CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

USO DE TÉCNICAS DE MUDANÇA DE PERIODICIDADE DE SÉRIES TEMPORAIS PARA PREVISÃO DA TEMPERATURA DE SUPERFÍCIE DO MAR DO OCEANO ATLÂNTICO

Patrícia Mattos Teixeira

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Tecnologia, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientadores

Leonardo Silva de Lima

Eduardo Soares Ogasawara

Rio de Janeiro

Maio de 2014

USO DE TÉCNICAS DE MUDANÇA DE PERIODICIDADE DE SÉRIES TEMPORAIS PARA PREVISÃO DA TEMPERATURA DE SUPERFÍCIE DO MAR DO OCEANO ATLÂNTICO

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Tecnologia, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Patrícia Mattos Teixeira

Aprovada por:

Presidente, Prof. Leonardo Silva de Lima, D.Sc. (orientador)

Prof. Eduardo Soares Ogasawara, D.Sc. (co-orientador)

Prof. Rafael Garcia Barbastefano, D.Sc.

Profª.Aline Guimarães Monteiro, D.Sc.

Prof. Daniel Cardoso Moraes de Oliveira, D.Sc. (UFF)

Rio de Janeiro

Maio de 2014

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

T266 Teixeira, Patrícia Mattos
Uso de técnicas de mudança de periodicidade de séries temporais para previsão da temperatura de superfície do mar do Oceano Atlântico / Patrícia Mattos Teixeira.—2014.
x, 63f. + apêndice : il.color. , grafs. , tabs. ; enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2014.
Bibliografia : f. 61-63
Orientadores : Leonardo Silva de Lima
Eduardo Soares Ogasawara

1. Oceano – Temperatura – Atlântico, Oceano. 2. Análise de séries temporais. 3. Redes neurais (Computação). I. Lima, Leonardo Silva de (Orient.). II. Ogasawara, Eduardo Soares (Orient.). III. Título.

CDD 551.52463

RESUMO

USO DE TÉCNICAS DE MUDANÇA DE PERIODICIDADE DE SÉRIES TEMPORAIS PARA PREVISÃO DA TEMPERATURA DE SUPERFÍCIE DO MAR DO OCEANO ATLÂNTICO

Patrícia Mattos Teixeira

Orientadores:

Leonardo Silva de Lima

Eduardo Soares Ogasawara

Resumo da Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Tecnologia, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Eventos ambientais extremos como as secas afetam a vida de bilhões de pessoas em todo o mundo. Embora não se possa impedir que este tipo de evento aconteça, a sua previsão possibilita mitigar parte do dano resultante de sua ocorrência. Uma das variáveis importantes para identificação de ocorrências de secas é a temperatura de superfície do mar (TSM). No Oceano Atlântico Tropical, os dados de TSM são coletados e providos pelo projeto *Pilot Research Moored Array in the Tropical Atlantic* (PIRATA), que é uma rede de observação composta por boias dispostas nesta região. Este trabalho compara a previsão de TSM utilizando duas técnicas de previsão diferentes: os modelos ARIMA e as redes neurais. Além disso, são utilizados dados de TSM em três frequências distintas, isto é, diária, semanal e mensal, a fim de avaliar os impactos da transformação de dados nos horizontes de previsão de TSM. Os experimentos computacionais apresentaram resultados positivos e apontam que a utilização da série em baixa frequência traz uma melhora de 71% na previsão de longo prazo.

Palavras-Chave:

ARIMA; Redes Neurais; Temperatura de Superfície do Mar

Rio de Janeiro

Maio de 2014

ABSTRACT

USE OF FREQUENCY CHANGE TECHNIQUES OF TIME SERIES FOR PREDICTION OF SEA SURFACE TEMPERATURE OF THE ATLANTIC OCEAN

Patrícia Mattos Teixeira

Advisor(s):

Leonardo Silva de Lima

Eduardo Soares Ogasawara

Abstract of dissertation submitted to Programa de Pós-graduação em Tecnologia - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

Extreme environmental events such as droughts affect billions of people around all the world. Although it is not possible to prevent this type of event happens, its prediction enables mitigate some of the damage resulting from its occurrence. One important variable for identifying occurrences of droughts is the sea surface temperature (SST). In the Tropical Atlantic Ocean, the SST data are collected and provided by the Pilot Research Moored Array in the Tropical Atlantic (PIRATA) Project, which is an observation network composed of buoys arranged in this region. This study aims compares the prediction of SST using two different prediction techniques: ARIMA models and neural networks. Furthermore, SST data are used in three different frequencies of time, ie, daily, weekly and monthly, in order to evaluate the impact of data transformation in forecasting SST horizons. The computational experiments showed positive results and suggest that the use of low-frequency series brings an improvement of 71% in the long-term forecast.

Keywords:

ARIMA; Neural Network; Sea Surface Temperature

Rio de Janeiro

May 2014

Sumário

Capítulo I - Introdução 1
Capítulo II - As secas 6
II.1 As secas no mundo 6
II.2 As secas no Brasil 7
Capítulo III - Temperatura de superfície do mar 11
III.1 TSM e as secas 11
III.2 TSM e o clima da região amazônica 12
III.3 TSM e os ciclones tropicais 13
III.4 TSM e as mudanças climáticas 14
III.5 O projeto PIRATA 14
Capítulo IV - Metodologia de previsão 18
IV.1 Séries temporais 19
IV.2 Técnicas de pré-processamento dos dados 20
IV.2.1 Limpeza dos dados 21
IV.2.2 Transformação da série em estacionária 21
IV.2.2.1 Diferenciação 21
IV.2.2.2 Remoção de tendência 22
IV.2.3 Mudança de periodicidade 22
IV.3 Modelos Auto-regressivos Integrados de Médias Móveis (ARIMA) 22
IV.3.1 Modelos AR 23
IV.3.2 Modelos MA 23
IV.3.3 Modelos ARMA 23
IV.3.4 Modelos ARIMA 24
IV.3.5 Metodologia *Box-Jenkins* 25
IV.3.5.1 Identificação do modelo 25
IV.3.5.2 Estimação do parâmetro 26
IV.3.5.3 Checagem do diagnóstico 26
IV.3.5.4 Previsão 26

IV.4 Modelos baseados em Redes Neurais Artificiais 27
IV.4.1 Perceptron de múltiplas camadas 28
IV.4.2 Modelagem de uma Rede Neural Artificial 29
IV.4.2.1 Arquitetura da rede 29
IV.4.2.2 Função de ativação 30
IV.4.2.3 Algoritmo de treinamento 31
IV.4.2.4 Particionamento dos dados 31
IV.5 Medidas de desempenho 32
Capítulo V - Experimentos computacionais de previsão para TSM 34
V.1 Preparação do experimento 34
V.1.1 Seleção e limpeza dos dados 35
V.1.2 Mudança de periodicidade 37
V.1.3 Transformação da série em estacionária 38
V.1.4 Particionamento dos dados 39
V.2 Experimento comparativo entre os modelos RNA e ARIMA 40
V.2.1 Modelo RNA 41
V.2.2 Modelo ARIMA 43
V.2.3 Análise dos resultados 44
V.2.3.1 Previsões utilizando dados diários 44
V.2.3.2 Previsões utilizando dados semanais 45
V.2.3.2 Previsões utilizando dados mensais 46
V.3 Modelos ARIMA 48
V.3.2 Previsão utilizando dados diários 49
V.3.3 Previsão utilizando dados semanais 51
V.3.4 Previsão utilizando dados mensais 52
V.3.5 Análise dos resultados 54
Capítulo VI - Conclusão 58
Referências Bibliográficas 61
Apêndice I - Mecanismo de treinamento por *backpropagation* 64

Lista de Figuras

Figura II.1 Regiões que foram afetadas pelas secas no Brasil e em outros países da América do Sul durante o período de 1980 a 2011. 8

Figura II.2 A posição da ZCIT (representada pela linha verde) nos meses de abril de 2011, quando se encontra mais ao sul (a) e setembro, quando se localiza mais ao norte (b). 9

Figura III.1 A região de localização das boias do PIRATA 16

Figura IV.1 Gráfico da função cosseno, exemplo de série estacionária. 20

Figura IV.2 Exemplo de uma rede neural artificial do tipo MLP e suas camadas..... 28

Figura V.1 Séries temporais das boias mais próximas à costa brasileira (a) 0ºN 35ºW, (b) 4ºN 38ºW, (c) 8ºS 34ºW, (d) 14ºS 32ºW (e) 19ºS 34ºW. 36

Figura V.2 Série de TSM da boia 19ºS 34ºW após os procedimentos de seleção e limpeza de dados. 37

Figura V.3 Arquitetura da rede do modelos para as séries dos dados diários(a), semanais(b) e mensais(c). 42

Figura V.4 Comparação entre as séries resultantes das previsões utilizando dados diários e a série real de TSM. 45

Figura V.5 Comparação entre as séries resultantes das previsões utilizando dados semanais e a série real de TSM. 46

Figura V.6 Comparação entre as séries resultantes das previsões utilizando dados mensais e a série real de TSM. 47

Figura V.7 Função de autocorrelação da amostra de dados de treinamento dos dados diários(a), semanais(b) e mensais(c). 49

Figura V.8 Comparação entre a série resultante da previsão utilizando o modelo ARIMA com dados diários e a série real de TSM. 50

Figura V.9 Comparação entre a série resultante da previsão utilizando o modelo ARIMA com dados semanais e a série real de TSM. 52

Figura V.10 Comparação entre a série resultante da previsão utilizando o modelo ARIMA com dados mensais e a série real de TSM. 53

Figura V.11 Comparação entre a série real de TSM e as séries resultantes das previsões utilizando o modelo ARIMA com dados diários e dados semanais. 55

Figura V.12 Comparação entre a série real de TSM e as séries resultantes das previsões utilizando o modelo ARIMA com dados diários e dados mensais..... 56

Figura Apêndice I.1: Funcionamento de um neurônio j 64

Figura Apêndice I.2: Exemplo de uma rede com fluxo *feed forward*..... 65

Figura Apêndice I.3: Exemplo de uma rede com fluxo *feed backward* 67

Lista de Tabelas

Tabela II.1 Estatísticas sobre as ocorrências de secas no mundo no período de 1980 a 2011 6

Tabela II.2 Estatísticas sobre as ocorrências de secas no Brasil no período de 1980 a 2011 7

Tabela V.1 Número de observações das séries diárias, semanais e mensais 38

Tabela V.2 Particionamento dos dados das séries de dados diários, semanais e mensais entre os conjuntos de teste e de treinamento 39

Tabela V.3 Estatística descritiva do conjunto de treinamento correspondente ao período de 6 anos para os dados diários, semanais e mensais. 40

Tabela V.4 Arquitetura das Redes Neurais Artificiais para os dados diários, semanais e mensais. 41

Tabela V.5 Configurações das Redes Neurais Artificiais para os dados diários, semanais e mensais. 43

Tabela V.6 Resumo dos resultados das previsões utilizando dados diários. 44

Tabela V.7 Resumo dos resultados das previsões utilizando dados semanais. 46

Tabela V.8 Resumo dos resultados das previsões utilizando dados mensais. 47

Tabela V.9 Configurações do modelo ARIMA para os dados diários 50

Tabela V.10 Resultados do modelo ARIMA utilizando dados diários. 51

Tabela V.11 Configurações do modelo ARIMA utilizando dados semanais. 51

Tabela V.12 Resultados do modelo ARIMA utilizando dados semanais. 52

Tabela V.13 Configurações do modelo ARIMA utilizando dados mensais. 53

Tabela V.14 Resultados do modelo ARIMA utilizando dados mensais. 54

Tabela V.15 Comparação entre os resultados do modelo ARIMA utilizando dados diários e dados semanais. 55

Tabela V.16 Comparação entre os resultados do modelo ARIMA utilizando dados diários e mensais. 56

Capítulo I - Introdução

Na natureza ocorrem diversos fenômenos que afetam a vida de bilhões de pessoas, como é o caso dos chamados eventos extremos. Apesar de não haver um consenso com relação a sua definição, um evento natural é caracterizado como extremo baseando-se normalmente em três critérios: raridade, intensidade e severidade (BENISTON *et al.*, 2007). O primeiro critério diz respeito à frequência em que um evento ocorre, enquanto a intensidade refere-se a eventos caracterizados por valores relativamente altos ou baixos, isto é, que possuam grande variação em relação ao valor normal e, por fim, o critério severidade considera os impactos socioeconômicos de um evento, ou seja, os prejuízos causados à sociedade (BENISTON *et al.*, 2007).

A despeito de haver uma forte associação entre eventos extremos e desastres naturais, a ocorrência de um evento extremo não resulta, necessariamente, em um desastre. Para que a ocorrência de eventos extremos incorra em desastres naturais são necessárias duas condições: (i) a exposição de comunidades a tais eventos; (ii) tal exposição aos eventos extremos esteja acompanhada de um nível elevado de vulnerabilidade, isto é, a predisposição à perda e ao dano (FIELD, 2012). Dessa forma, para que um evento seja caracterizado como um desastre é necessário que haja o envolvimento de seres humanos e que estes sofram algum tipo de impacto em virtude da sua ocorrência.

Dentre os eventos extremos que tiveram maior impacto sobre a sociedade, as secas, os ciclones e as enchentes (ou inundações) destacam-se como os três desastres naturais que mais causam danos. Juntos estes eventos causaram um prejuízo estimado em 1,22 trilhões de dólares e afetaram cerca de 4,7 bilhões de pessoas, causando a morte de mais de um milhão de pessoas em todo o mundo desde 1980, de acordo a pesquisa realizada no banco de dados de desastres EM-DAT (EM-DAT, 2013).

