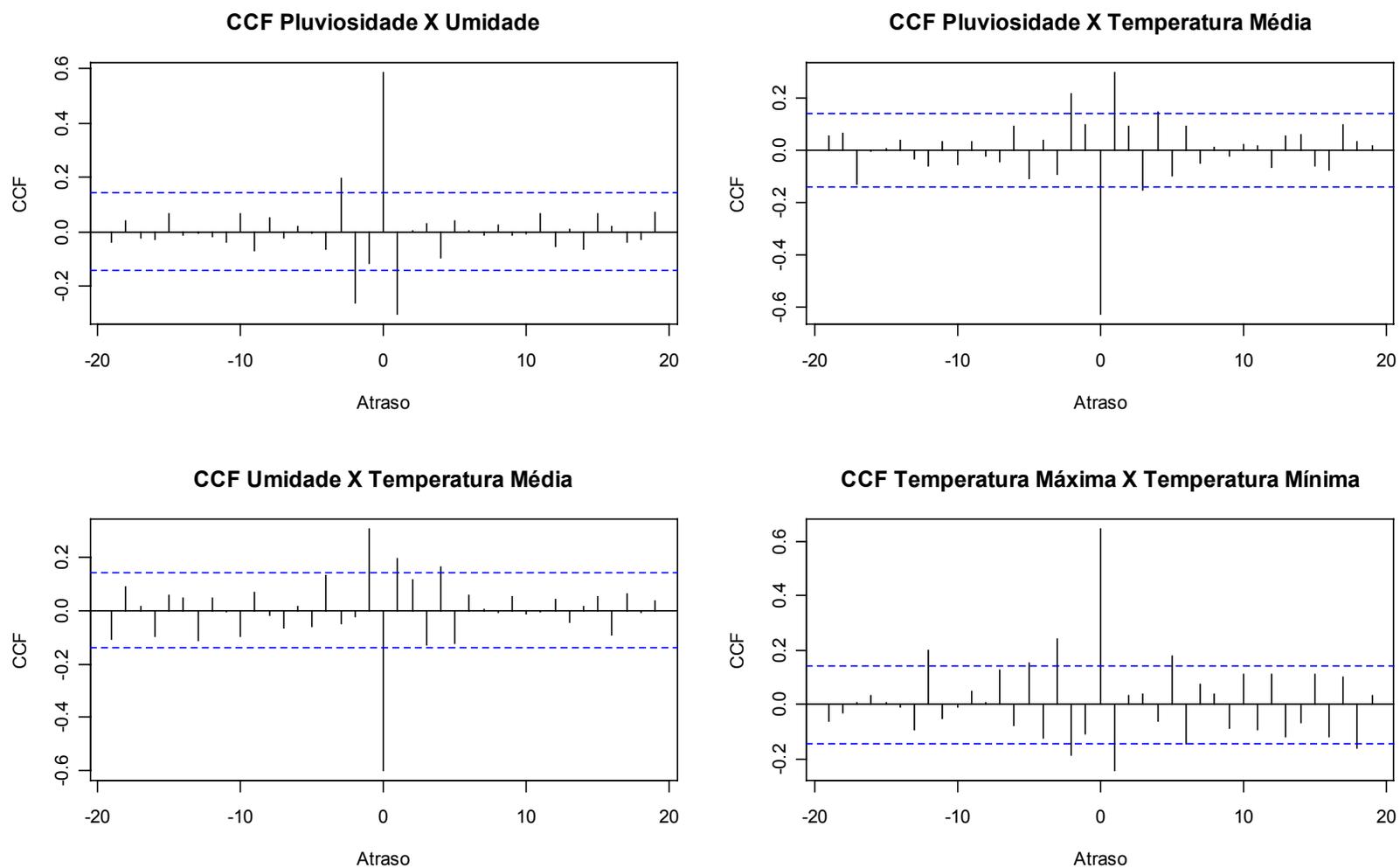
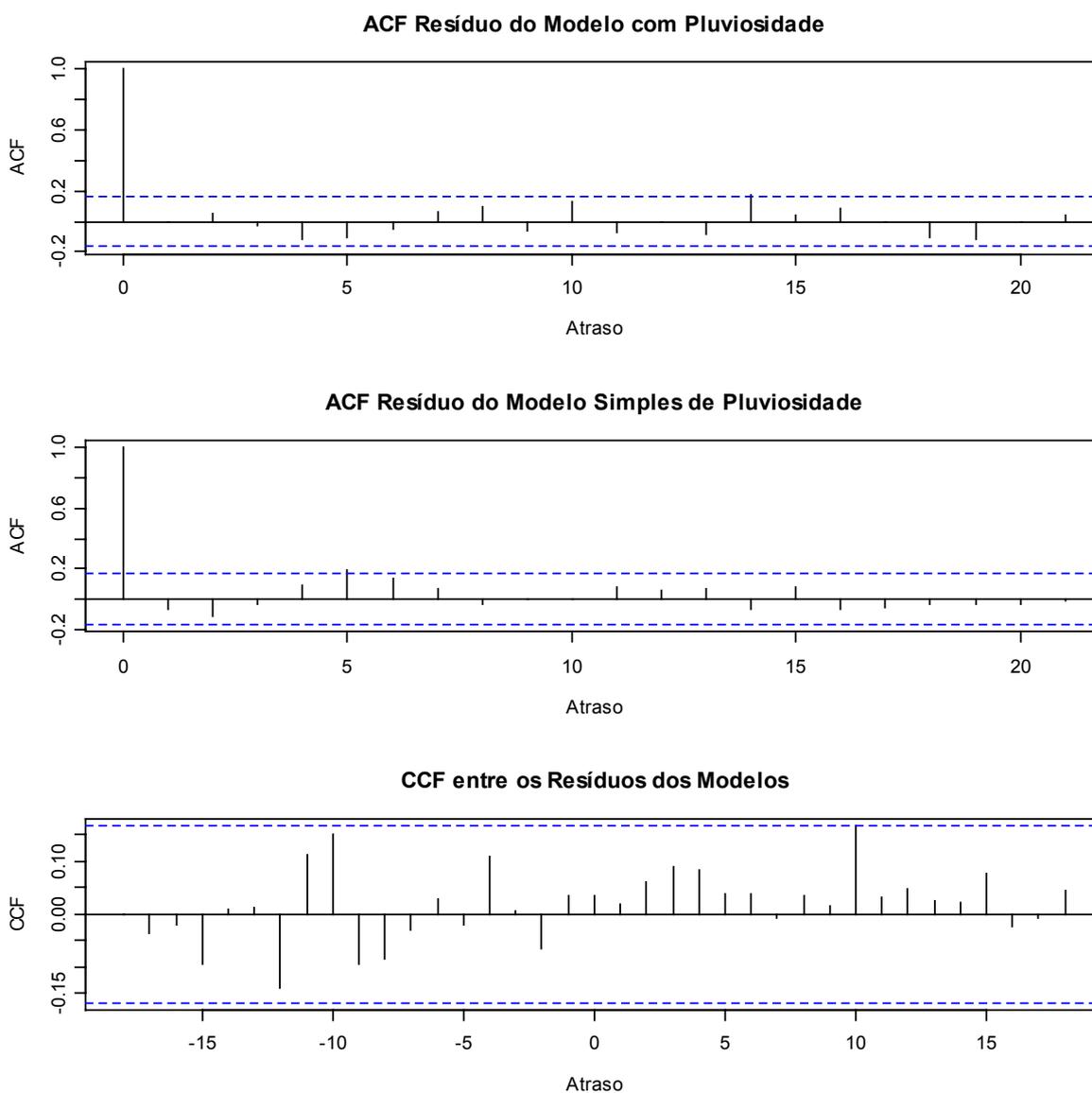
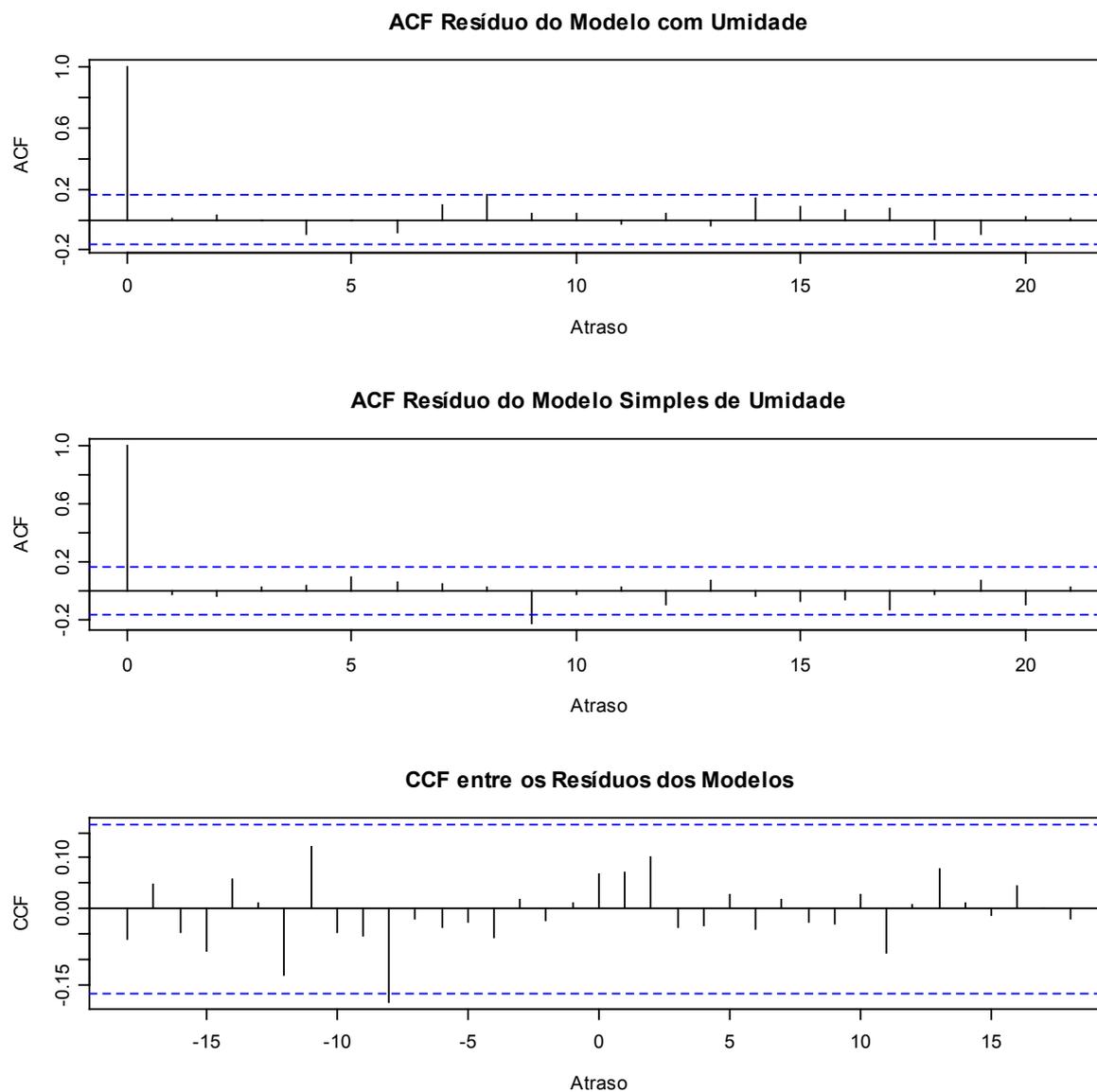


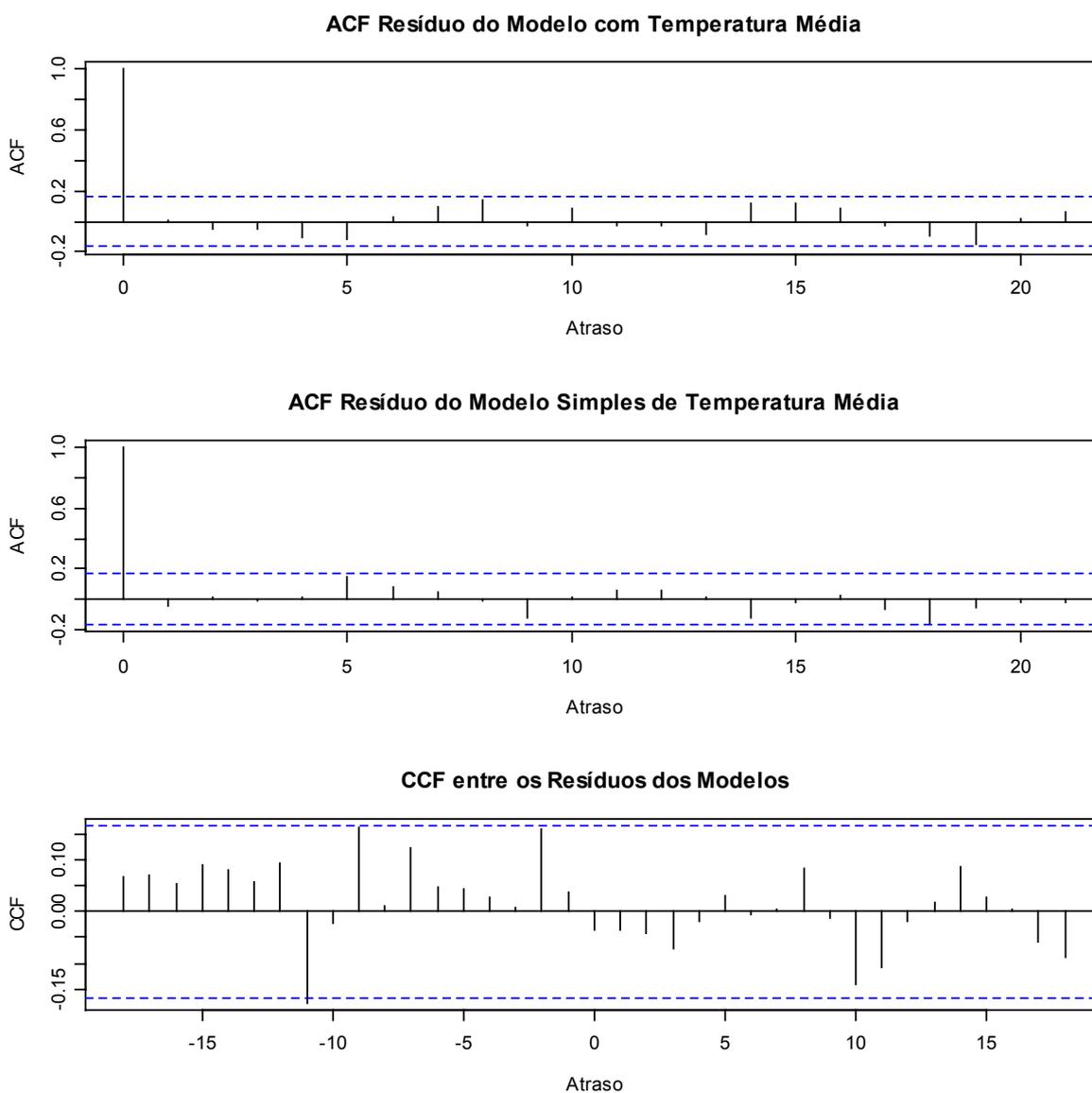
Figura 24 – Correlogramas das CCFs e IC 95% para as Séries Diferenciadas de (a) Pluviosidade e Umidade; (b) Pluviosidade e Temperatura Média; (c) Umidade e Temperatura Média; (d) Temperatura Mínima.



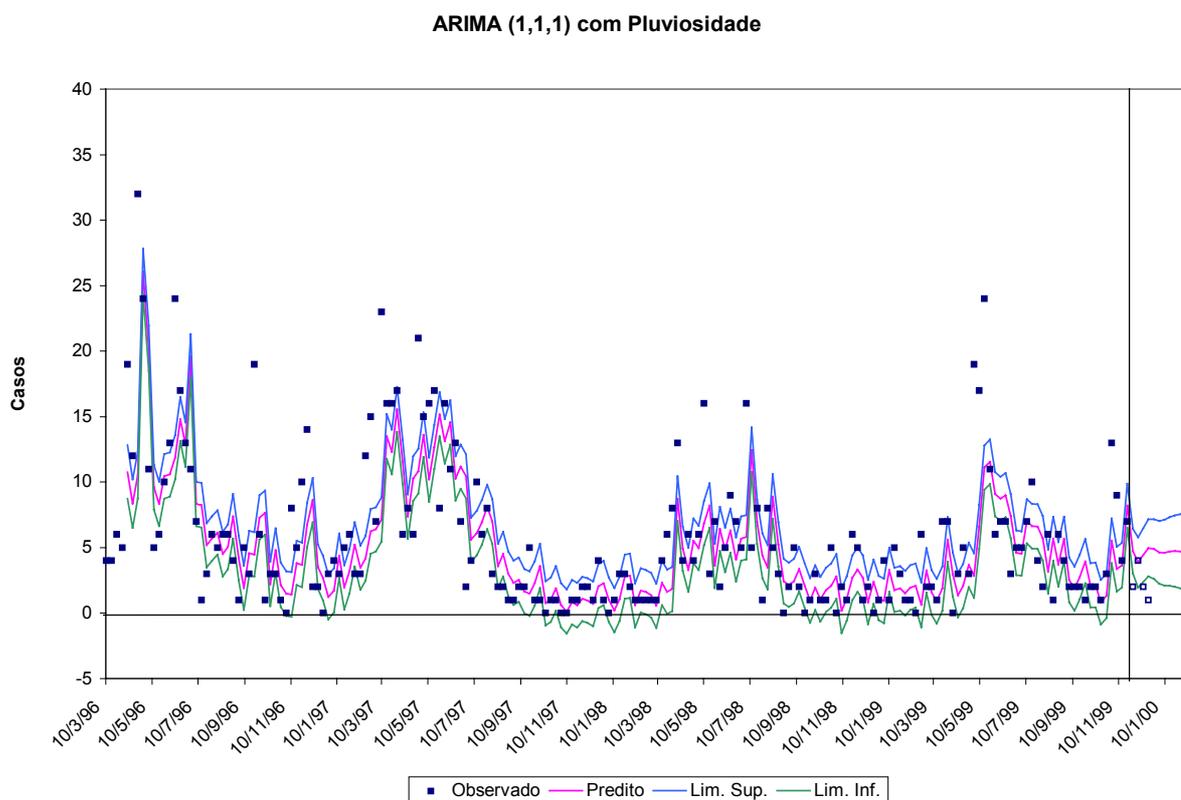
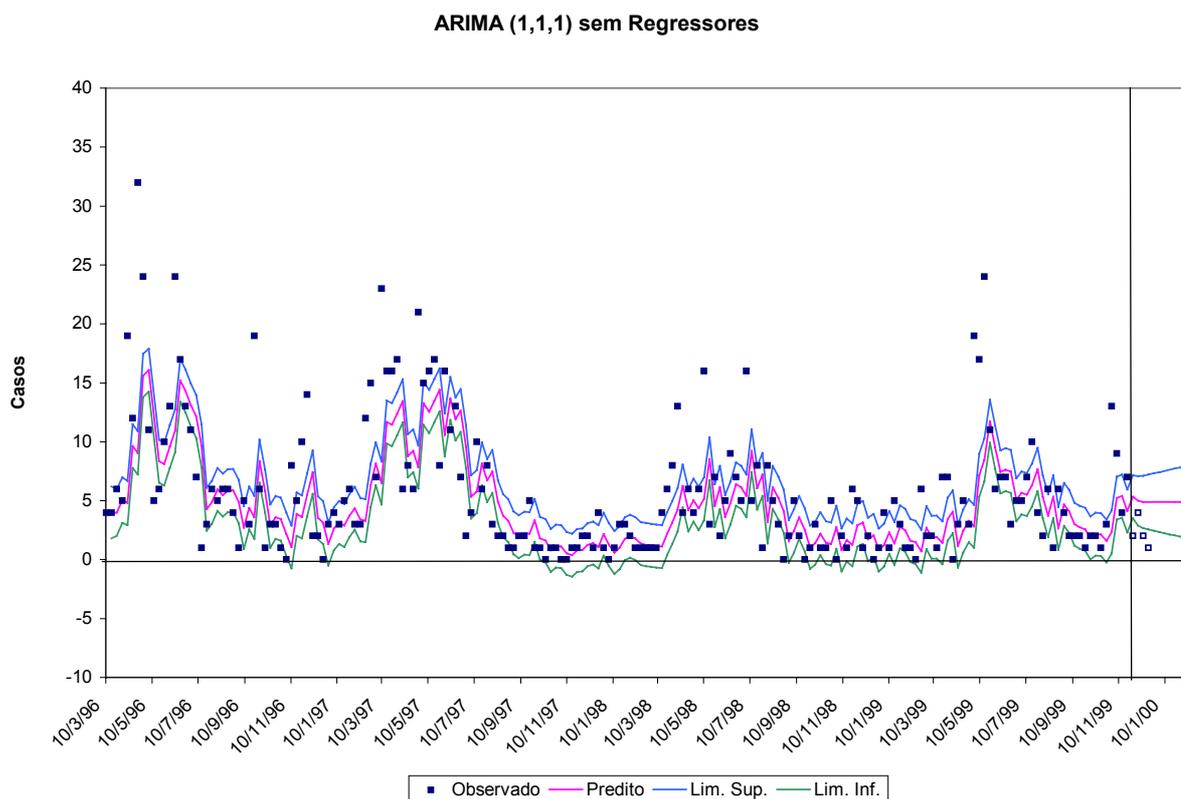
**Figura 25 – Diagnósticos para os Modelos com Regressores Dinâmicos – Modelo SARIMA (1,1,1)x(0,1,1) com Pluviosidade como Entrada. Correlogramas e IC 95% para: (a) ACF dos Resíduos do Modelo com Pluviosidade como Entrada; (b) ACF Modelo Pré-Branqueado de Pluviosidade - AR (3) e (c) CCF entre os Resíduos do Modelo com Pluviosidade como Entrada e os Resíduos do Modelo Pré-Branqueado de Pluviosidade - AR (3).**



**Figura 26 – Diagnósticos para os Modelos com Regressores Dinâmicos – Modelo SARIMA (1,0,1)x(0,1,1) com Umidade como Entrada. Correlogramas e IC 95% para: (a) ACF dos Resíduos do Modelo com Umidade como Entrada; (b) ACF Modelo Pré-Branqueado de Umidade - ARMA (1,1) e (c) CCF entre os Resíduos do Modelo com Umidade como Entrada e os Resíduos do Modelo Pré-Branqueado de Umidade - ARMA (1,1); (d).**