As secas se distinguem por causarem impactos mais severos, uma vez que ocasionaram um número maior de mortes quando comparadas aos demais desastres naturais, tendo causado cerca de 11,7 milhões de fatalidades entre 1900 a 2013, mais do que juntos causaram os terremotos, ciclones tropicais e inundações (EM-DAT, 2013). Este evento extremo pode ser definido como uma condição de umidade insuficiente causada por uma redução da quantidade de precipitação recebida durante um determinado período de tempo (MCKEE; DOESKEN; KLEIST, 1993; MISHRA; SINGH, 2010), que é variável de acordo com cada caso, podendo durar vários meses, estações do ano ou mesmo anos.

O crescimento da população mundial e o consequente aumento da demanda por água - principalmente devido à expansão dos setores de energia, agrícola e industrial - têm tornado a seca um problema ainda mais preocupante (MISHRA; SINGH, 2010), uma vez que a escassez

de água já é uma realidade para muitas regiões do mundo e afeta milhões de pessoas todos os anos.

Um agravante a ser considerado quando se trata de secas é a dificuldade de avaliá-las e prevê-las, dado que a avaliação e previsão de secas, segundo Svoboda *et al*. (2002), podem ser consideradas mais difíceis do que as de outros desastres naturais. Entretanto, apesar da complexidade inerente ao processo de avaliação e previsão de secas, visto que sua duração, extensão e impacto variam de uma região pra outra, as estratégias e princípios implícitos ao gerenciamento do risco de seca são semelhantes e independem das situações específicas de cada país (TADESSE et al., 2004).

Há uma grande quantidade de dados dos diversos índices e variáveis climáticos e hidrológicos existentes, assim como uma ampla diversidade de escala de tempo e espaço que precisam ser considerados na modelagem de previsão de secas (TADESSE *et al.*, 2004), sendo a identificação de preditores efetivos um componente principal para um modelo de previsão (FAROKHNIA; MORID; BYUN, 2011). E, neste contexto, que a chamada Temperatura de Superfície do Mar (TSM) destaca-se como um preditor importante para previsões de longo prazo de secas.

A temperatura de superfície do mar é a temperatura que a água do mar apresenta na sua proximidade com a atmosfera, isto é, quando medida em pequena profundidade. A TSM é considerada um dos fatores mais importantes no que diz respeito ao condicionamento da variabilidade climática e de eventos extremos, inclusive das secas, na qual sua relevância para a elaboração de modelos de previsão é ressaltada por Mishra e Singh (2011). A TSM desempenha um papel bastante significativo no sistema climático no sudoeste do Atlântico Tropical, região que abrange a costa do Nordeste do Brasil, uma vez que a sua variabilidade modula padrões de anomalias de precipitação, resultando nas secas nesta região (LINS *et al.*, 2013).

Além das secas da região Nordeste brasileira, há fortes indícios de que a TSM exerça um papel decisivo em outros fenômenos ocorridos no Oceano Atlântico Tropical, dentre os quais destacamos: (i) o processo que dá origem aos ciclones tropicais no Oceano Atlântico (SUKOV *et al.*, 2008); (ii) as chuvas na região Amazônica (FU *et al.*, 2001); (iii) a quantidade verde da Amazônia (CHO *et al.*, 2010) e também (iv) o sequestro de carbono no oceano (GRUBER; KEELING; BATES, 2002).

O desenvolvimento deste estudo foi motivado pela importância da TSM para o sistema oceano-atmosférico, bem como pela dificuldade associada ao processo de previsão de eventos extremos.

Dessa forma, o presente estudo visa contribuir para o aprimoramento das previsões de eventos extremos e suas variáveis relacionadas através da geração de dados de temperatura de superfície do mar baseada em dados passados a partir do uso de dois modelos de previsão

já consolidados na literatura: as Redes Neurais Artificiais (RNAs) e os modelos Auto-regressivos de Médias Móveis (ARIMA). Além disso, uma especial atenção é atribuída ao tratamento dos dados antes de serem utilizados para alimentarem os modelos de previsão, através das chamadas técnicas de pré-processamento, entre as quais se destaca a mudança de periodicidade dos dados.

A mudança de periodicidade tem como objetivo identificar se a previsão de uma série transformada pode trazer benefícios frente à previsão direta com a série sem nenhuma intervenção, principalmente na avaliação dos possíveis impactos que estas mudanças podem causar em horizontes de previsão mais extensos.

A metodologia da pesquisa leva em consideração a implementação e avaliação dos dados gerados pelos modelos de previsão de RNA e ARIMA. As implementações foram realizadas nas ferramentas WEKA, para a RNA, e Matlab, para o ARIMA. Os dados de entrada para os modelos são obtidos *online* no sítio do projeto *Pilot Research Moored Array in the Tropical Atlantic* (PIRATA), que disponibiliza dados de diferentes variáveis do Oceano Atlântico, como a temperatura do mar em diversas profundidades, salinidade do mar, intensidade do vento, radiação solar, precipitação, umidade relativa e temperatura do ar.

Considerando os impactos causados por um evento extremo, esta pesquisa justifica-se pelo fato de que todo estudo que procure contribuir para melhorar as técnicas de monitoramento e previsão desses eventos estará colaborando para que se consiga reduzir, dentro do que for possível, a vulnerabilidade da sociedade como um todo perante os eventos extremos e os seus indesejáveis efeitos.

Portanto, o estudo desenvolvido nesta dissertação tem como principal objetivo contribuir para a mitigação dos impactos de eventos extremos através da realização de previsões acuradas da temperatura de superfície do mar, fazendo uso de técnicas para mudança de periodicidade dos dados de TSM, a fim de verificar os benefícios que podem ser obtidos ao serem utilizados dados de menores frequências, e de modelos de previsão redes neurais artificiais e ARIMA, permitindo assim a comparação entre os resultados obtidos pelas duas técnicas ao serem utilizados os dados em diferentes periodicidades. Este estudo mostra-se, portanto, interessante tanto do ponto de vista tecnológico quanto social e ambiental.

Os estudos relacionados são importantes para o desenvolvimento de uma pesquisa, na medida em que mostram a relevância do que está sendo proposto perante a comunidade científica e, além disso, se mostram úteis, pois ajudam a guiar novas pesquisas, seja pelas contribuições metodológicas ou de qualquer outro cunho, que seja relevante para uma pesquisa.

Outros estudos também realizaram previsões de temperatura de superfície do mar, no entanto poucos trabalhos realizados sobre a TSM do oceano Atlântico Tropical foram encontrados, sendo mais comum encontrar estudos envolvendo a TSM do oceano Pacífico,

como Wu *et al.* (2006), que utiliza as redes neurais para realizar previsões de TSM do oceano Pacífico Tropical. Aguilar-Martinez e Hsieh (2009) também realizam previsões de TSM do oceano Pacífico e, assim como o estudo apresentado nesta dissertação, utilizam duas técnicas diferentes para comparar seus resultados de previsão, sendo a primeira o método conhecido como *Support Vector Machine* (SVM) e a segunda técnica as redes neurais.

Lins *et al.* (2013) também utiliza a técnica SVM para realizar previsões de TSM e, assim como nesta pesquisa, utilizada dados de TSM do oceano Atlântico Tropical para realiza previsões diárias desta mesma variável para o horizonte de um ano à frente. Neste estudo de Lins *et al.* (2013) são definidos cenários distintos e o modelo assume que há uma relação entre o dado previsto e os dados do dia anterior ao desejado de anos anteriores, utilizando tais dados como entrada para este modelo.

Outro estudo que se assemelha ao que é proposto nesta dissertação é o realizado por Ho *et al.* (2002), em que as técnicas redes neurais artificiais e o ARIMA têm seus desempenhos de previsão comparados. Ho *et al.* (2002) procura avaliar o desempenho de modelos de séries temporais como ferramenta para análise de falhas de sistema reparáveis. São comparados resultados ao utilizar redes neurais do tipo *feed-forward perceptron* de múltiplas camadas e redes neurais do tipo parcialmente recorrentes e ARIMA em previsões tanto para curto prazo quanto para longo prazo.

Já no estudo realizado por Chattopadhyay (2007) utilizou-se as redes neurais artificiais também para realizar previsões um evento extremo, conhecido por monções. Neste estudo, as redes neurais são utilizadas para prever média das chuvas de monções de verão na Índia, que utiliza como entrada a temperatura de superfície do mar, entre outras variáveis, e depois compara os resultados obtidos pelas RNAs com os resultados de dois métodos estatísticos, a regressão linear múltipla e a persistência.

No entanto, é importante ressaltar que, apesar de haver estudos semelhantes ressaltando a importância de se prever a temperatura de superfície do mar, nenhum estudo de previsão de temperaturas de superfície do mar do Oceano Atlântico Tropical que contemple previsões utilizando dados de diferentes periodicidades foi encontrado durante a pesquisa bibliográfica realizada ao longo deste trabalho.

Os cinco próximos capítulos desta dissertação compreendem uma abordagem sobre os principais conceitos necessários ao entendimento da contribuição científica proposta por este estudo. O Capítulo II retrata as secas com mais detalhes, levantando os impactos deste evento extremo no mundo e, em especial, no Brasil, onde são contempladas suas possíveis causas. Enquanto o Capítulo III aborda o papel desempenhado pela TSM tanto na ocorrência de secas como em outros eventos extremos. Além disso, ainda no Capítulo III é realizado um breve apanhado sobre o programa PIRATA, responsável por capturar e disponibilizar dados de variáveis de grande relevância para as pesquisas no oceano Atlântico Tropical. O Capítulo IV

apresenta os principais conceitos relacionados às técnicas de pré-processamento de dados e às redes neurais artificiais e aos modelos ARIMA como modelos para a previsão de TSM, abordando suas principais características, bem como as métricas que auxiliam na percepção do desempenho destes modelos. Já o Capítulo V procura explicitar os experimentos computacionais realizados, detalhando as técnicas de pré-processamento utilizadas e os modelos propostos e discutindo e analisando os resultados obtidos a partir dos experimentos realizados com as séries de diferentes periodicidades. Finalmente, nas Conclusões mostra-se que os objetivos do trabalho foram atingidos e direções futuras de pesquisa são indicadas.

Capítulo II - As secas

As secas ocorrem em diversas regiões do mundo e trazem com elas consequências de grande escala para sociedade e, por isso, muitos estudos podem ser encontrados sobre esse fenômeno, suas características e seus impactos. Dentre os principais autores que serviram de base para o desenvolvimento deste capítulo estão Mishra e Singh (2010), que em seu estudo fizeram uma revisão dos conceitos mais importantes sobre secas; Hastenrath (2012), que abordou os problemas climáticos enfrentados pelo Brasil desde o início da sua colonização; Moura e Shukla (1981) e Liu e Juarez (2001), que também falam dos problemas das secas no Brasil, abordando as suas causas; e, ainda, Servain *et al.* (1998) e Bourlès *et al.* (2008) contribuem ao ressaltar as características e motivações do projeto PIRATA.

A fim de proporcionar um claro entendimento sobre o problema das secas, este capítulo busca, primeiramente, explicitar a importância de seu estudo, levando em conta os problemas inerentes a esse evento climático. A seguir, é feito um levantamento sobre as consequências das ocorrências de secas no mundo e também no Brasil, onde se procura investigar as causas das secas na região Nordeste do país.

II.1 As secas no mundo

A ocorrência de secas se dá em diferentes regiões do mundo, abrangendo dezenas de países e causando impactos severos na vida de milhões de pessoas. A natureza desses impactos é tanto econômica quanto social, na medida em que a escassez de água afeta diretamente a qualidade de vida da população bem como prejudica atividades primárias da economia como a agricultura, por exemplo, fazendo com que medidas governamentais para mitigar os seus efeitos sejam requeridas. A gravidade das consequências das secas fica mais evidente ao se considerar que apenas nas últimas três décadas mais de 1,75 bilhões de pessoas foram afetadas em todo mundo, causando um prejuízo estimado em 4,36 trilhões de dólares (EM-DAT, 2013), conforme mostra a Tabela II.1.

Tabela II.1 Estatísticas sobre as ocorrências de secas no mundo no período de 1980 a 2011

Nº de ocorrências	477
Nº pessoas afetadas	1.744.602.166
Nº de óbitos	558.540
Prejuízo estimado (US$ Milhões)	4.363.266

Fonte: EM-DAT (2013)

A seriedade dos problemas envolvendo esse desastre natural e a abrangência de suas consequências tornam as secas um objeto de estudo importante para cientistas de variadas áreas, como ambientalistas, ecologistas, meteorologistas, geólogos e economistas, entre outros (MISHRA; SINGH, 2010). Assim como o presente estudo, as pesquisas nessa área procuram oferecer maneiras mais eficientes de lidar com as ocorrências desses desastres naturais, de forma a se conseguir mitigar ao máximo seus efeitos.

II.2 As secas no Brasil

No Brasil, os registros de secas e suas consequências socioeconômicas existem desde o início da colonização portuguesa, mas foi no século XVII que os governos começaram a tomar iniciativas para mitigar seus efeitos (HASTENRATH, 2012).. Estima-se que 32,8 milhões de pessoas tenham sido afetadas pelas secas nos últimos trinta anos e que um prejuízo de aproximadamente 2,4 bilhões de dólares tenha sido proporcionado ao país (EM-DAT, 2013). O resumo destes dados é apresentado no Tabela II.2.

Tabela II.2 Estatísticas sobre as ocorrências de secas no Brasil no período de 1980 a 2011

Nº de ocorrências	12
Nº pessoas afetadas	32.812.000
Nº de óbitos	20
Prejuízo estimado (US$ Milhões)	2.423

Fonte: EM-DAT (2013)

As secas exerceram também um papel importante em alguns eventos que marcaram a história do país, pois foi principalmente devido à sua ocorrência que a população do Nordeste migrou para outras regiões do país à procura de trabalho e melhores condições de vida (HASTENRATH, 2012).

Apesar de as secas afetarem grande parte do continente brasileiro, a região do Nordeste do Brasil é a mais prejudicada, sendo inclusive considerada uma região com propensão à secas (HASTENRATH, 2012; LIU; JUÁREZ, 2001), como mostra a região alaranjada na Figura II.1 Nessa região, as chuvas são raras e, em boa parte dela, costumam ocorrer apenas em um pequeno período do ano, mais especificamente nos meses de março e abril (HASTENRATH, 2012; MOURA; SHUKLA, 1981).

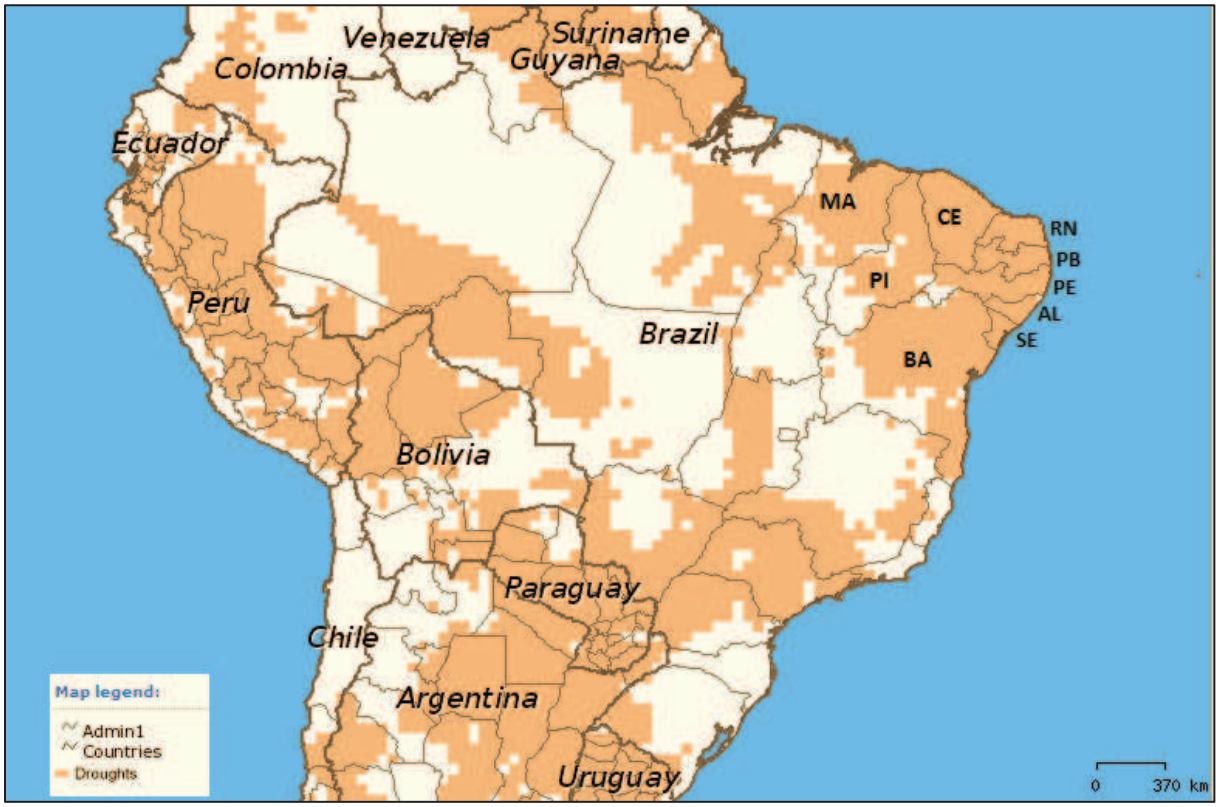


Figura II.1 Regiões que foram afetadas pelas secas no Brasil e em outros países da América do Sul durante o período de 1980 a 2011.

Fonte: Adaptado de PREVIEW Global Risk Data Plataform (2013).

As causas para que a região Nordeste seja tão severamente atingida pelas secas vêm sendo estudadas há bastante tempo e, atualmente, existe uma concordância no que diz respeito aos principais fatores que contribuem para a ocorrência desse desastre natural. Os fenômenos conhecidos por *El Niño* Oscilação Sul (ENOS) e Zona de Convergência Intertropical (ZCIT) foram fortemente associados à ocorrência de secas no Nordeste por diversos estudos realizados nessa região (DURAND *et al.*, 2005; HASTENRATH, 2012; LIU; JUÁREZ, 2001; MOURA; SHUKLA, 1981).

O *El Niño* é uma corrente oceânica quente que ocorre ao longo da costa da América do Sul, desde o Equador até o Peru (LIU; JUÁREZ, 2001). E é “caracterizado pelas temperaturas do oceano excepcionalmente quentes no Pacífico Equatorial” (NOAA, 2013a). Como o componente atmosférico empregado para investigar o *El Niño* chama-se Oscilação Sul, o fenômeno que envolve tanto os componentes atmosféricos quanto os oceânicos é conhecido como *El Niño* Oscilação Sul – ENOS (LIU; JUÁREZ, 2001).

Alguns estudos procuram mostrar que os anos em que *El Niño* acontece, há também a ocorrência de secas no Nordeste, demonstrando assim que há algum tipo de relacionamento entre os eventos, como apontam Liu e Joarez (2001) e Durand *et al.* (2005) . Entretanto, não é

somente o fenômeno *El Niño* o responsável pela ocorrência das secas na região nordestina do Brasil, há também outros fatores que contribuem para as secas, como a Zona de Convergência Intertropical.

De acordo com a Fundação Cearense de Meteorologia e Recursos Hídricos (FUNCEME), um dos órgãos brasileiros mais respeitados de pesquisa sobre a ocorrência de secas na região Nordeste, a Zona de Convergência Intertropical é o sistema meteorológico mais importante no que concerne à indicação do volume de chuvas no Nordeste. A ZCIT, ainda segundo a FUNCEME, é “uma banda de nuvens que circunda a faixa equatorial do globo terrestre, formada principalmente pela confluência dos ventos alísios do hemisfério norte com os ventos alísios do hemisfério sul”.

A ZCIT movimenta-se ao longo do ano e migra sazonalmente mais para o norte do oceano Atlântico no período de agosto a setembro, enquanto no período de março a abril migra mais para o sul do oceano (HASTENRATH, 2012), conforme mostra a Figura II.2. A relação entre ZCIT e as secas no Nordeste se dá a partir dessa movimentação da zona de convergência ao longo do ano, de modo a favorecer a ocorrência de chuvas no momento em que a zona de convergência se encontra posicionada mais ao sul e tornar mais propícia à ocorrência de secas quando posicionada ao norte, principalmente se tal posicionamento ao norte ocorrer no período correspondente aos meses de março e abril, que é quando espera-se que a ZCIT esteja localizada ao sul (DURAND *et al.*, 2005; HASTENRATH, 2012).

(a) (b)

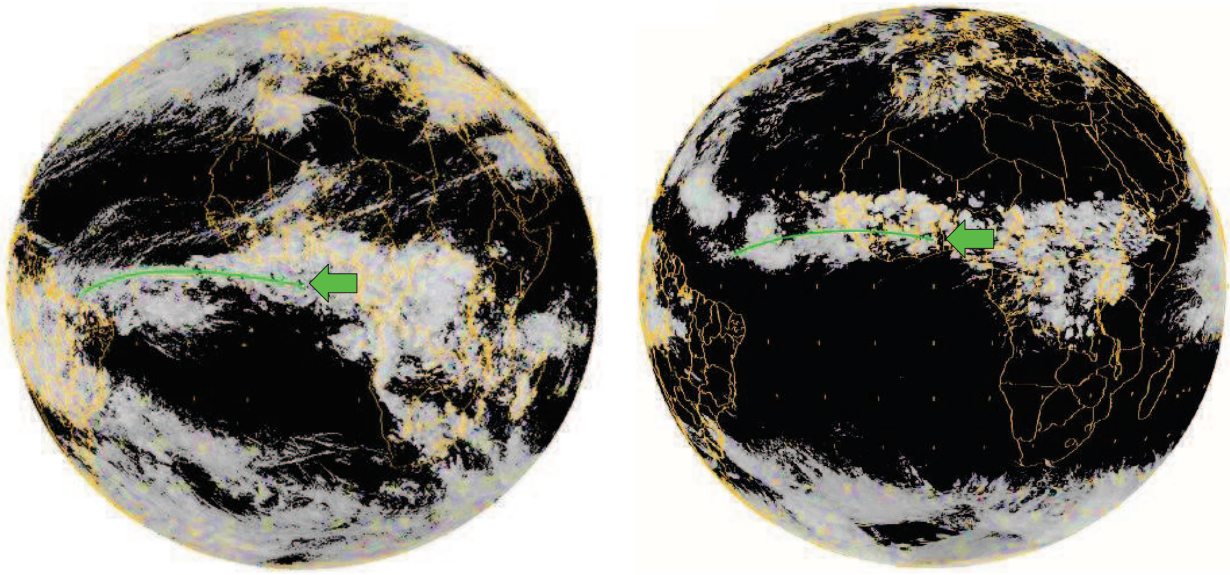


Figura II.2 A posição da ZCIT (representada pela linha verde) nos meses de abril de 2011, quando se encontra mais ao sul (a) e setembro, quando se localiza mais ao norte (b).
Fonte: FUNCEME (2013)

A ZCIT é mais significativa sobre os oceanos e por isso as suas movimentações e sua localização são fortemente afetadas pela temperatura de superfície do mar, que exerce um

papel importante na dinâmica oceano-atmosférica, influenciando por consequência na ocorrência ou não de chuvas na região Nordeste (HASTENRATH, 2012; MOURA; SHUKLA, 1981). A temperatura de superfície do mar, bem como o detalhamento do papel esta que exerce nas secas e também em outros fenômenos climáticos será abortado a seguir, no Capítulo III.

Capítulo III - Temperatura de superfície do mar

A temperatura de superfície do mar (TSM) é a temperatura da água do oceano em sua superfície, sendo considerada um indicador de acúmulo de calor do oceano (LINS *et al.*, 2013). Os dados de TSM utilizados nesta dissertação correspondem, mais especificamente, aos valores da temperatura da água do mar obtidos entre um metro e um metro e meio de profundidade em relação à superfície, medidos em graus Celsius.

A TSM nos oceanos tropicais Pacífico e Atlântico é a principal variável física influenciadora das condições climáticas em várias áreas do globo terrestre (ALVES; DE SOUZA; CAMPOS, 2006). A TSM do Atlântico Tropical desempenha um papel significativo no sistema climático no sudoeste deste Oceano (LINS *et al.*, 2013), região que abrange a costa do Nordeste do Brasil e também parte da região Amazônica. Esta variável é especialmente significativa para a região dos trópicos devido à sensibilidade da atmosfera às condições de superfície do continente e do oceano, sendo influenciadora, portanto, da variabilidade do clima (YOON; ZENG, 2010).

Vários artigos na literatura apontam a correlação entre a temperatura de superfície do mar e a ocorrência de diferentes eventos na natureza, entre os quais podem ser destacados as secas, os ciclones tropicais (ou furacões) e as chuvas na região amazônica (CHO *et al.*, 2010; MISHRA; SINGH, 2011; MOURA; SHUKLA, 1981; VIANNA *et al.*, 2010). Além disso, a TSM tem sido associada também à processos relacionados mais diretamente com as mudanças climáticas, como nos processos de sequestro de carbono (GRUBER; KEELING; BATES, 2002).

III.1 TSM e as secas

Em diversas partes do mundo, a precipitação tem sido associada a fenômenos atmosféricos, como o *El Niño* Oscilação Sul e Zona de Convergência Intertropical e a temperatura de superfície do mar, onde tem se observado um forte relacionamento estatístico entre esses fenômenos e as chuvas (MISHRA; SINGH, 2011), despertando o interesse dos pesquisadores em buscar modelos para realizar previsões de dados de TSM de longo prazo. Entretanto, a influência de cada um desses fatores, pode variar de uma região para outra, podendo ser mais relevantes em algumas regiões do que outras.

No Brasil, conforme fora abordado no Capítulo II, as secas na região Nordeste têm sido vinculadas mais fortemente à alteração do posicionamento da Zona de Convergência Intertropical (ZCIT). Segundo Cho *et al.* (2010), o Oceano Atlântico Tropical exerce um papel crítico na determinação do posicionamento da Zona de Convergência Intertropical, o que por sua vez afeta as chuvas em regiões equatoriais. Ainda de acordo com Cho *et al.* (2010), a TSM

pode ser entendida como um dos principais agentes do Oceano Atlântico no que concerne às alterações do posicionamento da ZCIT.

Após analisar dados de TSM em um período de 25 anos (de 1948 a 1972), Moura e Shukla (1981) observaram que a ocorrência simultânea de anomalias no TSM ao norte e ao sul da linha do Equador, de forma que ao norte o TSM apresentasse temperaturas anormalmente mais quentes e ao sul temperaturas mais frias, provocaram a intensificação e o deslocamento da ZCIT para o norte, ocasionando assim períodos de seca no Nordeste do Brasil.