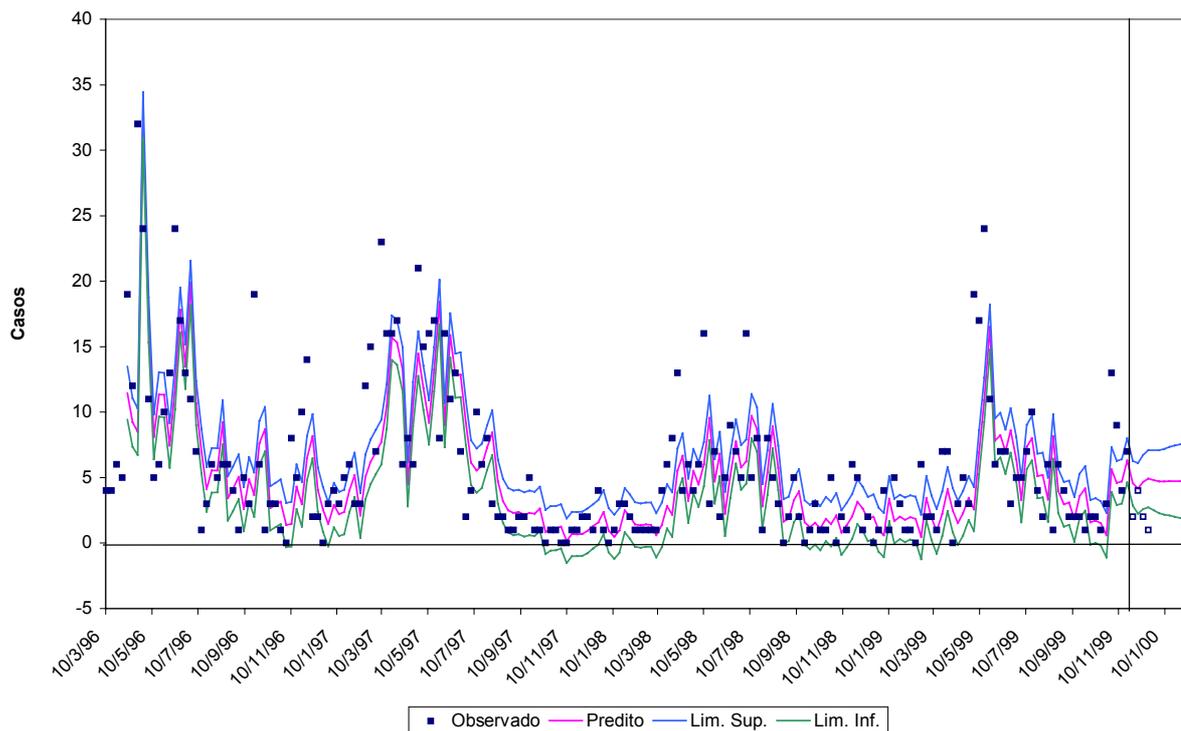


**Figura 27 – Diagnósticos para os Modelos com Regressores Dinâmicos – Modelo SARIMA (1,0,1)x(0,1,1) com Temperatura Média como Entrada Correlogramas e IC 95% para: (a) ACF dos Resíduos do Modelo; (b) ACF Modelo Pré-Branqueado de Temperatura Média - ARIMA (2,1,3) e (c) CCF entre os Resíduos do Modelo com Temperatura Média como Entrada e os Resíduos do Modelo Pré-Branqueado de Temperatura Média - ARIMA (2,1,3).**



**Figura 28 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados). Acima: Modelo ARIMA (1,1,1) sem Regressores Dinâmicos. Abaixo: Modelo ARIMA (1,1,1) com Pluviosidade (2,0,0) como Regressor Dinâmico.**

ARIMA (1,1,1) com Umidade



ARIMA (1,1,1) com Temperatura Máxima

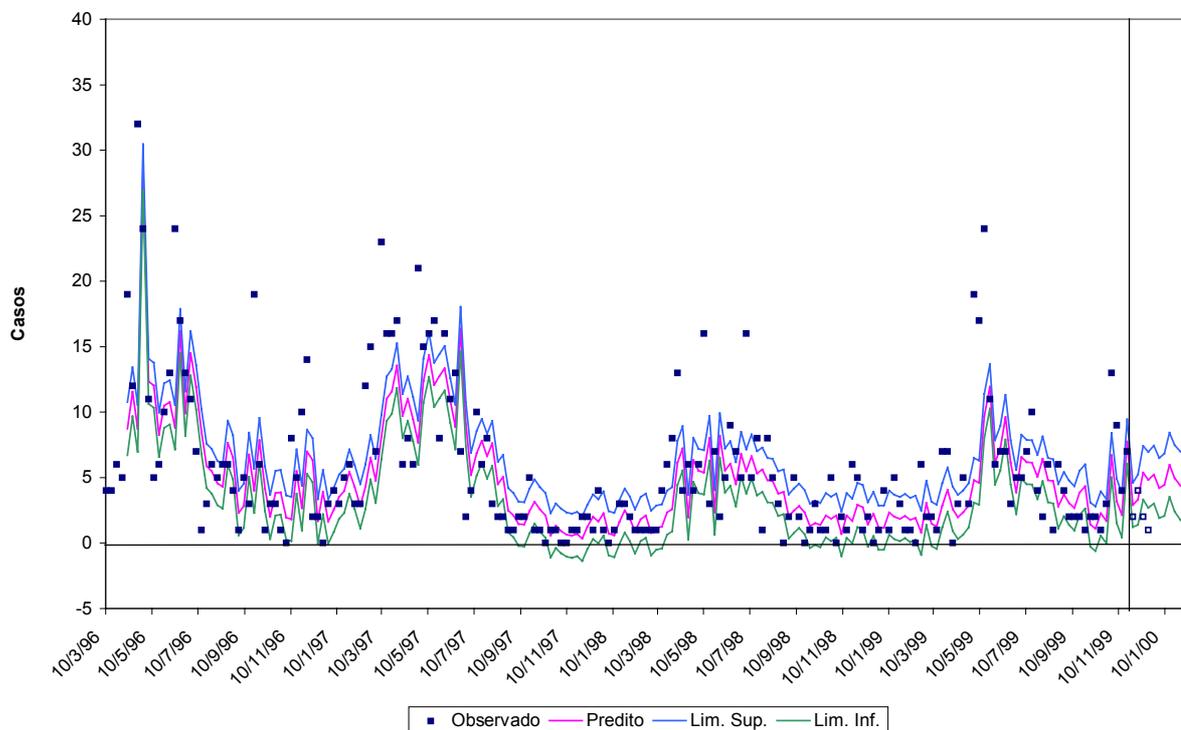
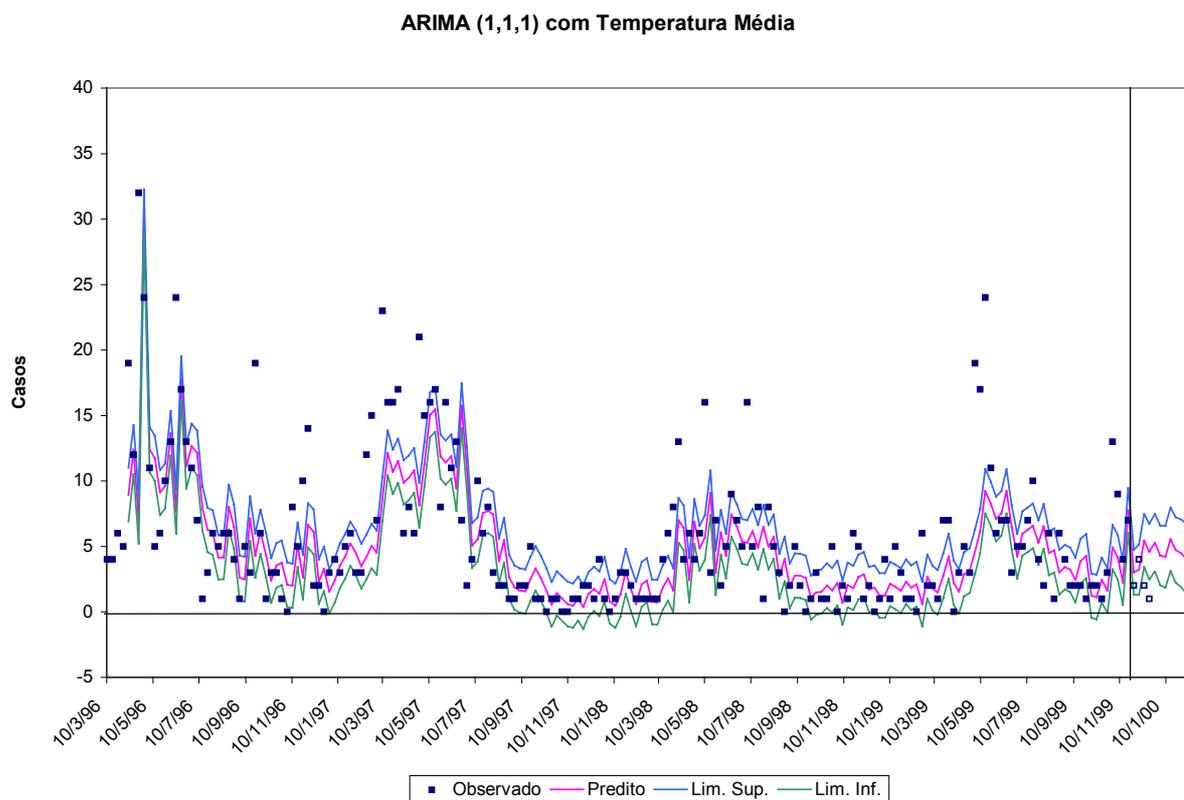
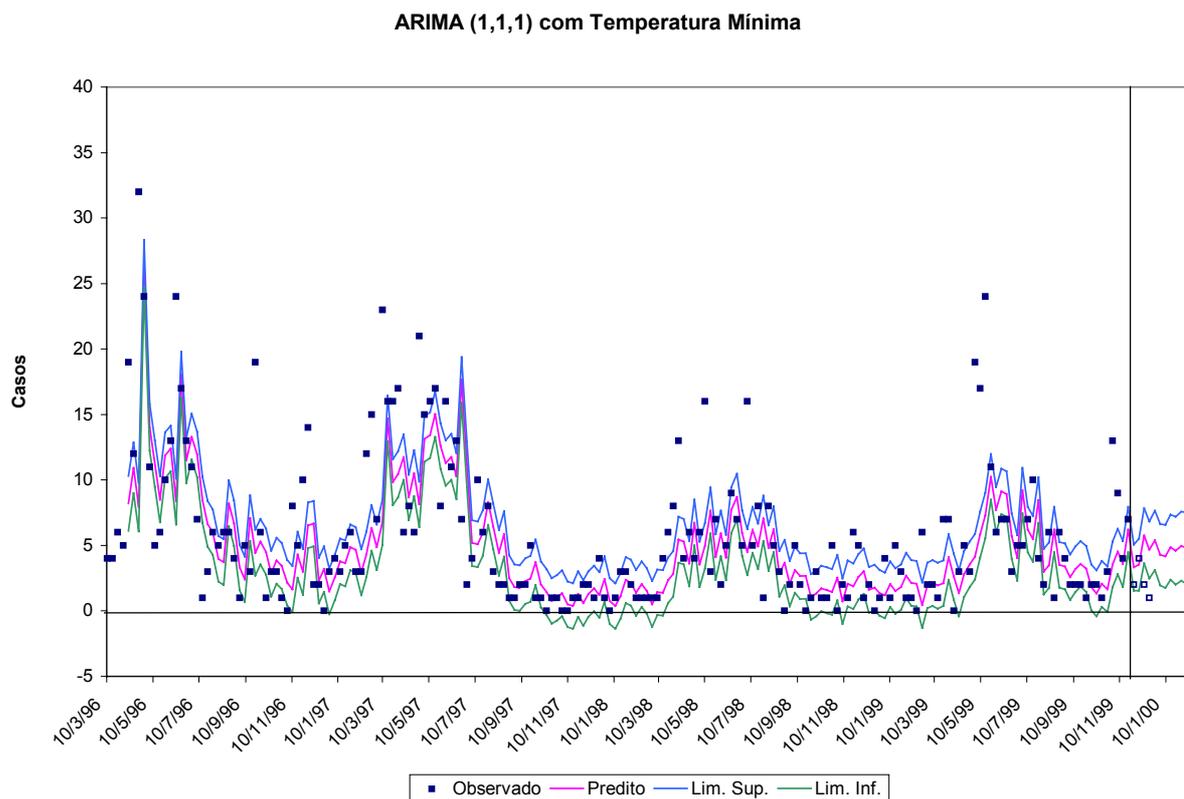
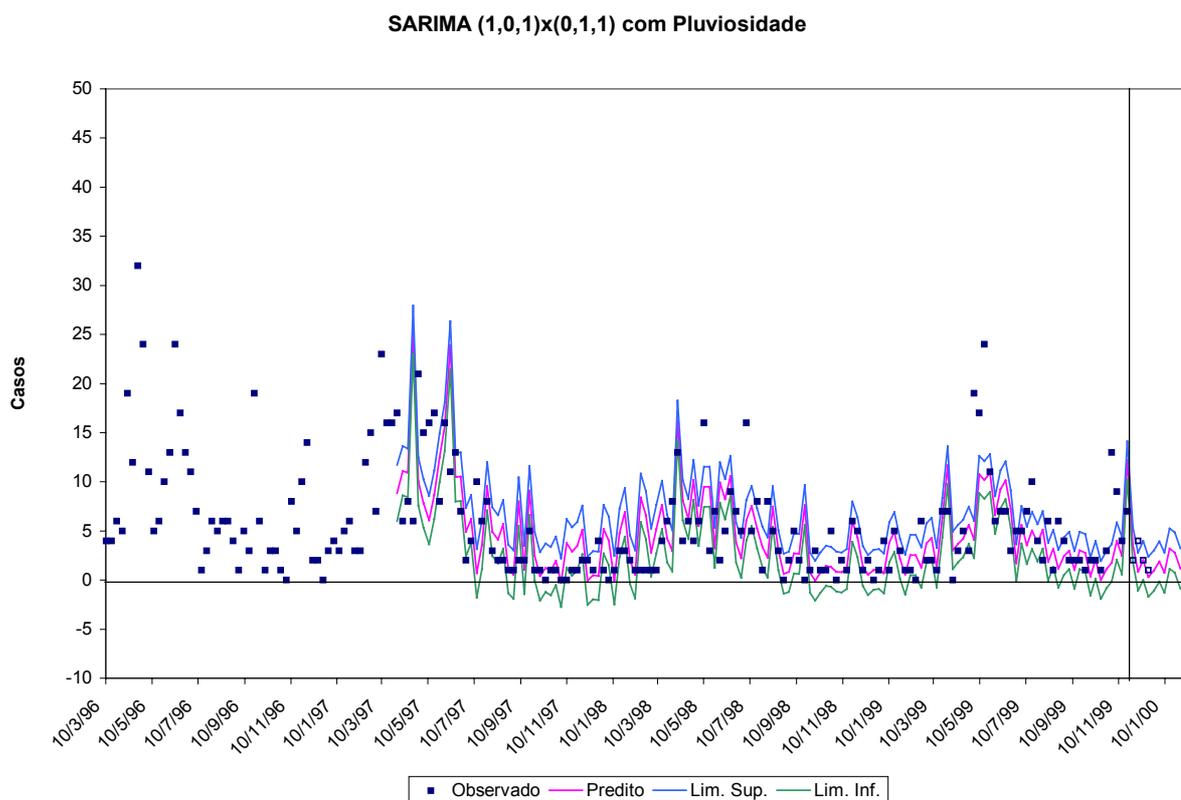
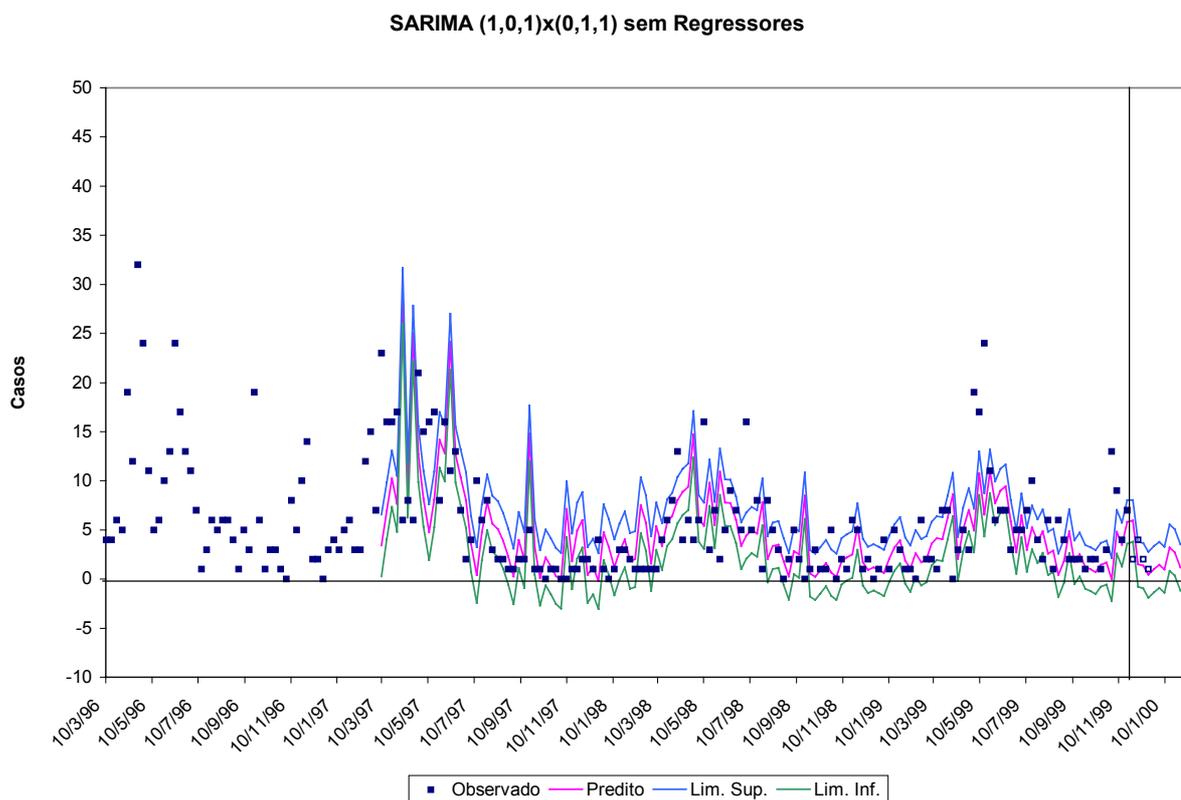