A relação entre as secas na região Nordeste do Brasil e a TSM no Atlântico Tropical também é abordada por outros autores, como por Hastenrath (2012), que explica que “a região Nordeste experimenta uma curta estação chuvosa em torno de março e abril, quando o gradiente inter-hemisférico da TSM no Atlântico Tropical é mais fraco e a ZCIT atinge a sua posição mais meridional no decurso do seu ciclo anual.”.

Sendo assim, os estudos já realizados até o momento parecem indicar que a temperatura de superfície do mar é uma variável relevante para a dinâmica das secas e também de outros eventos extremos como quantidade de verde de florestas, ciclones, tempestades e outros que são mais bem detalhados a seguir.

III.2 TSM e o clima da região amazônica

No Brasil, a influência da TSM na ocorrência de eventos climáticos não se limita às secas na região do Nordeste, mas atinge também outras regiões, como a região Amazônica, no Norte do país. Fu *et al.* (2001) realizaram simulações a fim de melhor compreender a influência da sazonalidade da TSM nos oceanos Pacífico e Atlântico na precipitação da região e concluíram que a TSM de ambos oceanos tem importante participação nas chuvas da região durante os períodos de equinócio e, principalmente, durante a chamada primavera austral, isto é, a primavera no Hemisfério Sul.

Além disso, a TSM tem sido associada também à variabilidade da quantidade de verde na Amazônia. Tal fato por ser justificado devido aos ecossistemas de plantas tropicais serem especialmente sensíveis às flutuações da umidade do solo e à radiação, como é o caso da região Amazônica, que por sua vez são fortemente dependentes de condições climáticas como a quantidade de nuvens (CHO *et al.*, 2010).

O papel da TSM se faz determinante devido ao fato de, segundo Cho *et al.* (2010), esta variável poder afetar consideravelmente a quantidade de nuvens nas regiões tropicais localizadas próximas ao oceano, uma vez que pode alterar padrões e intensidade do transporte de umidade da atmosfera para o continente, afetando, a variabilidade da vegetação e da precipitação na região Amazônica.

III.3 TSM e os ciclones tropicais

O ciclone tropical é um fenômeno conhecido principalmente por sua capacidade de devastação, visto que a sua ocorrência acarreta em grandes perdas econômicas e, em muitos casos, em fatalidades. Nas últimas três décadas, cerca de meio bilhão de pessoas foram afetadas pelos ciclones tropicais, causando um prejuízo de aproximadamente 632 milhões de dólares e a morte de 412 mil pessoas (EM-DAT, 2013), o que reforça o poder de devastação dos ciclones tropicais.

Dessa forma, desenvolver modelos para prever a ocorrência e os impactos dos ciclones tropicais se mostra uma tarefa de grande valor científico e social. Sukov *et al.* (2008), por exemplo, procura desenvolver em seu estudo um algoritmo para detectar um ciclone tropical em seu estágio inicial. Para tanto, utiliza dados de características ambientais tais como a temperatura de superfície do mar, precipitação, umidade relativa do ar e salinidade.

Para que o ambiente se torne propício para a origem de um ciclone é necessário que a temperatura de superfície do mar seja superior a 26,5°C, pois dessa forma contribui para o desencadeamento do processo convectivo que origina os ciclones tropicais (BLAIR, 1964). Segundo Miguens (2000), a capacidade do ar de conter umidade é diretamente proporcional à sua temperatura, e por isso, nas regiões onde a temperatura de superfície do mar seja mais elevada, o ar é mais quente e úmido. O ar mais quente também é menos denso, o que faz com que o ar quente e úmido suba e se encontre com o ar mais frio, característico de camadas atmosféricas mais elevadas. Como consequência do encontro entre o ar quente e úmido com o ar frio, tem-se a condensação da umidade. O processo anteriormente descrito é chamado de convecção. A condensação libera calor, que contribui para o aquecimento da massa de ar que está seguindo o mesmo processo descrito anteriormente e, dessa forma, o processo de convecção vai sendo intensificado. A importância da TSM neste contexto se deve ao fato de que o processo de convecção só consegue ser mantido em regiões onde a TSM seja superior a 26,5°C e, por isso, a TSM pode ser entendida como um elemento influenciador para formação de ciclones tropicais.

Os ciclones tropicais são formados principalmente sobre o oceano tropical e afetam regiões próximas à linha do equador com maior frequência. Este desastre natural recebe diferentes denominações de acordo com a região em que ocorre, podendo ser chamado de furacão, tufão ou apenas ciclone. No Oceano Atlântico os ciclones costumam acontecer no Hemisfério Norte, tendo havido um único registro no Hemisfério Sul conhecido como furacão Catarina (VIANNA *et al.*, 2010).

O furacão Catarina ocorreu em 2004, próximo à costa da região Sul do Brasil, afetando cerca de 150 mil pessoas e causando um prejuízo estimado em 350 milhões de dólares (EM-DAT, 2013). Um aspecto interessante com relação à formação deste fenômeno é que, ao

contrário do que usualmente ocorre, o furacão Catarina se formou com temperaturas de superfície do mar de 24ºC, fato que instigou muitos pesquisadores. De acordo com Vianna *et al.* (2010), uma explicação para a interação entre a TSM no Atlântico Sul e a formação do Catarina pode ser atribuída parcialmente à considerável diferença entre os valores de temperatura do ar e da superfície do mar.

III.4 TSM e as mudanças climáticas

Mais recentemente a TSM também tem sido entendida como um indicador das mudanças climáticas. Apesar da temperatura média global na superfície do planeta muitas vezes ser abordada como se fosse o único indicador de mudanças climáticas, há outros indicadores baseados em observações da atmosfera e oceanos também importantes e que podem ajudar a avaliar as alterações climáticas, tais como o nível do mar, temperatura do ar marinho e a temperatura de superfície do mar (SLINGO, 2013), reforçando mais uma vez o papel dessa variável no sistema climático.

A TSM tem sido relacionada também ao processo conhecido como sequestro de carbono. O sequestro de carbono é o nome dado ao fenômeno que ocorre principalmente nas florestas e oceanos, no qual o gás carbônico é capturado e, após um processo de fotossíntese, é lançado oxigênio na atmosfera. Apesar de não existirem muitos estudos deste processo no Oceano Atlântico Sul, é possível encontrar pesquisas de sequestro de carbono no Atlântico Norte, como em Gruber *et al.* (2002). Neste estudo, Gruber *et al.* (2002) avalia a variabilidade temporal do sequestro de carbono no Oceano Atlântico Norte e, após analisar uma série temporal de 18 anos, conclui que a variabilidade de carbono na região é fortemente impulsionada pelas anomalias de TSM e também pelas variações nas profundidades da chamada camada de mistura (do inglês, *mixed-layer depths*).

Portanto, é possível perceber que a TSM exerce um papel determinante na dinâmica oceano-atmosférica e está diretamente relacionada à ocorrência de diferentes fenômenos naturais. Dessa forma, conseguir entender e prever essa variável pode trazer grandes contribuições no que diz respeito à melhoria do monitoramento de importantes eventos climáticos e, assim, fazer com que se torne possível mitigar seus efeitos. Tendo isto em vista, foi desenvolvido um projeto de monitoramento do oceano Atlântico Tropical chamado PIRATA, que apresenta como um dos principais objetivos descrever e entender a evolução da temperatura de superfície do mar (GRODSKY *et al.*, 2005).

III.5 O projeto PIRATA

Eventos de grande impacto socioeconômicos acontecem na região abrangida pelo oceano Atlântico Tropical Sul. Entretanto, segundo Good *et al.* (2009), o processo que envolve estes eventos é bem menos compreendido do que aqueles que ocorrem em outras regiões próximas aos demais oceanos do mundo. A carência de estudos nesta região está relacionada à limitação de observações no oceano Atlântico Tropical Sul e, também, à importância que é dada ao oceano Atlântico Tropical Norte na literatura devido aos eventos ocorridos nesta região serem relativamente mais comuns e mais devastadores. Levando em consideração essa necessidade de melhor compreender os fenômenos ocorridos no oceano Atlântico Tropical, e a fim de tentar suprir essa lacuna de informação, foi proposto o projeto *Pilot Research Moored Array in the Tropical Atlantic* (PIRATA).

O PIRATA é uma rede de observação composta por boias espalhadas pelo Oceano Atlântico Tropical, planejada para monitorar uma série de variáveis dos processos de interação oceano-atmosfera e, dessa forma, proporcionar um melhor entendimento sobre a variabilidade desse sistema oceano-atmosférico (BOURLÈS *et al.*, 2008; SERVAIN *et al.*, 1998). Este projeto é semelhante a outros programas que já existiam em outras regiões, como o *Tropical Atmosphere-Ocean* (TAO), usado para estudar a variabilidade da *El Niño Oscilação-Sul* (ENOS) no Oceano Pacífico Equatorial.

O PIRATA teve seu início em 1997, a partir de uma parceria entre o Brasil, Estados Unidos e França. Nessa época o programa contava com doze boias e compunham a fase piloto do programa, que iria durar até 2001, quando estas boias iniciais foram completamente instaladas, passando então para fase de consolidação (BOURLÈS *et al.*, 2008; SERVAIN *et al.*, 1998). A fase de consolidação do programa procurou demonstrar que os dados do PIRATA poderiam contribuir para as pesquisas científicas, além de possuir aplicações operacionais ao fornecer os dados para previsões meteorológicas, por exemplo. Em 2006, o programa teve formalmente seu reconhecimento, passando a fazer parte do Sistema Global de Observações Oceânicas e do Sistema Global de Observações Climáticas (BOURLÈS *et al.*, 2008). A partir de então o programa passou para a fase de sustentação, quando ganhou novas extensões, adquirindo a configuração que mantém até hoje.

A definição da localização das boias, que pode ser observada pelos círculos amarelos na Figura III.1, foi estrategicamente pensada para que os dois principais modos de variabilidade climática do Oceano Atlântico Tropical pudessem ser observados: o modo equatorial e o modo meridional, também chamado de modo dipolo (SERVAIN; CLAUZET; WAINER, 2003).

As boias utilizadas pelo programa são conhecidas como *Autonomous Temperature Line Acquisition System* (ATLAS) e foram projetadas para medir variáveis meteorológicas através de sensores hidrológicos entre a superfície e quinhentos metros de profundidade. Cada boia é composta por dois sensores de pressão, onze sensores de temperatura e quatro sensores de

condutividade. Essas boias são capazes de armazenar as observações médias diárias, que são transmitidas para o sistema chamado Argos (BOURLÈS *et al.*, 2008; SERVAIN *et al.*, 1998).

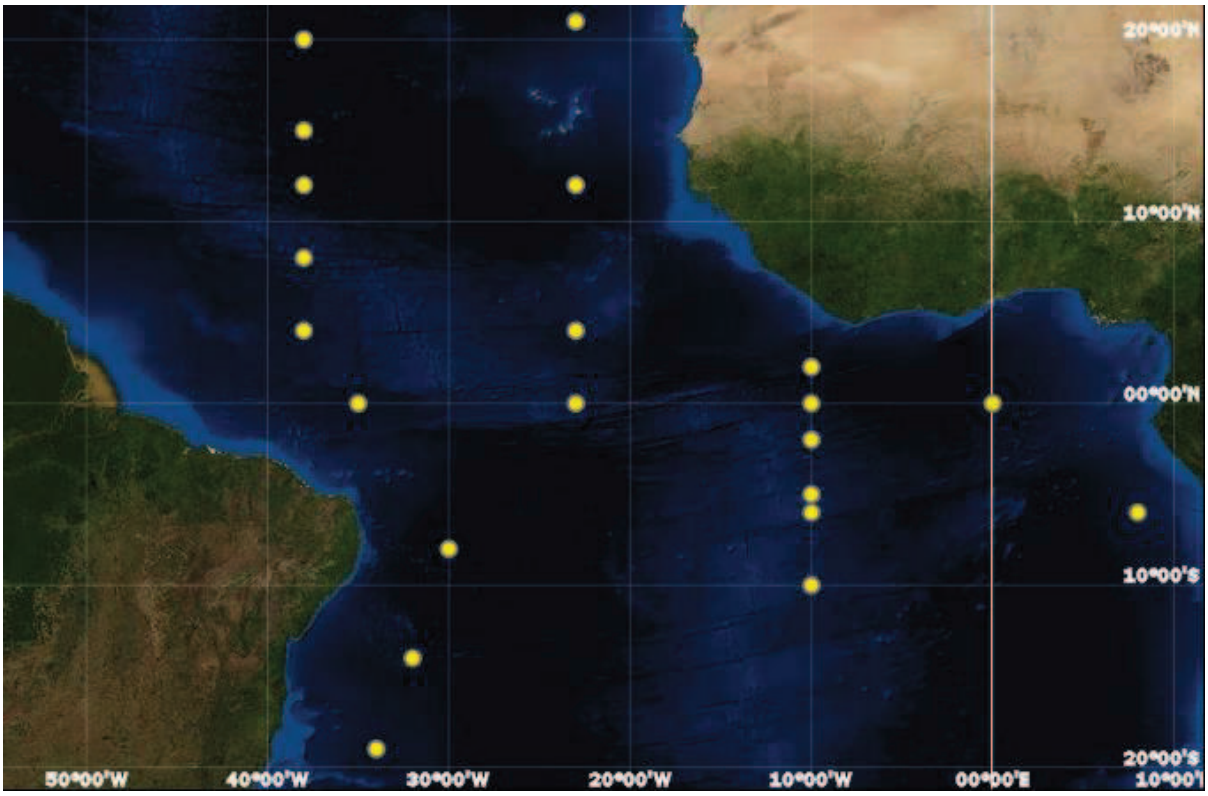


Figura III.1 A região de localização das boias do PIRATA

Fonte: GOOS-Brasil (2013)

As variáveis medidas e disponibilizadas para consulta são as temperaturas da água em quatorze profundidades diferentes situadas entre 1 e 500 metros da superfície - onde a temperatura medida entre 1 metro e 1,5 metros de profundidade é a referida temperatura de superfície do mar, salinidade em sete distintas profundidades (entre 1 e 12 metros da superfície), vento, radiação solar, precipitação, umidade relativa e temperatura do ar. Os diversos dados coletados são atualizados diariamente e estão disponíveis na Internet nas páginas do GOOS-Brasil (GOOS-BRASIL, 2013) e também na página do *National Oceanic & Atmospheric Administration* - NOAA (NOAA, 2013b), onde também é possível encontrar dados de outros programas.

Segundo Durand *et al.* (2005), a disponibilidade em tempo real dos dados é um aspecto muito positivo dos dados do PIRATA, o que faz com que tenham grande utilidade para aplicações em previsões. O PIRATA, assim como outros sistemas de observação oceânica, é essencial para descrever e entender o oceano e, por isso, seus dados têm sido usados para comparação de resultados numéricos e também para aumentar o conhecimento em regiões que são tidas como zonas importantes para a circulação nos oceanos (LINS *et al.*, 2013).

Dessa forma, os dados das diferentes variáveis obtidas e disponibilizados graças ao projeto PIRATA têm sido de grande utilidade para diversas pesquisas. O objetivo desse trabalho é utilizar dados históricos de TSM disponibilizados pelo projeto PIRATA para realizar novas previsões de TSM com diferentes horizontes de planejamento e utilizando diferentes frequências obtidas a partir da transformação dos dados orginalmente diários do PIRATA em dados semanais e mensais. Os próximos capítulos abordam os métodos de previsão e descrevem os experimentos computacionais realizados sobre o modelo de previsão.

Capítulo IV - Metodologia de previsão

A previsão exerce um papel de grande importância no gerenciamento de risco, na preparação da comunidade e também para a mitigação dos efeitos dos eventos extremos. Uma previsão acurada pode ajudar aos governos e órgãos responsáveis a tomarem decisões em tempo hábil, atenuando assim os impactos desses eventos sobre a economia e, principalmente, sobre a população.

No entanto, a falta de capacidade de prever com precisão e antecedência suficientes as condições relacionadas a estes eventos ainda é um empecilho para a mitigação de seus efeitos (MISHRA; SINGH, 2011). Neste contexto, ao considerar o importante papel que a TSM exerce nos diferentes eventos extremos, a previsão desta variável climática pode contribuir para melhoria das previsões desses fenômenos naturais.

Uma vez que uma série temporal é definida como um conjunto de observações ordenadas no tempo (MORETTIN; TOLOI, 1981; PARISI; SLIVA; SUBRAHMANIAN, 2013), o conjunto de dados de TSM fornecido pelo projeto PIRATA pode ser entendido como uma série temporal e, portanto, a escolha de modelos de previsão de séries temporais para realizar as previsões da temperatura de superfície do mar se mostra pertinente. Nas previsões de séries temporais, as observações passadas de uma mesma variável são coletadas e analisadas com intuito de desenvolver um modelo que consiga descrever o relacionamento existente entre elas e, dessa forma, tornar possível que seus valores sejam extrapolados para o futuro (ZHANG, 2003).

Um dos principais objetivos da modelagem de séries temporais é estimar uma função baseada nas informações presentes nos dados disponíveis, a fim de descrever o problema estudado com maior acurácia possível (LINS *et al.*, 2013). Considerando a relevância da previsão de séries temporais e suas diversas aplicações, tem sido despendido um grande esforço para o desenvolvimento e melhoria dos modelos de série temporais.

Dentre os modelos existentes, o *Autoregressive Integrated Moving Average* (ARIMA) é largamente utilizado para realizar previsões de séries temporais, sendo apontado como um dos modelos mais populares. Além do ARIMA, outras pesquisas sugerem o uso das Redes Neurais Artificiais (RNAs) como um modelo promissor para realizar este tipo de previsão (ZHANG, 2003). Desde então, os modelos ARIMA e RNAs têm sido frequentemente comparados, sem que haja uma conclusão definitiva com relação à superioridade de algum dos modelos no desempenho de previsão (ZHANG; PATUWO; HU, 1998).

Ao contrário da maioria dos métodos de previsão, os modelos ARIMA e RNAs são métodos que se diferenciam por procurarem extrair dos dados defasados da própria série estudada as informações necessárias para realizar suas previsões, buscando entender o

comportamento dos dados e aprender com eles, a fim de conseguir obter resultados de previsões mais próximos à realidade.

Este capítulo procura definir, primeiramente, os principais conceitos a respeito das séries temporais. A seguir, busca-se descrever as técnicas utilizadas para pré-processamento dos dados, inclusive a transformação de dados baseada em mudança de periodicidade das séries temporais. Depois são abordadas as principais características que tornam as duas técnicas utilizadas neste estudo, RNAs e ARIMA, ferramentas muito eficientes de previsão, bem como as medidas de desempenho mais frequentemente utilizadas para avaliar a acurácia de uma previsão são definidas. Por fim, alguns trabalhos relacionados à esta dissertação são brevemente apresentados.

IV.1 Séries temporais

Uma série temporal pode ser definida como um conjunto de observações, em que cada observação é registrada em um período *t* específico (BROCKWELL; DAVIS, 2010). As observações de uma série temporal devem ser coletadas em intervalos regulares, como, por exemplo, diariamente, mensalmente ou anualmente. Além disso, uma série temporal pode ser discreta ou contínua, mas para fins deste estudo, estas são consideradas sempre da forma discreta, onde o conjunto de observações T é representado pela equação (1).

$$T = \{x_1, x_2, \ldots, x_t\} \quad (1)$$

Outra característica importante a respeito de uma série temporal é identificar se esta apresenta comportamento estacionário ou não estacionário. Informalmente, uma série é dita estacionária quando apresenta um comportamento estável, ou seja, com médias e flutuações ao redor da média mais ou menos constantes (MORETTIN; TOLOI, 1981). De modo mais formal, uma série temporal é estacionária quando a sua média, variância e covariância permanecem as mesmas ao longo do tempo (GUJARATI; PORTER, 2008). Já as séries não estacionárias não apresentam estas características e representam a maior parte dos problemas encontrados na prática (MORETTIN; TOLOI, 1981). A estacionariedade ou os desvios de estacionariedade podem ser observados com auxílio de um gráfico da série temporal (BROCKWELL; DAVIS, 2010), conforme pode ser observado na Figura IV.1 onde é apresentado o gráfico da função cosseno, que é um exemplo muito conhecido de série estacionária.

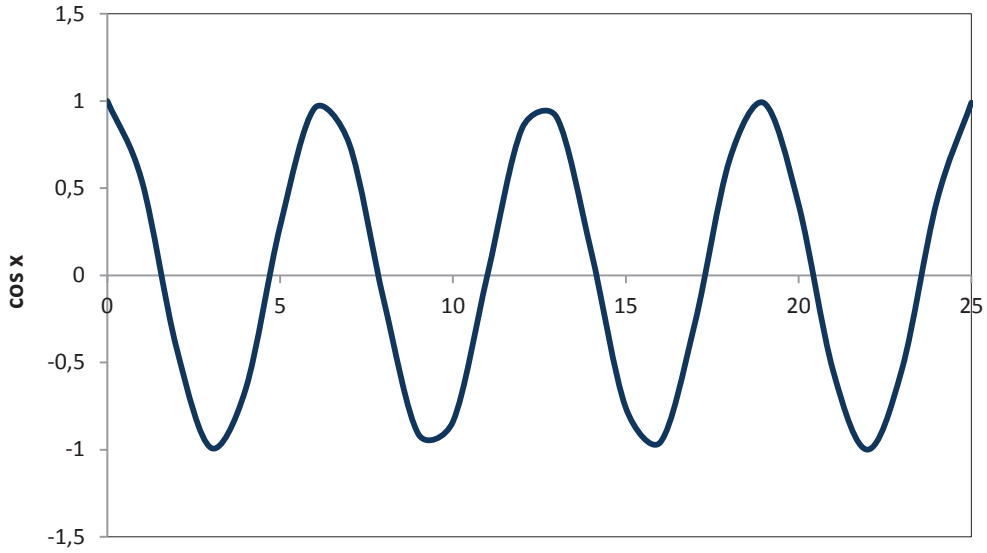


Figura IV.1 Gráfico da função cosseno, exemplo de série estacionária.

Outra maneira utilizada para identificar uma série estacionária é a função de autocorrelação (BROCKWELL; DAVIS, 2010). A autocorrelação pode ser definida como a correlação entre integrantes de séries de observações ordenadas no tempo ou no espaço (GUJARATI; PORTER, 2008). Sendo assim, quando há presença de autocorrelação em uma série, é possível observar uma correlação entre os valores da série com os valores desta mesma série defasada em algumas unidades de tempo (GUJARATI; PORTER, 2008).

Além de serem utilizadas para realizar previsões meteorológicas, as séries temporais apresentam diversas outras aplicações, sendo utilizadas, por exemplo, para realizar previsões de demanda, no censo demográfico e nas taxas de desemprego, destacando-se suas aplicações na área de econometria (GUJARATI; PORTER, 2008).

IV.2 Técnicas de pré-processamento dos dados

As séries temporais estão suscetíveis a diversos tipos de imperfeições, isto é, podem apresentar interrupções na série devido à falta parcial de dados ou dados que não representem a realidade da série, como ruídos. No entanto, para que as previsões possam obter resultados satisfatórios, é preciso que estas imperfeições sejam tratadas durante a preparação ou pré-processamento dos dados.

Para que o pré-processamento seja bem-sucedido, é preciso que haja uma noção geral dos dados. Para tanto, o conhecimento sobre as características básicas estatísticas do conjunto de dados, tais como média, desvio padrão e variância, se mostra muito relevante, uma vez que podem auxiliar a identificar as propriedades do conjunto de dados analisado e ressaltar possíveis necessidades de tratamento nos dados (HAN; KAMBER; PEI, 2011). Além

disso, o uso de gráficos também pode ajudar a perceber os padrões presentes nos dados com mais clareza e por isso é uma ferramenta bastante utilizada (HAN; KAMBER; PEI, 2011).

IV.2.1 Limpeza dos dados

Os dados do mundo real tendem a serem incompletos, inconsistentes e apresentarem ruídos. Pensando nisso foram desenvolvidas técnicas de limpeza para preencher os dados faltantes e suavizar possíveis ruídos e corrigir inconsistências (HAN; KAMBER; PEI, 2011).

No caso do programa PIRATA, a limpeza dos dados se faz um procedimento diferencial devido ao contexto em que os dados de TSM são receptados. Dado que as boias estão localizadas no oceano, a mesmas estão sujeitas a diversos tipos de interferências externas tais como pirataria ou ataques de animais. Isto pode ocasionar a interrupção da coleta dos dados por dias ou longos períodos de tempo e, por isso, na maioria dos casos, a disponibilidade efetiva dos dados não chega a ser superior a 90% (BOURLÈS *et al.*, 2008).

Para o caso de indisponibilidade de dados, a depender da especificidade de cada caso, os seguintes métodos são indicados: (i) ignorar o dado faltante; (ii) preencher manualmente os dados faltantes; (iii) utilizar uma constante global para substituir o dado faltante; (iv) utilizar uma medida de tendência central para preencher o valor faltante; (v) utilizar a mesma média ou mediana para todas as amostras pertencentes a uma mesma classe; e (vi) utilizar o valor mais provável para preencher o valor faltante (HAN; KAMBER; PEI, 2011).

IV.2.2 Transformação da série em estacionária

A maioria dos modelos de previsão necessitam trabalhar com séries estacionárias. Como diversos fenômenos são regidos em ambientes não estacionários, em muitos casos é necessário que a série temporal não estacionária seja transformada em uma série estacionária, a fim de que seja possível realizar previsões mais acuradas (ZHANG, 2003). Duas técnicas comumente utilizadas para este fim são a diferenciação e a remoção de tendência.

IV.2.2.1 Diferenciação

Uma técnica utilizada para transformar séries não estacionárias em séries estacionárias é chamada de diferenciação (LINS *et al.*, 2013; MORETTIN; TOLOI, 1981). A diferenciação consiste em tomar sucessivas diferenças da série de dados estudada até que seja obtida uma série estacionária. A diferenciação de uma série é dada pela equação (2).

$$\mathbf{w_t} = x_t - x_{t-d} \qquad (2)$$

Onde $\mathrm{w_t}$ é o operador diferença, x_t corresponde à observação da série temporal no tempo *t* e *d* representa o número de defasagens. Em muitos casos, quando a série temporal original não é estacionária a sua primeira diferença já é estacionária (GUJARATI; PORTER, 2008), por isso comumente adota-se $d = 1$.

IV.2.2.2 Remoção de tendência

Outra técnica utilizada para transformar séries não estacionárias em séries estacionárias é conhecida como remoção de tendência. Nesta técnica, busca-se obter o resíduo da série, que é a série sem tendência (BOX; JENKINS; REINSEL, 2008). A série sem tendência ou a série de resíduos é obtida a partir da diferença entre a regressão linear simples da série de dados e a própria série de dados (GUJARATI; PORTER, 2008).