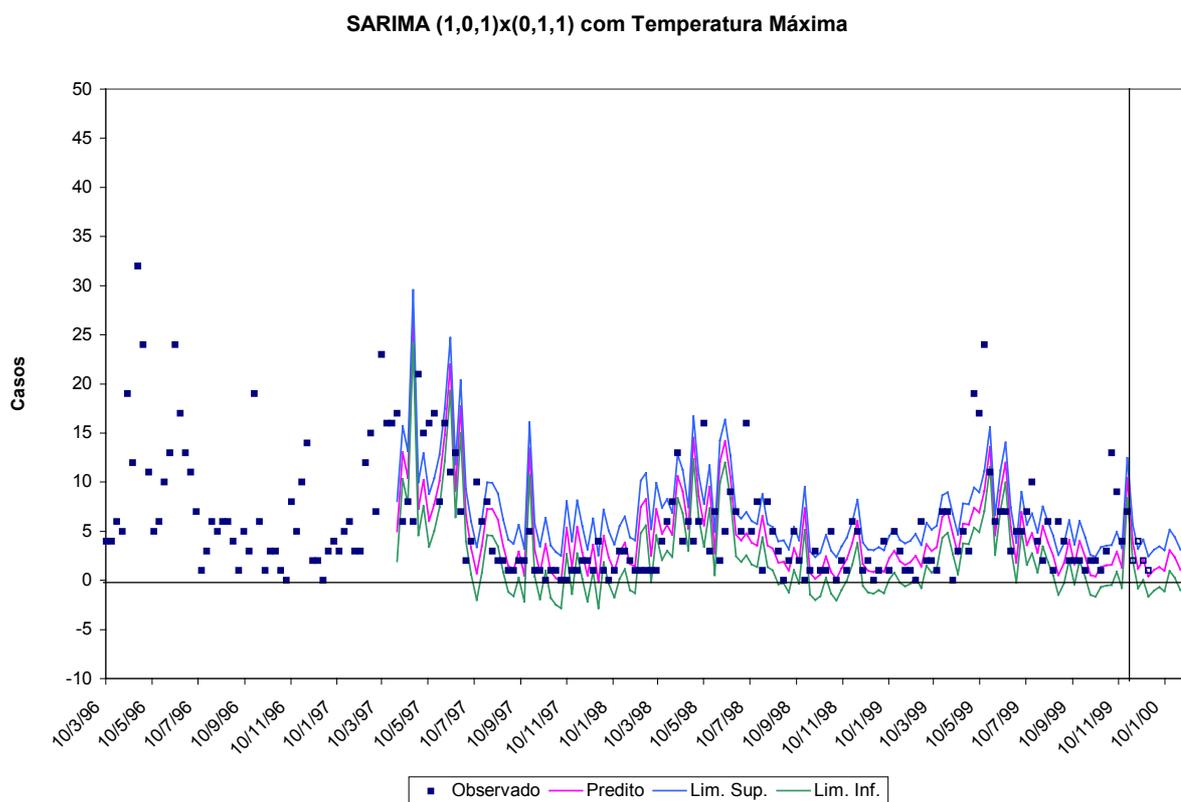
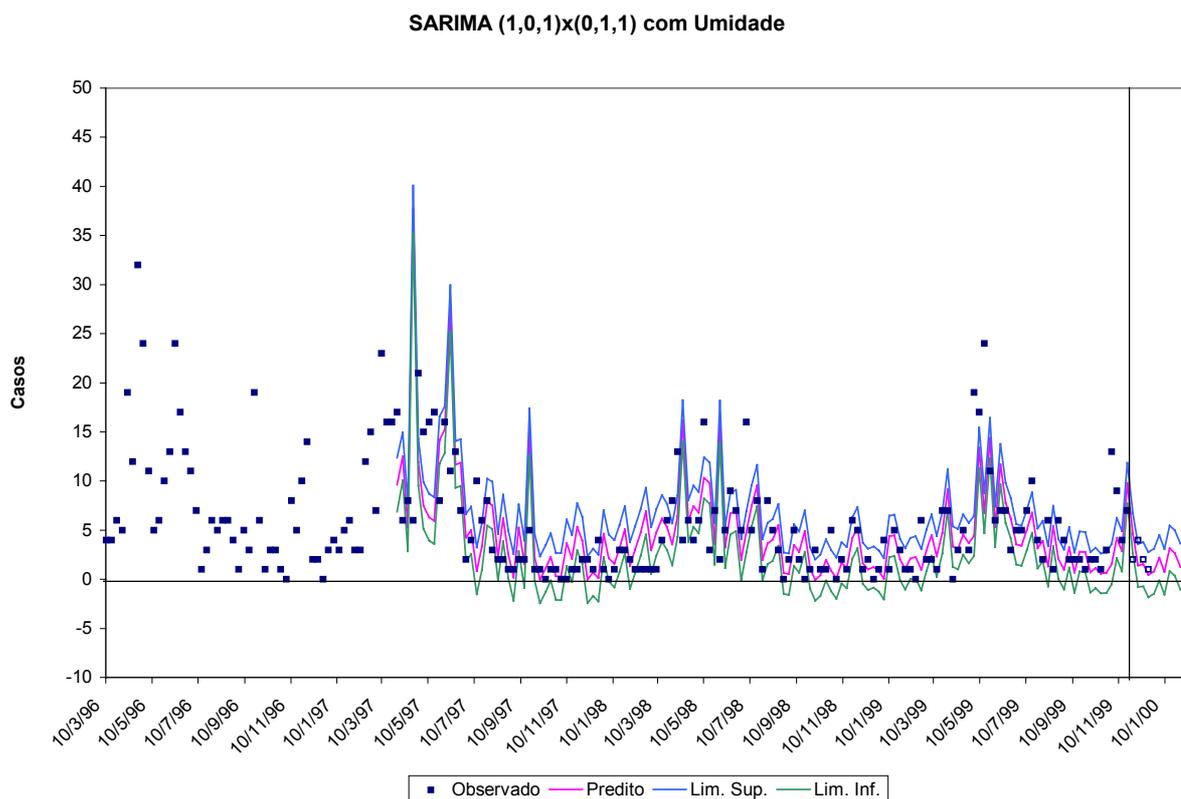
Figura 29 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados). Acima: Modelo ARIMA (1,1,1) com Umidade (2,0,0) como Regressor Dinâmico. Abaixo: Modelo ARIMA (1,1,1) com Temperatura Máxima (0,1,1) como Regressor Dinâmico.