IV.2.3 Mudança de periodicidade

Define-se por periodicidade a frequência na qual uma série temporal é coletada. É muito comum existirem séries diárias, semanais, mensais, trimestrais e anuais. A mudança de periodicidade consiste em transformar uma série de alta frequência em uma série de baixa frequência, como, por exemplo, transformar uma série diária em série mensal. Esta transformação é feita a partir de uma operação de agregação, também conhecida como operação de resumo (HAN; KAMBER; PEI, 2011). No caso das previsões, a agregação pode ser útil ao possibilitar que sejam utilizadas diferentes janelas de tempo e, por conseguinte, horizontes diferentes de previsão.

Além disso, são utilizados na mudança de periodicidade os seguintes métodos: (i) média, em que calcula-se a média aritmética dos valores da série no período selecionado; (ii) acumulado, onde os valores da série no período selecionado são somados; e (iii) fim de período, que corresponde à última observação da série no período selecionado (IPEADATA, 2013).

IV.3 Modelos Auto-regressivos Integrados de Médias Móveis (ARIMA)

O modelo ARIMA faz parte dos denominados modelos estatísticos. Eles são amplamente utilizados nas previsões de séries temporais devido à relativa simplicidade em que são implementados e que também podem ser entendidos. Um modelo estatístico é definido como descrição matemática de um sistema ou processo em termos de variáveis aleatórias, onde um sistema complexo é representado de maneira simplificada (SLINGO, 2013).

O modelo ARIMA é uma combinação de diferentes modelos, sendo constituídos pelos modelos auto-regressivos (AR), modelos de média móvel (MA) e a diferenciação integrada (I). Desta forma, para mais bem compreender o modelo ARIMA é importante entender suas componentes, que são apresentadas nas próximas subseções. Além disto, a processo para previsão por meio destes modelos é também um componente muito importante e, por isso, a subseção IV.3.5 apresenta a metodologia Box-Jenkins.

IV.3.1 Modelos AR

O modelo auto-regressivo é representado por $AR(p)$, em que *p* refere-se ao número de termos auto-regressivos (GUJARATI; PORTER, 2008). Este modelo é caracterizado por incluir em sua análise um ou mais valores defasados, isto é, de períodos passados da própria variável, ajustando-os conforme mostra a equação (3).

$$AR(p) = \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + \cdots + \emptyset_p x_{t-p} + e_t \quad (3)$$

Onde x_t corresponde à observação da série temporal no tempo t, e_t representa o erro de eventos aleatórios que não podem ser explicados pelo modelo no período t e $\emptyset_{\mathrm{p}}$ corresponde ao parâmetro do modelo AR de ordem p.

IV.3.2 Modelos MA

O modelo de média móvel é representado por $MA\,(q)$, em que q denota o número de termos de médias móveis (GUJARATI, 2006). Neste modelo, a previsão depende dos valores dos erros observados em cada período passado, sendo definido conforme a equação (4).

$$MA\,(q) = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \quad (4)$$

Onde e_t representa o erro de eventos aleatórios que não podem ser explicados pelo modelo no período t e θ_q corresponde ao parâmetro do modelo MA de ordem q.

IV.3.3 Modelos ARMA

O modelo auto-regressivo de média móvel é representado por $ARMA\,(p, q)$, em que p refere-se ao número de termos auto-regressivos e q denota o número de termos de médias

móveis. Este modelo pode ser entendido como uma combinação entre os modelos $AR(p)$ e $MA\ (q)$, sendo definido conforme mostra a equação (5).

$$ARMA\ (p,q) = \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + \cdots + \emptyset_p x_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \quad (5)$$

Onde x_t corresponde à observação da série temporal no tempo t, e_t representa o erro de eventos aleatórios que não podem ser explicados pelo modelo no período t e $\emptyset_{\mathrm{p}}$ corresponde ao parâmetro do modelo AR de ordem p e θ_q corresponde ao parâmetro do modelo MA de ordem q.

IV.3.4 Modelos ARIMA

Os três modelos descritos anteriormente, AR, MA e ARMA, adequam-se melhor a conjuntos de observações que sejam geradas, necessariamente, por uma série temporal estacionária (BROCKWELL; DAVIS, 2010). Entretanto, nos casos em que é necessário realizar uma transformação dos dados para gerar uma nova série que tenha as características e estacionariedade, os modelos auto-regressivos integrados de médias móveis (ARIMA) se mostram mais adequados.

Os modelos ARIMA combinam modelos auto-regressivos com modelos de média móvel, permitindo uma modelagem não estacionária através da diferenciação dos dados da série temporal (BARUA; NG; PERERA, 2012). O modelo é representado por $ARIMA\ (p,d,q)$, onde (i) p denota o número de termos auto-regressivos; (ii) d refere-se ao número de vezes que a série temporal precisa sofrer diferenciação antes de tornar-se estacionária; e (iii) q denota o número de termos de médias móveis (GUJARATI; PORTER, 2008). Dessa forma, quando $ARIMA\ (p,0,0)$ o modelo é caracterizado como puramente auto-regressivo e, de maneira similar, quando $ARIMA\ (0,0,q)$, puramente um modelo de média móvel. O ARIMA é uma generalização do modelo ARMA, que se diferencia pela abrangência de séries temporais que precisem sofrer diferenciação, isto é, d diferenciações para tornar-se estacionária. Logo, no caso em que $d = 0$, os dois modelos se igualam.

Ao contrário dos modelos de regressão onde a variável prevista é explicada pelos seus *k* regressores, o ARIMA permite que esta seja explicada por valores passados, ou defasados da própria variável e seus termos de erro estocástico (GUJARATI; PORTER, 2008). A ênfase desse método não é construir modelos de equação única ou equações simultâneas, mas analisar as propriedades probabilísticas ou estocásticas da série temporal, deixando que os dados falem por eles mesmos (GUJARATI; PORTER, 2008).

Dessa forma, os modelos ARIMA são capazes de descrever o comportamento de séries onde os erros observados são autocorrelacionados e influenciam a evolução do processo, podendo descrever tanto séries estacionárias quanto séries não estacionárias de maneira satisfatória (MORETTIN e TOLOI, 1981). A definição do modelo $ARIMA\ (p, d, q)$ pode ser observada na equação (6)

$$ARIMA\ (p, d, q) = \emptyset_0 + \emptyset_1 w_{t-1} + \emptyset_2 w_{t-2} + \cdots + \emptyset_p w_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \quad (6)$$

Tal que

$$w_t = x_t - x_{t-d} \quad (7)$$

Onde w_t é o operador diferença, x_t corresponde à observação da série temporal no tempo t, *d* representa o número de defasagens e_t representa o erro de eventos aleatórios que não podem ser explicados pelo modelo no período t e $\emptyset_p$ corresponde ao parâmetro do modelo AR de ordem p e θ_q corresponde ao parâmetro do modelo MA de ordem q.

IV.3.5 Metodologia *Box-Jenkins*

A chamada metodologia *Box-Jenkins* foi desenvolvida pelos pesquisadores que dão nome à metodologia, Box e Jenkins, com o intuito estabelecer uma abordagem prática para construir os modelos ARIMA (ZHANG, 2003). A metodologia *Box-Jenkins* é utilizada para auxiliar na identificação do processo que segue uma série temporal, isto é, se uma série segue o processo puramente auto-regressivo ou puramente de média móvel ou ainda se a série temporal segue um processo ARMA ou ARIMA (GUJARATI; PORTER, 2008).

De uma maneira geral, o desenvolvimento do modelo ARIMA utilizando a metodologia Box-Jenkins segue três passos iterativos: (i) identificação do modelo; (ii) estimação do parâmetro; (iii) checagem do diagnóstico (BARUA; NG; PERERA, 2012; ZHANG, 2003) e há ainda uma etapa seguinte, que é a própria previsão com base no modelo definido (GUJARATI; PORTER, 2008).

IV.3.5.1 Identificação do modelo

A primeira etapa da metodologia *Box-Jenkins* consiste em encontrar valores adequados para cada um dos parâmetros *p*, *d* e *q* (GUJARATI; PORTER, 2008). A ideia básica da

identificação do modelo é que se uma série temporal é gerada por um processo ARIMA, ela deve ter algumas propriedades de autocorrelação. Dessa forma, a metodologia Box-Jenkins propõe que seja utilizada uma função de autocorrelação e uma função de autocorrelação parcial nos dados da amostra como ferramentas básicas para identificar a ordem do modelo ARIMA (ZHANG, 2003). Neste primeiro passo de identificação, muitas vezes é necessário que a série temporal seja transformada em uma série não estacionária, de forma que o modelo ARIMA consiga realizar melhor a previsão.

IV.3.5.2 Estimação do parâmetro

Uma vez escolhido o modelo ARIMA, é necessário então estimar os parâmetros dos termos auto-regressivos e de médias móveis incluídos no modelo, de maneira a minimizar o erro global do modelo. Para isso, podem ser utilizados diferentes métodos, entre eles estão o método dos mínimos quadrados e métodos de estimação não-lineares (GUJARATI; PORTER, 2008). Entretanto, a estimação do parâmetro torna-se uma tarefa mais fácil com o apoio de pacotes estatísticos e por isso não é preciso muito aprofundamento nos conceitos relacionados a esta estimativa (GUJARATI; PORTER, 2008).

IV.3.5.3 Checagem do diagnóstico

A checagem do diagnóstico consiste em verificar a adequação do modelo utilizado. Para tanto, deve-se checar se as premissas sobre os erros foram satisfeitas.

Quando a conclusão desta etapa mostrar que o modelo não está adequado à série estudada, o processo deve se repetir, e mais uma vez as etapas anteriores devem ser realizadas até que seja encontrado o modelo mais adequado para realizar a previsão (ZHANG, 2003).

IV.3.5.4 Previsão

Uma das razões para a modelagem do ARIMA ser tão popular é o sucesso obtido por suas previsões (GUJARATI; PORTER, 2008). Dessa forma, ao definir um modelo ARIMA utilizando a metodologia proposta por Box e Jenkins, há uma expectativa maior de se conseguir resultados de previsões acuradas.

IV.4 Modelos baseados em Redes Neurais Artificiais

As redes neurais artificiais foram inspiradas por sistemas biológicos, mais especificamente pelo funcionamento do cérebro humano, o que justifica o nome dado a essa técnica. Uma rede neural é composta por vários nós interconectados, que também podem ser chamados de neurônios ou unidades de processamento e é pelas conexões entre os nós, também chamadas de arcos, que as informações entre eles são passadas (ZHANG; PATUWO; HU, 1998).

As redes neurais artificiais são capazes de aprender e generalizar a partir de experiências vivenciadas e, por isso, têm se mostrado uma ferramenta com grande habilidade para classificar e reconhecer padrões (ZHANG; PATUWO; HU, 1998). A rede neural se tornou uma técnica muito popular para realizar previsões em diferentes áreas de interesse, encontrando aplicações em finanças, geração de energia, medicina, recursos naturais e ciências do ambiente (MAIER; DANDY, 2000).

Tendo em vista que os modelos de previsão utilizam como entrada valores passados e que, consequentemente, sobre os quais já se tem conhecimento, e, além disso, que a partir do uso desses modelos espera-se conseguir saídas que correspondam aos valores futuros desejados, as redes neurais artificiais encontram nas previsões uma de suas maiores aplicações.

Zhang *et al.* (1998) enumera algumas características das RNAs que justificam a sua escolha como ferramenta de previsões em detrimento a outras: (i) RNAs são métodos autoadaptativos orientados a dados, que aprendem a partir de exemplos e conseguem perceber relacionamentos sutis entre os dados, mesmo quando estes são desconhecidos ou difíceis de serem descritos; (ii) RNAs conseguem generalizar, ou seja, depois de aprender com a amostra utilizada, conseguem inferir parte desconhecida da população, ainda que haja ruídos nos dados utilizados como amostra; (iii) RNAs são aproximadores universais de função, isto é, demonstrou-se que uma rede neural é capaz de aproximar funções contínuas com acurácia satisfatória; (iv) RNAs são não-lineares e conseguem realizar modelagens não-lineares sem que haja um conhecimento *a priori* dos relacionamentos entre as variáveis de entrada e de saída.

A maioria dos estudos de RNA busca fazer previsões e reconhecer padrões e, além disso, comparar o uso de redes neurais com outras abordagens tradicionais em problemas relacionados a ciências ecológicas e atmosféricas (CHATTOPADHYAY, 2007), como é o caso dos estudos realizados por Wu *et al.* (2006) e Aguilar-Martinez e Hsieh (2009), onde o primeiro utiliza as redes neurais para fazer previsões de temperatura de superfície do mar do oceano Pacífico Tropical, enquanto o segundo, além de utilizar as redes neurais, faz uso também de outro método conhecido como *Support Vector Machine* (SVM) e compara seus resultados.

IV.4.1 Perceptron de múltiplas camadas

Diferentes tipos de redes neurais artificiais foram propostos ao longo dos estudos sobre essa técnica, mas, para Morid *et al.* (2007), um dos modelos utilizados em pesquisas da área de hidrologia é o chamado *perceptron* de múltiplas camadas (MLP). As redes neurais MLP são utilizadas para diferentes finalidades, mas destacam-se nos problemas que envolvem análises de processos complexos, como é o caso das previsões de secas, onde o relacionamento entre as variáveis de entrada e saída precisam descobertos (MORID; SMAKHTIN; BAGHERZADEH, 2007).