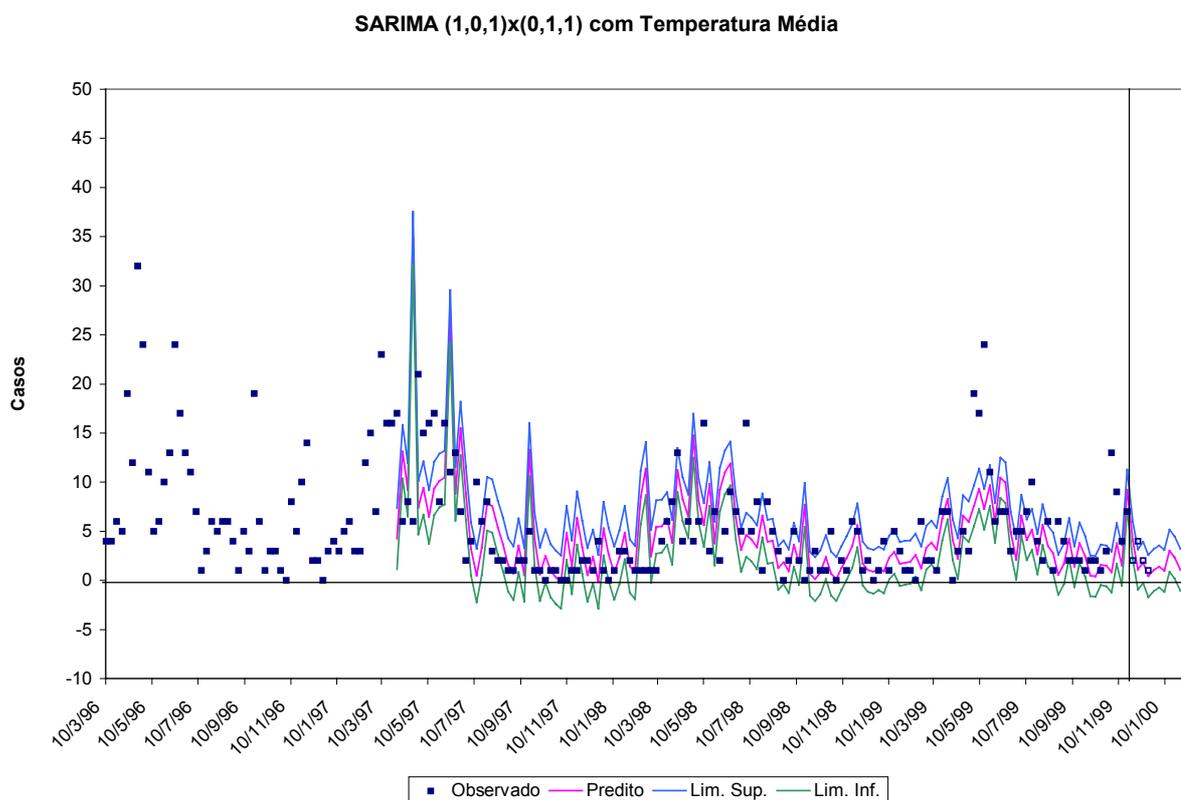
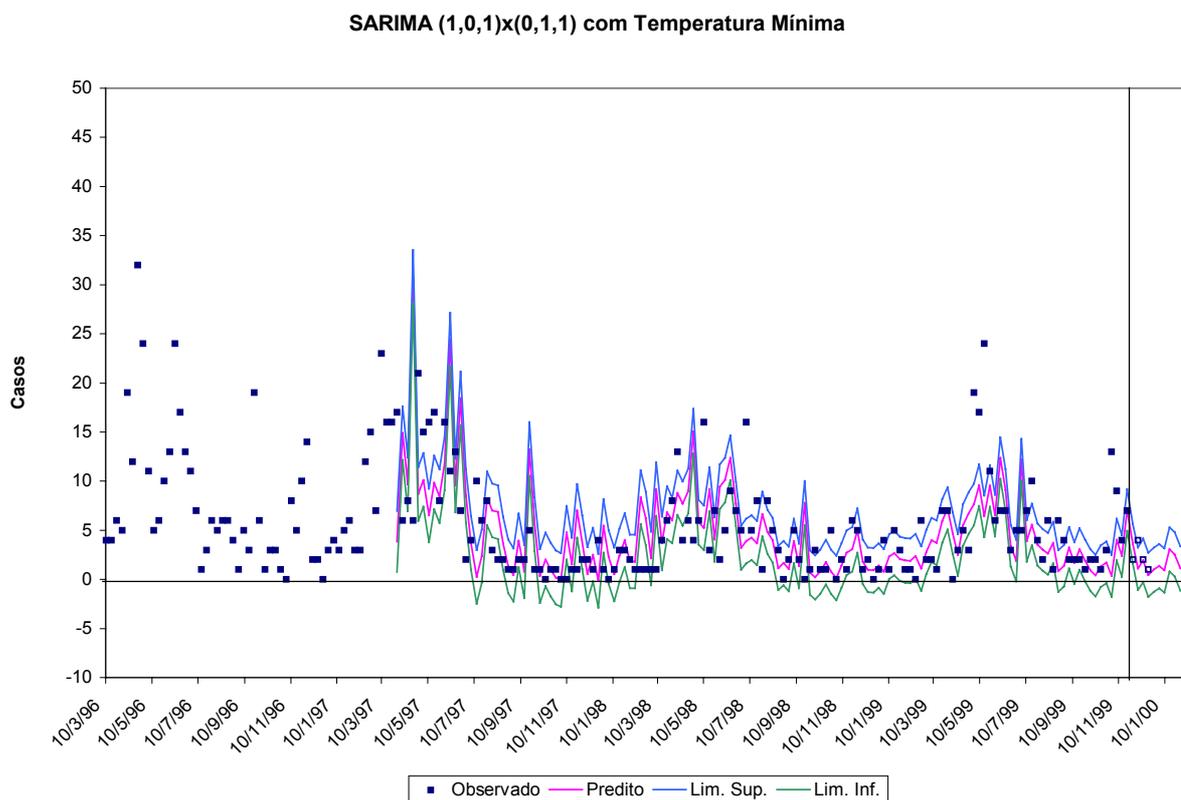
**Figura 30 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados). Acima: Modelo ARIMA (1,1,1) com Temperatura Mínima (0,1,1) como Regressor Dinâmico. Abaixo: Modelo ARIMA (1,1,1) com Temperatura Média (0,1,1) como Regressor Dinâmico.**



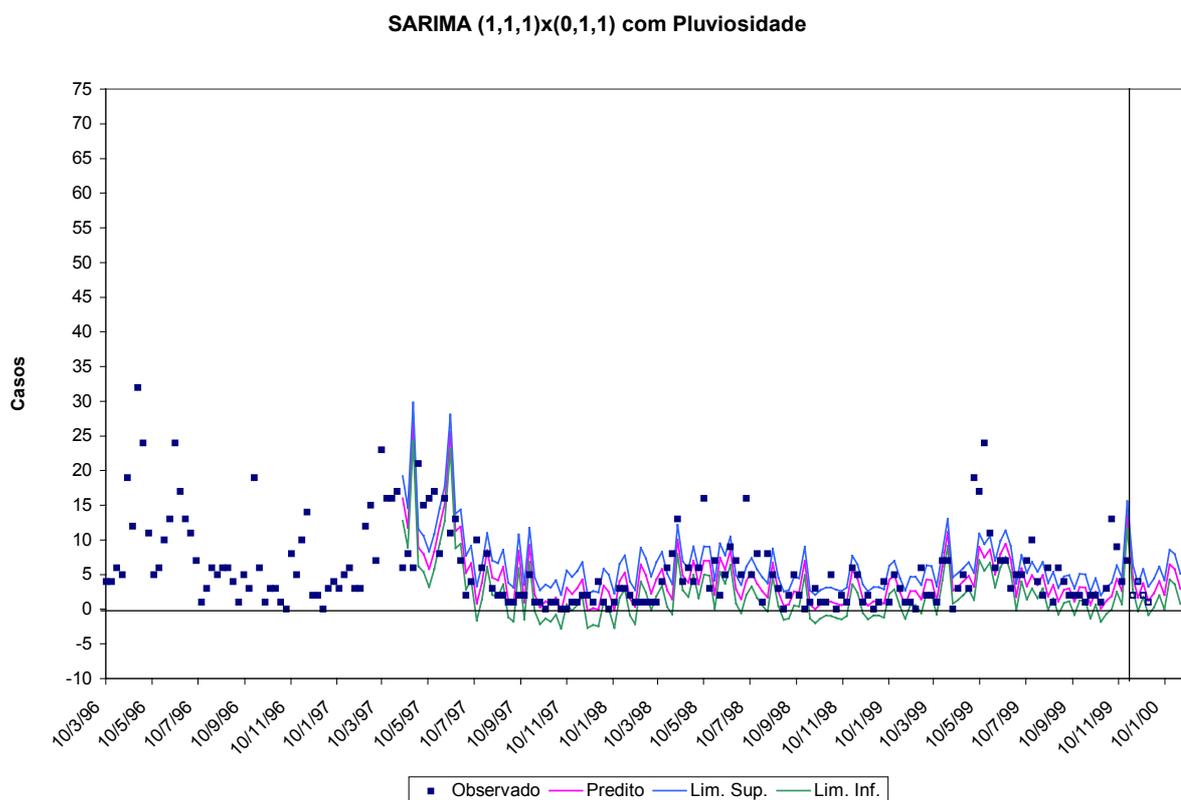
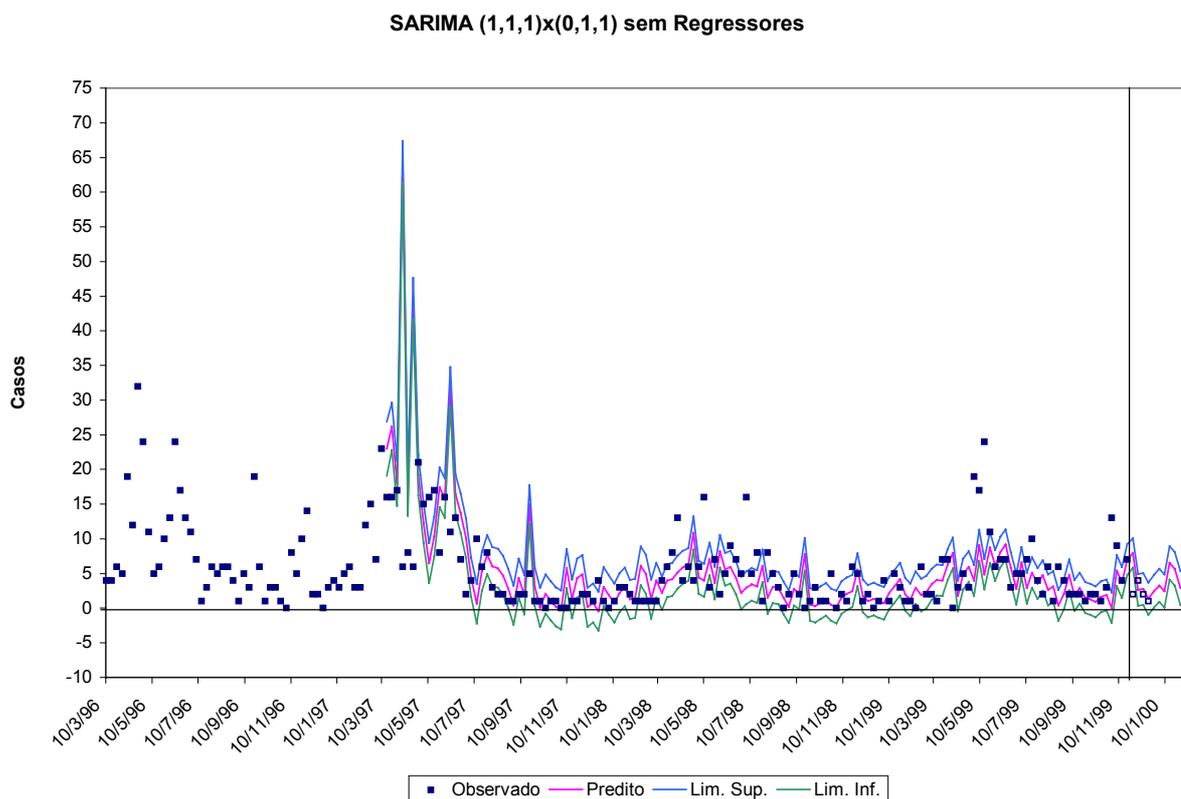
**Figura 31 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados). Acima: Modelo SARIMA (1,0,1)x(0,1,1) sem Regressores Dinâmicos. Abaixo: Modelo SARIMA (1,0,1)x(0,1,1) com Pluviosidade (2,0,0) como Regressor Dinâmico.**



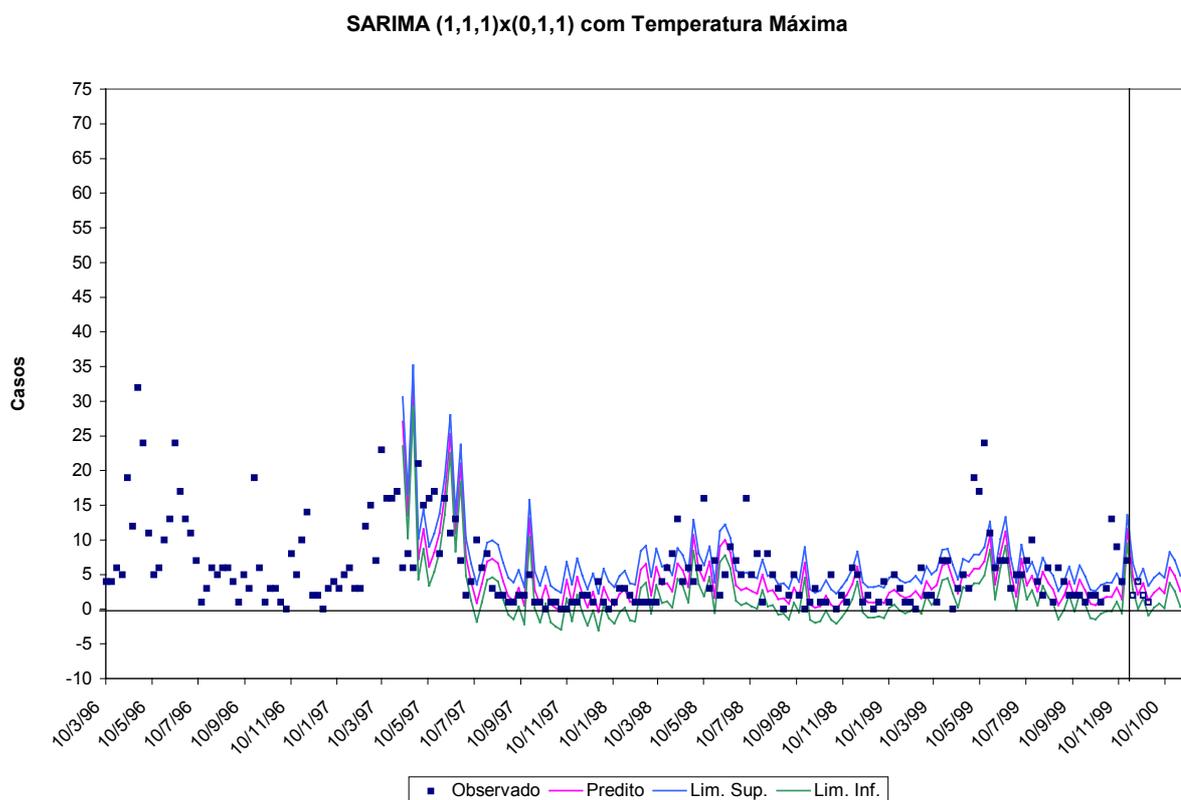
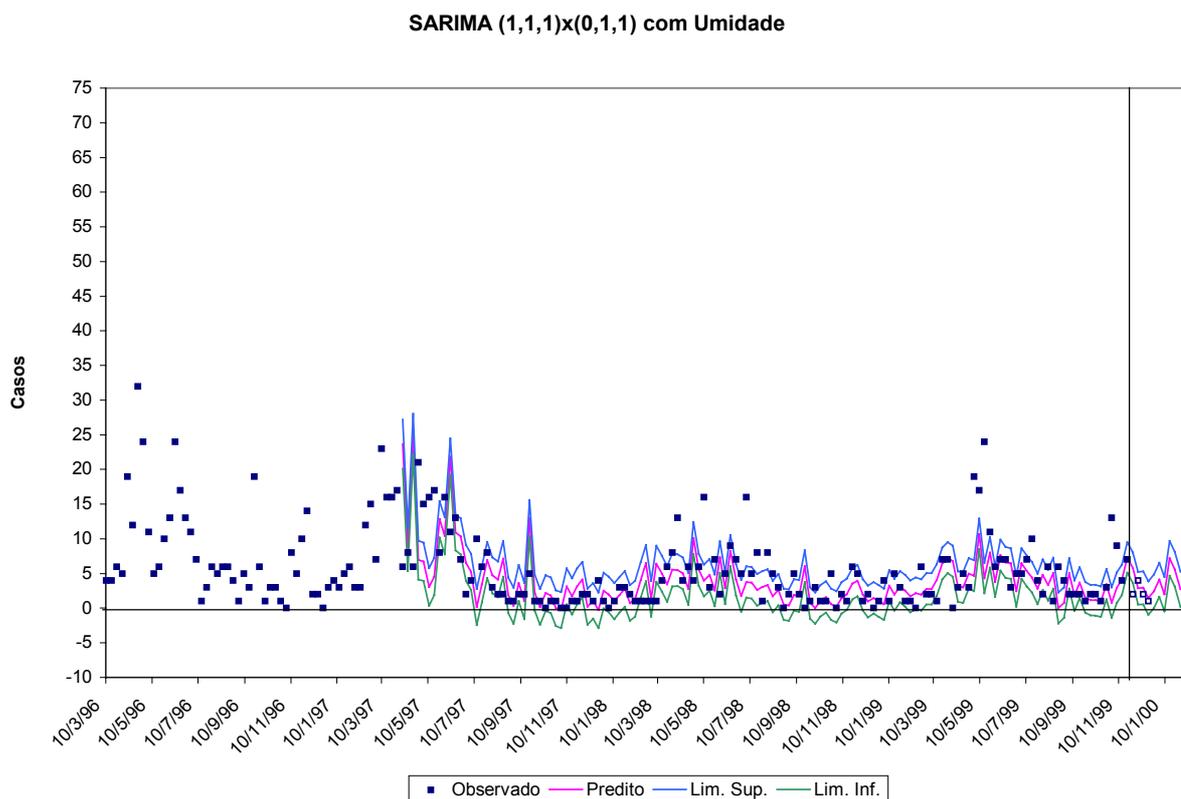
**Figura 32 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados). Acima: Modelo SARIMA (1,0,1)x(0,1,1) com Umidade (2,0,0) como Regressor Dinâmico. Abaixo: Modelo SARIMA (1,0,1)x(0,1,1) com Temperatura Máxima (0,1,1) como Regressor Dinâmico.**



**Figura 33 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados). Acima: Modelo SARIMA (1,0,1)x(0,1,1) com Temperatura Mínima (0,1,1) como Regressor Dinâmico. Abaixo: Modelo SARIMA (1,0,1)x(0,1,1) com Temperatura Média (0,1,1) como Regressor Dinâmico.**

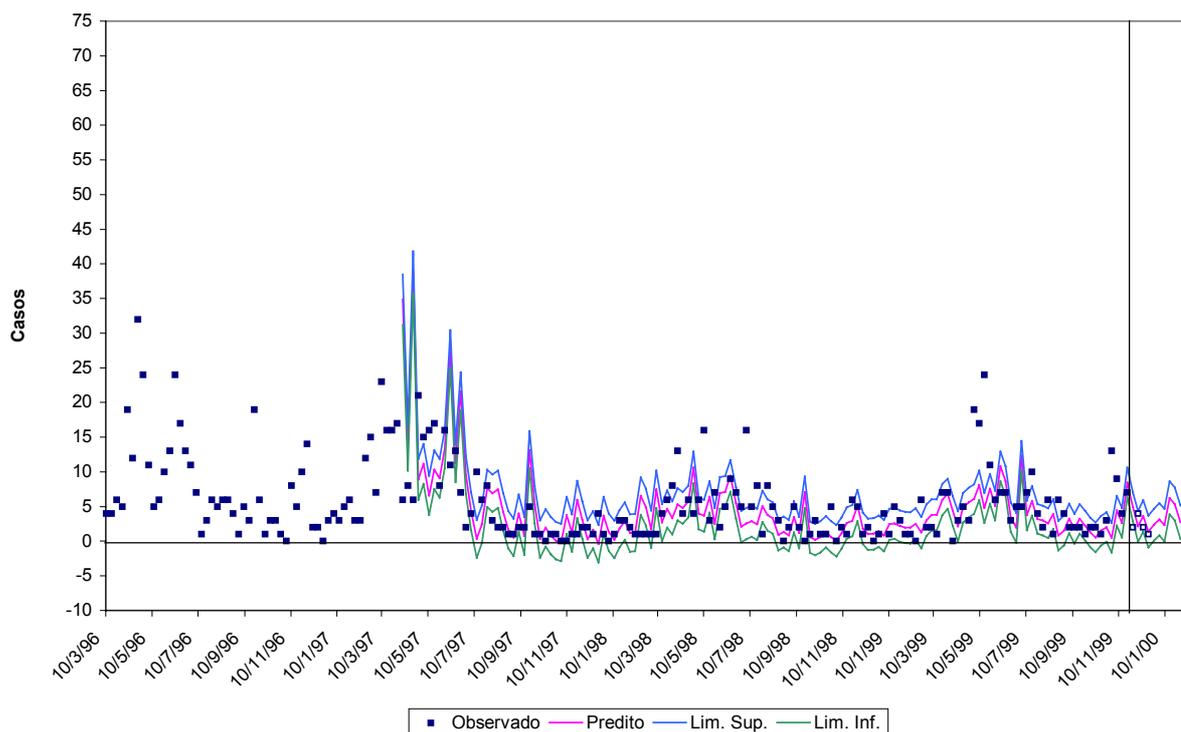


**Figura 34 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados). Acima: Modelo SARIMA (1,1,1)x(0,1,1) sem Regressores Dinâmicos. Abaixo: Modelo SARIMA (1,1,1)x(0,1,1) com Pluviosidade (2,0,0) como Regressor Dinâmico.**