Uma rede neural artificial MLP é constituída por uma camada de entrada, uma ou mais chamadas escondidas e uma camada de saída, onde cada camada é composta por unidades de processamento ou nós (HAN; KAMBER; PEI, 2011; ZHANG; PATUWO; HU, 1998). A primeira camada, chamada de entrada, é por onde as informações externas são recebidas, ou seja, por onde os dados são introduzidos na rede, enquanto a última camada, de saída, é por onde os resultados para uma dada entrada são produzidos (MORID; SMAKHTIN; BAGHERZADEH, 2007). E entre as camadas de entrada e saída, há uma ou mais camadas escondidas, que é onde os dados são processados. Dessa forma, os nós pertencentes à camada de entrada se conectam aos nós da camada escondida por meio de ligações ou arcos, por onde passam as informações entre eles. Os nós da camada escondida, por sua vez, conectam-se aos nós da camada de saída. A Figura IV.2 mostra um exemplo de uma rede neural MLP constituída por uma camada escondida.

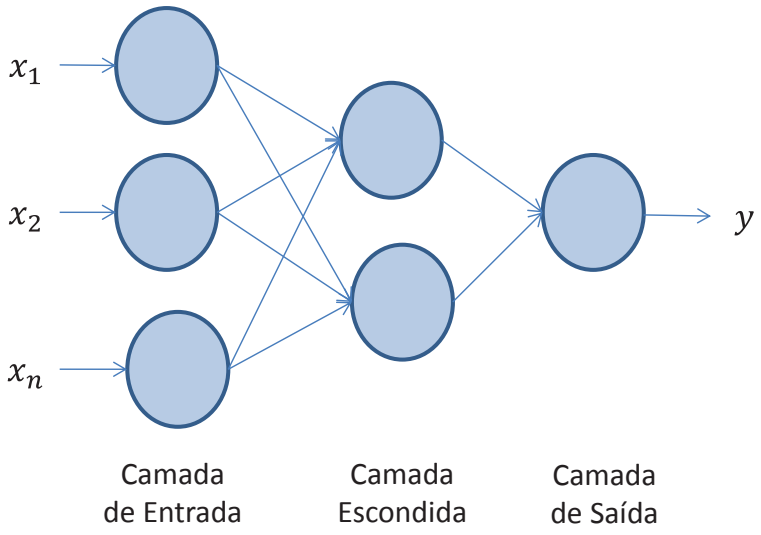


Figura IV.2 Exemplo de uma rede neural artificial do tipo MLP e suas camadas

IV.4.2 Modelagem de uma Rede Neural Artificial

Há diversos fatores que são de grande importância para modelar uma rede neural artificial, de forma a conseguir bons resultados com esta técnica. Entretanto, modelar uma RNA não se trata de uma tarefa trivial, uma vez que envolve fatores que afetam o seu desempenho e que, por isso, devem ser cuidadosamente considerados (ZHANG; PATUWO; HU, 1998). Entre os fatores que mais influenciam o desempenho de uma rede neural estão: a arquitetura da rede, a função de ativação utilizada, o algoritmo de treinamento a ser escolhido, o treinamento e teste do modelo utilizado. E, por fim, faz-se uso de medidas que auxiliem a medir o desempenho obtido pela rede modelada para o problema em questão.

IV.4.2.1 Arquitetura da rede

A definição da arquitetura da rede consiste em determinar o número de nós em cada uma das camadas de uma RNA. Apesar de existirem métodos que procuram otimizar a arquitetura de uma RNA, eles costumam ser complexos e difíceis de implementar e, além disso, não garantem uma solução ótima (ZHANG; PATUWO; HU, 1998). Logo, se tratando de arquitetura, na maioria dos casos, a tentativa e erro muitas vezes acaba sendo a melhor solução (HAN; KAMBER; PEI, 2011).

Os nós que compõem a camada de entrada correspondem ao número de variáveis no vetor utilizado como entrada para realizar as previsões. Para Zhang *et al.* (1998), essa decisão é a decisão mais crítica para um problema de previsão de séries temporais. Essa importância se deve ao fato de conter informações sobre a complexa estrutura de correlação entre os dados. Nesses problemas de previsões de séries temporais, a quantidade de nós costuma corresponder ao número de observações passadas que serão utilizadas para descobrir os padrões e fazer previsões para os valores futuros (CHATTOPADHYAY, 2007; ZHANG; PATUWO; HU, 1998).

O número de camadas escondidas e quantidade de nós em cada uma delas exerce um papel muito importante na obtenção de bons resultados com o uso de RNA, uma vez que são os nós que fazem parte da camada escondida os responsáveis por perceber as características e padrões nos dados e, além disso, fazer um mapeamento não linear complexo entre as variáveis de entrada e saída (ZHANG; PATUWO; HU, 1998). A quantidade de camadas escondidas pode variar de acordo com a quantidade de nós que sejam necessários, no entanto, uma única camada costuma ser suficiente para a maioria dos casos estudados. O número de nós também é uma decisão complicada a se tomar, pois, de uma maneira geral, redes com poucos nós na camada escondida costumam ter melhores habilidades de generalização, além de reduzirem os problemas de *overfitting* - isto é, perda da capacidade de

generalização, mas por outro lado, redes com poucos nós podem reduzir sua capacidade de modelagem e aprendizado dos dados. Zhang *et al.* (1998) observaram em seus estudos que redes que apresentavam o mesmo número de nós na camada escondida e na camada de entrada mostraram ter melhores resultados.

Já no caso da camada de saída, a quantidade de nós está relacionada a cada caso especificamente estudado e para Maier e Dandy (2000) essa quantidade deve ser igual ao número de saídas que se deseja para o modelo. Dessa forma, para um problema de previsões de séries temporais, por exemplo, esse número corresponde ao horizonte de previsão. Além disso, segundo Zhang *et al.* (1998), há dois tipos de previsões e que influenciam na decisão da topologia da rede: *one-step-ahead* e *multi-step-ahead*. No primeiro caso utiliza-se apenas um nó, onde os valores previstos são iterativamente usados como entradas para outras previsões. No caso do *multi-step-ahead* costuma-se ter diversos nós, afim de que as previsões sejam feitas diretamente, ou seja, muitos passos adiante.

Além das decisões que envolvem os nós em cada camada de uma rede neural do tipo MLP, outro fator que faz parte da decisão da arquitetura da rede é a maneira como os nós das diferentes camadas se interconectam e, portanto, para Maier e Dandy (2000), a escolha da quantidade de conexões de peso é um aspecto de grande relevância.

IV.4.2.2 Função de ativação

O relacionamento entre os parâmetros utilizados como entrada e aqueles que são gerados como saídas de uma rede neural artificial precisam de uma função para acontecer, a esta função é dada o nome de função de ativação ou de transferência (CHATTOPADHYAY, 2007). Apesar de ser possível que os nós possuam diferentes funções de ativação, mesmo pertencendo a uma mesma camada, é comum que as redes utilizem a mesma função de ativação para todos os nós de uma mesma camada (ZHANG; PATUWO; HU, 1998), e, mais do que isso, normalmente a mesma função de transferência é usada para todas as camadas (MAIER; DANDY, 2000). Apesar de haver outras funções de ativação, as mais utilizadas são:

i) Função sigmoide ou logística:

$$f(x) = (1 + \exp(-x))^{-1} \quad (8)$$

ii) Função linear:

$$f(x) = x \quad (9)$$

iii) Função tangente hiperbólica:

$$f(x) = (\exp(x) - \exp(-x))/(\exp(x) + \exp(-x)\,) \qquad (10)$$

IV.4.2.3 Algoritmo de treinamento

O treinamento de uma rede neural pode ser entendido como um problema de minimização não linear, cujos pesos das conexões entre os nós são modificados de forma iterativa, a fim de que se consiga minimizar o erro quadrático médio ou erro médio total da rede entre os valores previstos obtidos como saída do modelo e os valores reais desejados (ZHANG; PATUWO; HU, 1998). Há diversos algoritmos de treinamento, que diferem entre si, principalmente, pela maneira como os pesos dos arcos são modificados, mas o mais popular é o chamado *Backpropagation* (BP), também conhecido por regra delta generalizada.

De maneira simplificada, o algoritmo de aprendizado *backpropagation* abrange a propagação de erros de trás para frente e funciona da seguinte maneira: os padrões são apresentados à camada de entrada e as atividades resultantes são propagadas para as demais camadas, até a camada de saída, então a resposta obtida é comparada com o valor desejado para que o erro seja calculado. Feito isso, o erro é retropropagado da camada de saída até camada de entrada, de forma que os pesos dos arcos sejam atualizados e, dessa forma, se consiga reduzir o erro.

IV.4.2.4 Particionamento dos dados

Na modelagem de uma RNA é necessário selecionar as amostras de treinamento e de teste, que podem influenciar no desempenho do modelo. A amostra de treinamento é usada durante a fase de desenvolvimento do modelo de previsão, enquanto a amostra de teste é utilizada para avaliar as previsões realizadas pelo modelo. Há ainda uma terceira amostra chamada de validação que é utilizada para evitar o problema de *overfitting*, mas é comum fazer uso da amostra de teste para acumular essa função, principalmente quando há poucos dados para serem utilizados (ZHANG; PATUWO; HU, 1998).

Dado o conjunto de dados para estudar o problema, é preciso dividi-lo entre as amostras de treinamento e teste. Apesar de não haver um valor ótimo de divisão, a maioria dos estudos destina a maior parte dos dados para a amostra de treinamento e uma quantidade menor para a amostra de teste. Os percentuais dos conjuntos de dados mais comumente

utilizados para treinamento e teste, respectivamente, são: 90% e 10%, 80% e 20%, 70% e 30%.

Maier e Dandy (2000) ressaltam que, tipicamente, as redes neurais artificiais não conseguem fazer previsões com valores que estejam fora do intervalo de dados pertencentes à amostra utilizada como treinamento. Outro fator que se deve observar é que os conjuntos de dados utilizados para as amostras precisam ser representativos da mesma população. Além disso, em geral, é esperado que quanto maior for o tamanho da amostra, também maior será a acurácia dos resultados obtidos (ZHANG; PATUWO; HU, 1998).

IV.5 Medidas de desempenho

É possível perceber que escolher a técnica certa para realizar previsões de séries temporais não é uma tarefa trivial, uma vez que é difícil na prática determinar o processo gerador de uma série temporal e, além disso, há um consenso no que diz respeito à inexistência de um único método que seja superior em todas as situações, principalmente pelo fato de que os problemas reais são frequentemente complexos e, por isso, um único modelo não seria capaz de capturar diferentes padrões igualmente bem (ZHANG, 2003). Por isso, uma prática adotada pelos pesquisadores é testar modelos diferentes e selecionar aquele que apresentar o resultado mais acurado.

Dessa forma, a etapa seguinte às previsões é avaliar a acurácia dos resultados obtidos, pois assim é possível comparar diferentes modelos e métodos de previsão e chegar à definição do melhor modelo, dentro todos os estudados. Neste contexto, a acurácia dos modelos é normalmente medida com auxílio de medidas estatísticas.

De uma forma geral, ao utilizar medidas de desempenho para um modelo, o que se espera é perceber a diferença entre o resultado esperado e o obtido pelo modelo de previsão utilizado. Há diferentes medidas de desempenho disponíveis, mas como cada uma delas apresenta suas vantagens e desvantagens, é comum utilizar mais de uma para medir a acurácia de uma previsão (ZHANG; PATUWO; HU, 1998).

Algumas das principais medidas desempenho utilizadas são apresentadas a seguir, onde o erro de previsão individual é representado por e_t, o valor real é representado por x_t e N representa o número de termos de erro.

i) Erro médio absoluto (EMA) - É definido como a média dos erros absolutos, podendo ser obtido através da equação (10).

$$\mathrm{EMA} = \sum|e_t| \,) \,/\, N \qquad (11)$$

ii) Soma dos erros quadráticos (SQE) - Trata-se do somatório dos quadradros dos erros e está representado na equação (11).

$$\text{SQE} = \sum(\text{e}_\text{t})^2 \qquad (12)$$

iii) Erro médio quadrático (EMQ) - É definido pela média dos erros quadráticos e pode ser obtido através da equação (12).

$$\text{EMQ} = \sum(\text{e}_\text{t})^2/\text{N} \qquad (13)$$

iv) Raiz do erro médio quadrático (REMQ) - Como o nome já diz, é obtido a partir da raiz quadrada do EMQ, conforme mostra a equação (13).

$$\text{REMQ} = \sqrt{\sum(\text{e}_\text{t})^2/\text{N}} \qquad (14)$$

v) Erro médio percentual absoluto (EMPA) - É definido pela média do erro absoluto entre os valores previstos e os valores reais expresso em percentual dos valores reais, podendo ser obtido pela equação (14).

$$\text{EMPA} = \frac{100}{N} \sum \left|\frac{e_t}{x_t}\right| \qquad (15)$$

Capítulo V - Experimentos computacionais de previsão para TSM

Um modelo de previsão é composto por três componentes básicos, que devem ser definidos durante a modelagem do problema: (i) variáveis de entrada; (ii) metodologia e (iii) variáveis de saída, isto é, o que se deseja obter como resultado da previsão. Neste trabalho, as variáveis de entrada corresponde à série temporal de temperatura de superfície do mar, que apresenta relevância em contextos práticos relacionados a previsão de eventos climáticos.