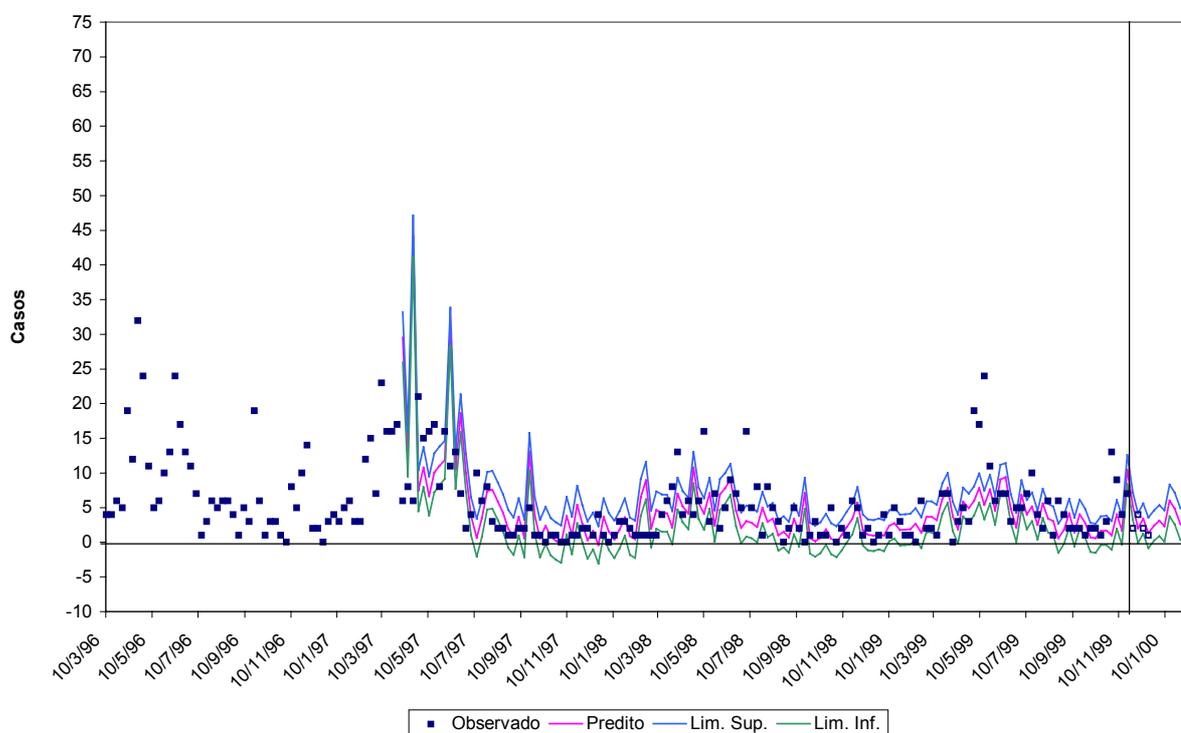


**Figura 35 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados). Acima: Modelo SARIMA (1,1,1)x(0,1,1) com Umidade (1,0,1) como Regressor Dinâmico. Abaixo: Modelo SARIMA (1,1,1)x(0,1,1) com Temperatura Máxima (0,1,1) como Regressor Dinâmico.**

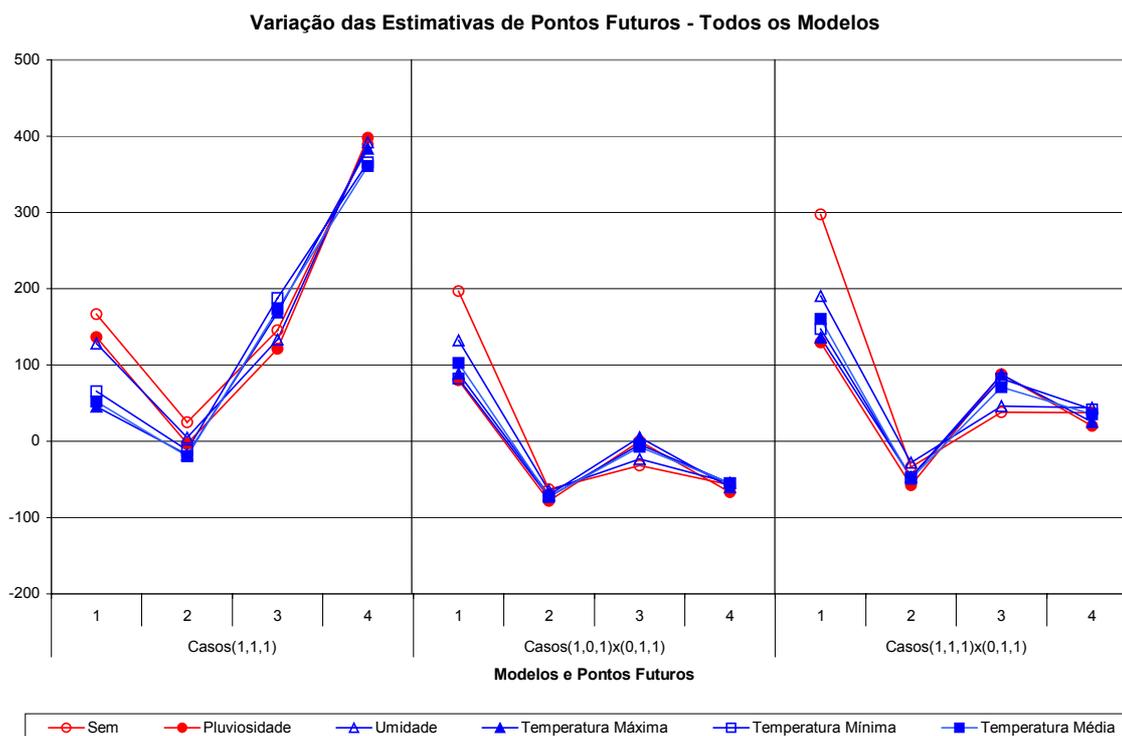
SARIMA (1,1,1)x(0,1,1) com Temperatura Mínima



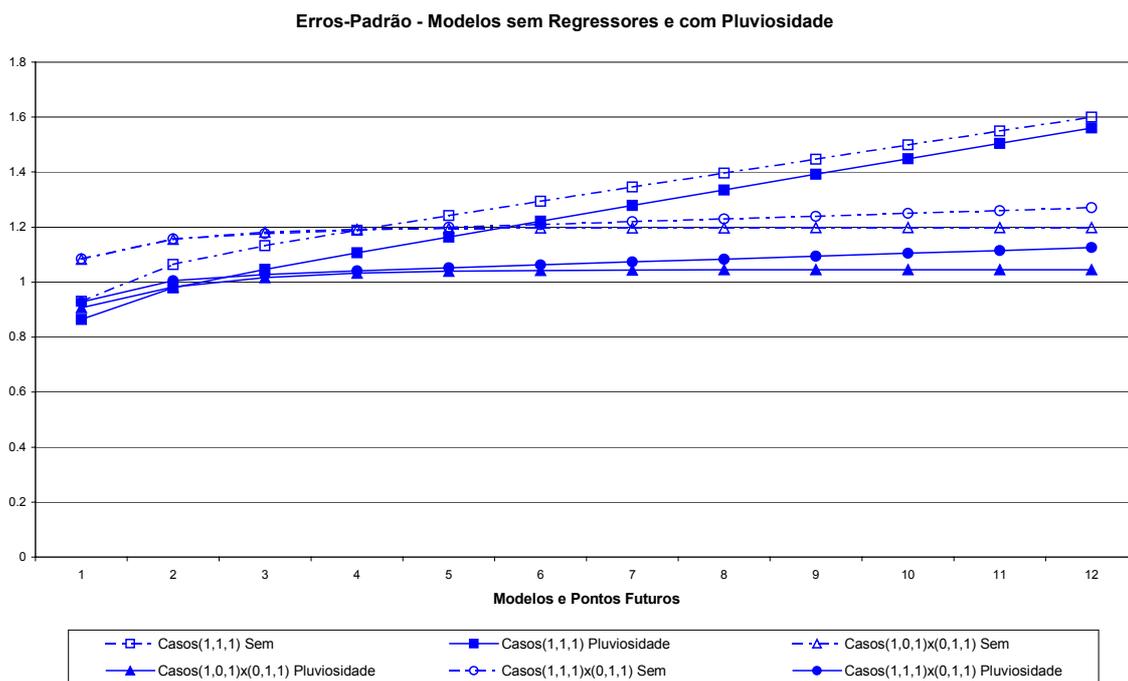
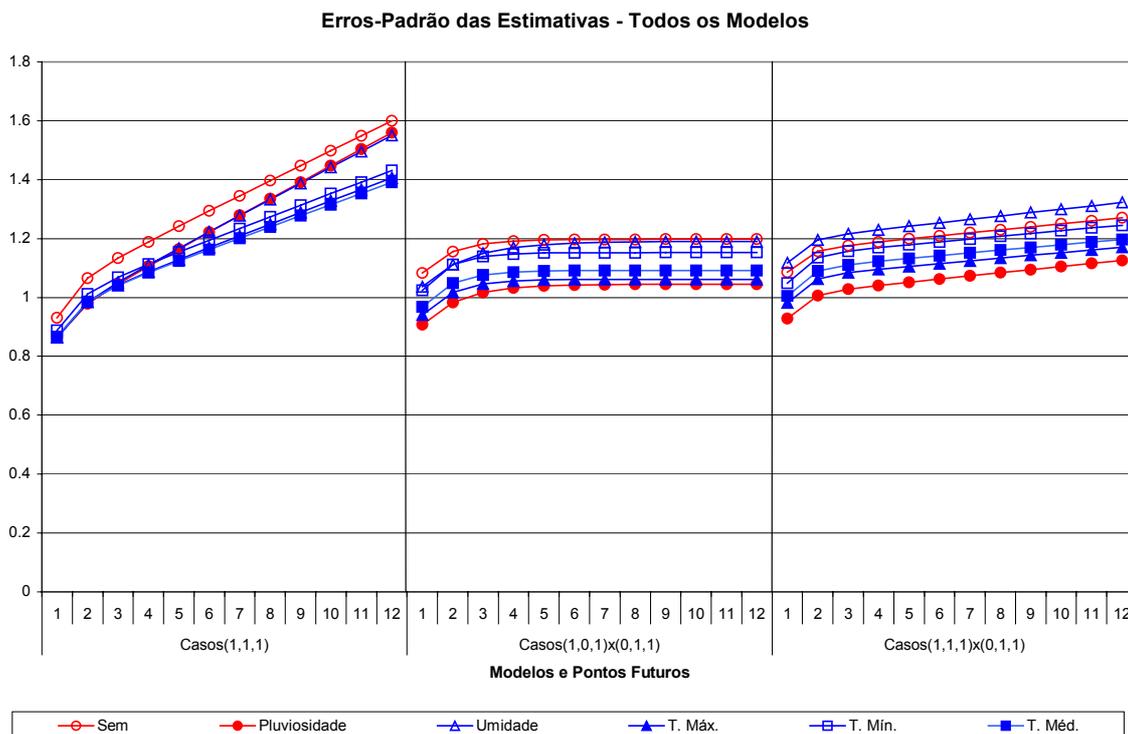
SARIMA (1,1,1)x(0,1,1) com Temperatura Média



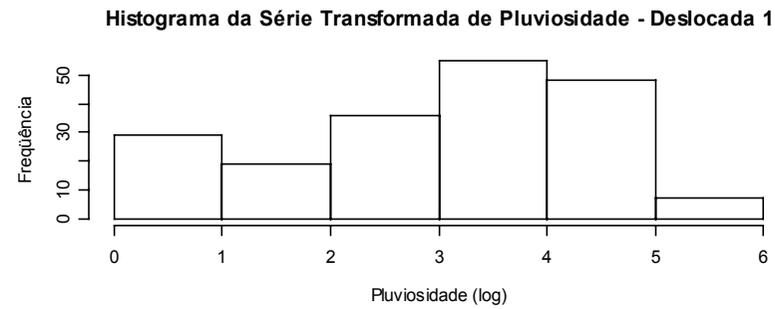
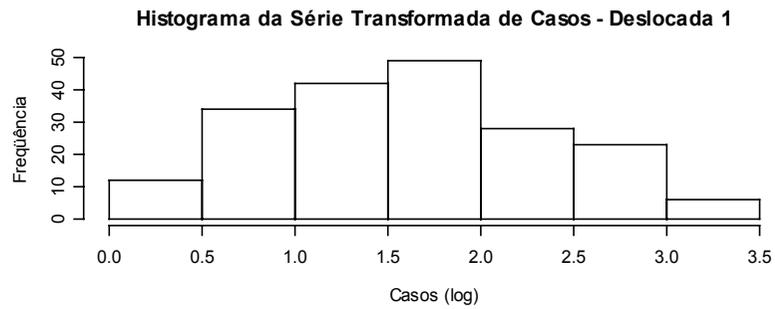
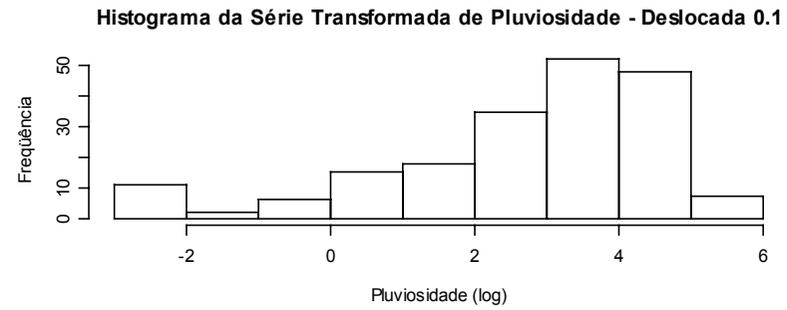
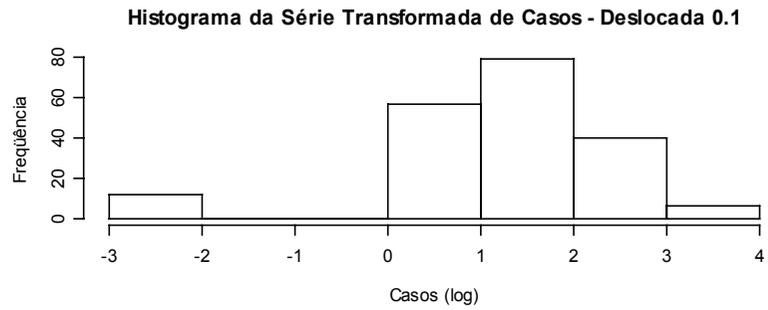
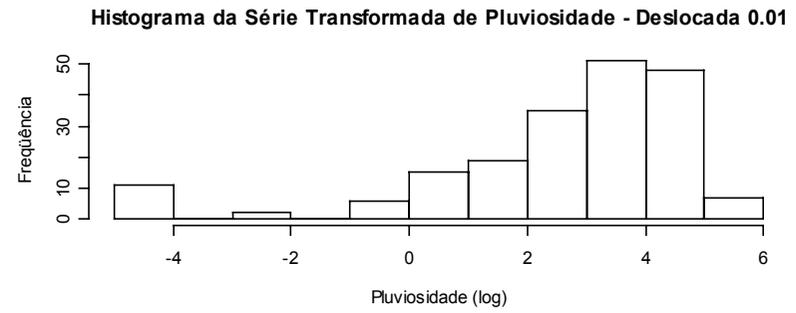
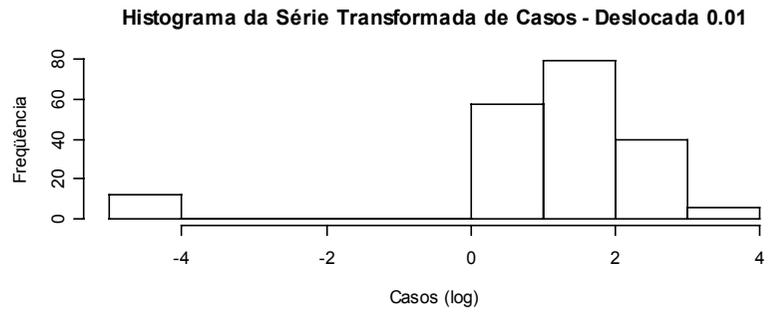
**Figura 36 – Gráficos Temporais dos Modelos – Observações, Curva Ajustada com IC 95% e Previsão com IC 95% para 12 Semanas (separados pela linha vertical) e 4 Observações Reais (quadrados vazados).  
 Acima: Modelo SARIMA (1,1,1)x(0,1,1) com Temperatura Mínima (0,1,1) como Regressor Dinâmico.  
 Abaixo: Modelo SARIMA (1,1,1)x(0,1,1) com Temperatura Média (0,1,1) como Regressor Dinâmico.**



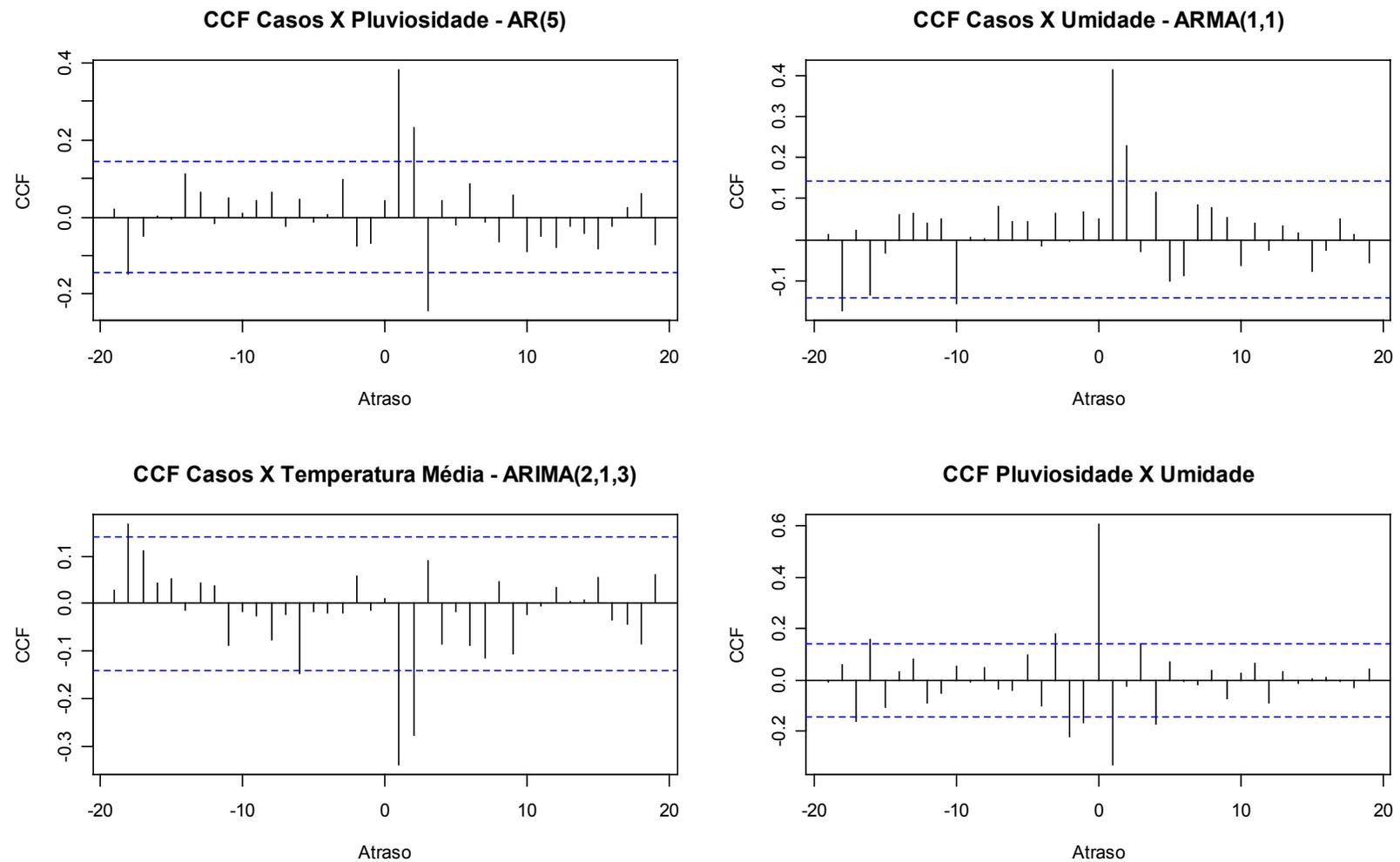
**Figura 37 – Gráficos da Variação das Estimativas de Pontos Futuros em Relação às Observações Reais em Percentual – Resultado de 18 Diferentes Modelos, Agrupados pelos Diferentes Modelos de Casos.**



**Figura 38 – Gráficos dos Comportamentos dos Erros-Padrão em 18 Modelos Diferentes. Gráfico Superior: Agrupados pelos Diferentes Modelos de Casos. Gráfico Inferior: Detalhe dos Modelos sem Regressores e com Pluviosidade.**



**Figura 39 – Histogramas das Séries de Casos e Pluviosidade Transformadas pela Função Logarítmica, Deslocadas de 0.01, 0.1 e 1 Unidade.**



**Figura 40 – Correlogramas das CCFs e IC 95% para: (a) Resíduos dos Casos e Pluviosidade Ajustados por AR(5); (b) Resíduos dos Casos e Umidade Ajustados por ARMA(1,1); (c) Resíduos dos Casos e Temperatura Média Ajustados por ARIMA(2,1,3); (d) Resíduos da Diferenciação da Pluviosidade e Umidade. As séries de casos e pluviosidades usadas nestes correlogramas são as originais, não transformadas.**