A metodologia de previsão definida no Capítulo IV permite que haja uma análise comparativa entre dois métodos diferentes, os modelos Auto-regressivos de Médias Móveis (ARIMA) e as Redes Neurais Artificiais (RNAs), que são apontados na literatura como adequados para a previsão de eventos climáticos (MISHRA; SINGH, 2010). Propõe-se aqui nesta etapa de metodologia, uma transformação prévia da série de TSM em diferentes frequências, isto é, a partir dos dados brutos diários a série pode ser adequada a frequências semanais, mensais, trimestrais ou anuais. O objetivo é verificar se esta transformação pode produzir bons resultados na previsão dos dados em diferentes horizontes de tempo. Ressalta-se que a avaliação desta técnica constitui uma das principais contribuições desta dissertação.

Dessa forma, este capítulo está organizado da seguinte maneira: primeiramente, a etapa de preparação do experimento é descrita; em seguida, os modelos ARIMA e RNA utilizados são definidos e os resultados de suas previsões são comparados. Por fim, um novo experimento utilizando apenas o modelo ARIMA é realizado, a fim de avaliar o impacto das transformações da série de TSM nos resultados da previsão, e os resultados obtidos são discutidos.

V.1 Preparação do experimento

É essencial que uma quantidade significativa de dados históricos seja utilizada para que se consiga identificar relacionamentos entre parâmetros climáticos distintos que possam ser utilizados na previsão de secas (TADESSE *et al.*, 2004). No entanto, em grandes volumes de dados é comum que haja muita informação de pouca relevância para o que se deseja estudar. Por isso, diversas técnicas têm sido desenvolvidas para que se consiga extrair as informações desejáveis das bases de dados, como é o caso da mineração de dados. A mineração de dados pode ser definida como um processo que procura extrair informações úteis e não triviais de grandes quantidades de dados (DHANYA; KUMAR, 2009; HAN; KAMBER; PEI, 2011) .

Com auxílio das técnicas de mineração de dados, alguns procedimentos foram realizados neste estudo a fim de tornar possível a extração de informações úteis dos dados de TSM e como consequência realizar previsões satisfatórias. Ainda que sejam utilizadas duas

técnicas diferentes de previsão, há diversas etapas na modelagem do problema que são comuns à ambas e os principais procedimentos para a preparação dos experimentos realizados nesta dissertação estão descritos nos itens a seguir.

V.1.1 Seleção e limpeza dos dados

As observações diárias de temperatura de superfície do mar do projeto PIRATA foram obtidas a partir do site do GOOS-Brasil (GOOS-BRASIL, 2013), conforme fora abordado no Capítulo III desta dissertação. Estes dados foram coletados pelas diversas boias localizadas no Oceano Atlântico Tropical. Cada boia tem um posicionamento geográfico no oceano (latitude, longitude) e fornece a série temporal de TSM daquele ponto. Além disso, visto que as boias se encontram no oceano, elas estão sujeitas a diversos tipos de interferências externas que podem ocasionar a interrupção da coleta de dados por longos períodos de tempo ou mesmo por poucos dias. Em função disto, foi necessário utilizar algumas técnicas de pré-processamento, como a limpeza dos dados.

Em um primeiro momento, o critério definido para selecionar as boias que iriam prover as séries temporais de TSM foi a proximidade com a costa brasileira. Posto isto, as boias cujas localizações são 0ºN 35ºW, 4ºN 38ºW, 8ºS 34ºW, 14ºS 32ºW e 19ºS 34ºW foram selecionadas. Além da localização, o segundo critério utilizado na escolha das boias foi a disponibilidade de dados das boias.

Ao analisar a série histórica de TSM de todas as boias que encontravam-se próximas à costa do Brasil, observou-se um período equivalente a sete anos em que cinco boias apresentavam a série com poucas interrupções, onde duas estão situadas no hemisfério norte – 0ºN 35ºW e 4ºN 38ºW, e outras três no hemisfério sul – 8ºS 34ºW, 14ºS 32ºW e 19ºS 34ºW.

A Figura V.1 apresenta a série relativa ao período de novembro de 2005 a outubro de 2012 de cada uma das boias citadas acima, de forma que as interrupções são representadas quando a série toca no eixo das abscissas, isto é, quando o valor da TSM é zero.

A partir da Figura V.1 é possível observar que ambas as boias situadas no hemisfério norte (0ºN 35ºW e 4ºN 38ºW) e a boia 8ºS 34ºW situada no hemisfério sul apresentaram diversos momentos de interrupção no recebimento de dados, tornando suas séries temporais descontínuas. Já ao analisar os gráficos das boias 14ºS 32ºW e 19ºS 34ºW, percebe-se que apesar da série temporal da boia 14ºS 32ºW apresentar alguns poucos períodos de indisponibilidade de dados, no ano de 2011 há uma interrupção na série equivalente a um período de cinco meses e meio sem observações de TSM, o que é uma perda considerável para a série.

Com isso, a boia 19ºS 34ºW se mostrou a mais adequada para os experimentos propostos nesta dissertação, uma vez que esta boia apresentou apenas quatro observações faltantes no período de sete anos analisado e, por isso, foi selecionada.

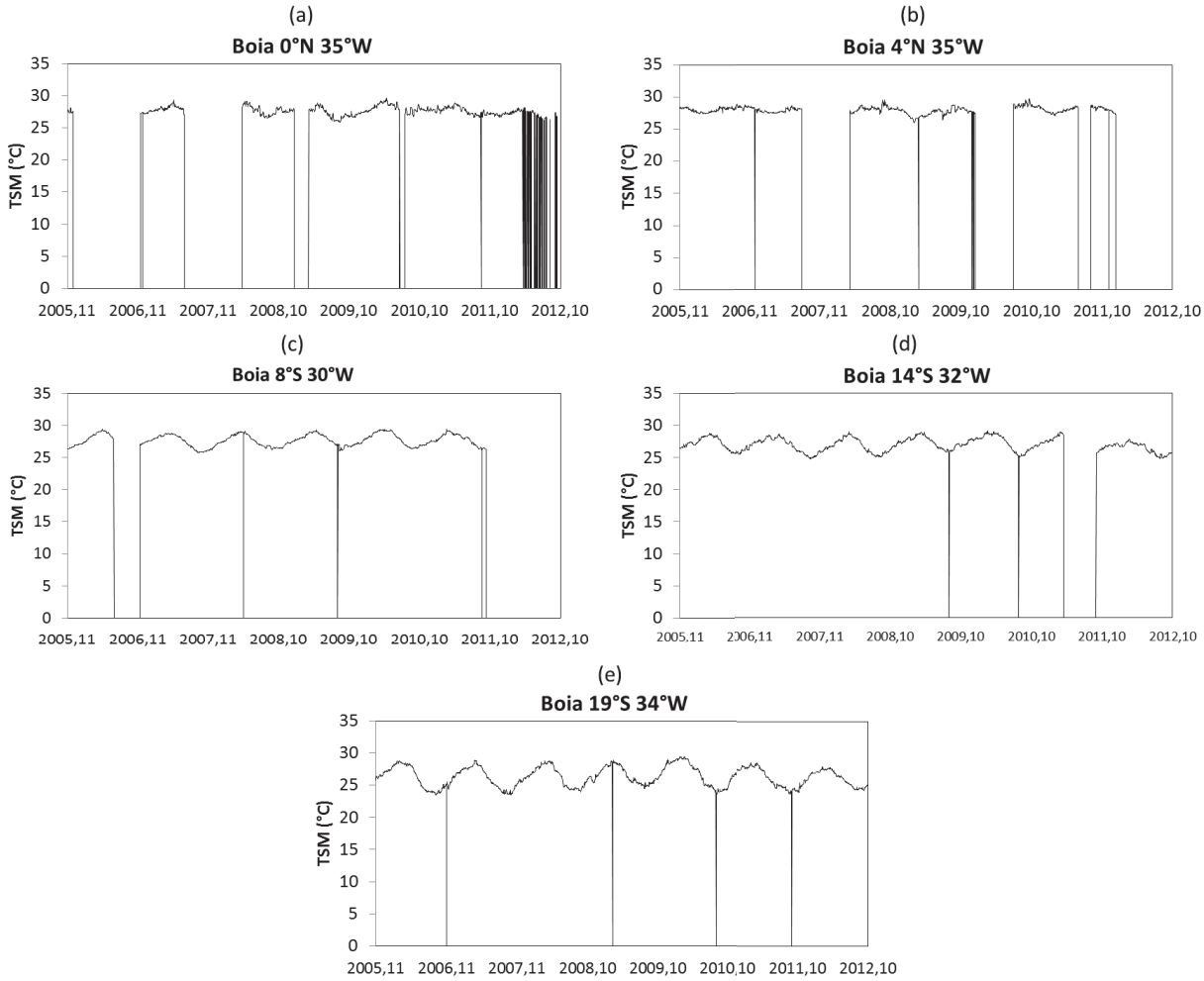


Figura V.1 Séries temporais das boias mais próximas à costa brasileira (a) 0ºN 35ºW, (b) 4ºN 38ºW, (c) 8ºS 34ºW, (d) 14ºS 32ºW (e) 19ºS 34ºW.

A fim de solucionar o problema da indisponibilidade de dados apresentada pela boia 19ºS 34ºW foram testadas duas técnicas diferentes: a reposição dos dados faltantes a partir do valor médio de toda a série e do preenchimento destes valores pela média dos valores anteriores e posteriores aos dados faltantes. No entanto, dado que a série apresenta uma variação de aproximadamente 4°C, a série obtida ao ser utilizada a média global da série não se mostrou satisfatória e, por isso, as observações faltantes foram preenchidas pela média entre o valor de TSM do dia anterior e do dia posterior à data em que o dado não foi coletado. Isto é, o valor do dia faltante pode ser calculado a partir da seguinte fórmula:

$$x_t = \frac{(x_{t-1} + x_{t+1})}{2} \qquad (16)$$

A Figura V.2 abaixo mostra a série temporal diária da boia 19ºS 34ºW correspondente ao período de novembro de 2005 a outubro de 2012 após o preenchimento das observações faltantes de TSM.

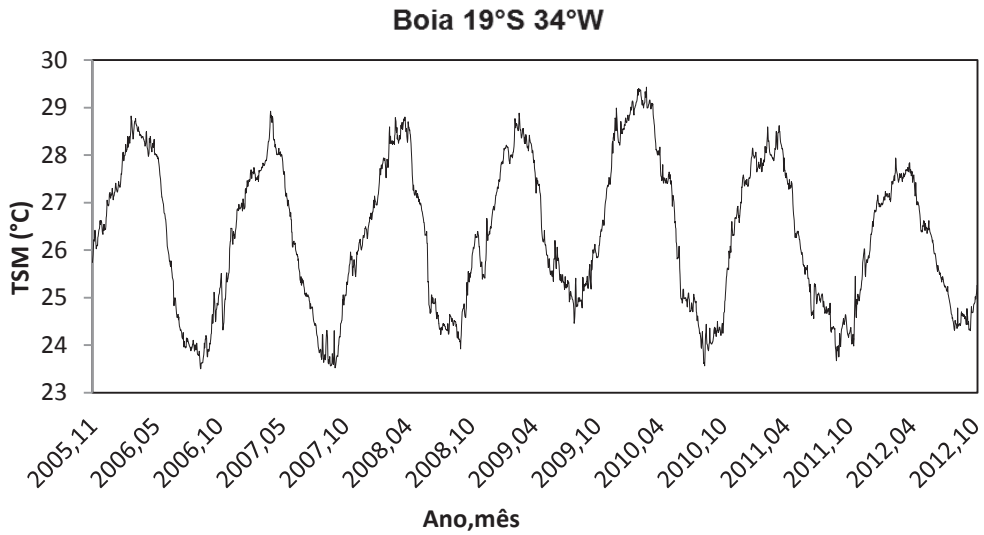


Figura V.2 Série de TSM da boia 19ºS 34ºW após os procedimentos de seleção e limpeza de dados.

V.1.2 Mudança de periodicidade

A mudança na periodicidade de uma série consiste em transformar as observações de alta frequência em observações de menor frequência por meio de técnicas de agregação. No caso do projeto PIRATA, os dados de TSM coletados e disponibilizados são todos diários Dessa forma, pode-se realizar a mudança de periodicidade para menor frequência, como observações semanais, mensais, trimestrais, semestrais e até mesmo anuais.

A realização da transformação da série de TSM foi viável, pois a partir da fonte de dados do PIRATA é possível ter, entre outras, as informações sobre o dia, o mês e o ano em que uma determinada observação foi coletada, o que foi essencial para a transformação dos dados neste experimento.

Dessa forma, levando em consideração as informações das observações fornecidas pelo PIRATA, cada observação da série transformada em dados mensais foi obtido a partir da média aritmética do período correspondente a um mês. Por exemplo, a série mensal correspondente ao mês de maio de 2006 foi obtida a partir da média dos valores dos 31 dados diários correspondentes à este mês e ano. A equação (16) apresenta a fórmula utilizada para a mudança de periodicidade para dados mensais.

$$x_j = \frac{(x_{j,1} + \cdots + x_{j,n_j})}{n_j} \quad (17)$$

Onde x_j é o valor da série mensal, $x_{j,i}$ é o valor do *i-ésimo* dia do mês *j* e n_j é a quantidade de dias no mês em questão.

Já as séries semanais foram obtidas a partir da média aritmética calculada a cada intervalo de sete valores diários, iniciando-se no primeiro dado da série diária, que corresponde ao dia 01 de novembro de 2005. Isto é, o primeiro dado da série semanal foi obtido através da média dos sete primeiros valores diários de TSM, o segundo a partir da média do oitavo ao décimo quinto valor e assim sucessivamente até o final da série de TSM da boia 19ºS 34ºW. Após a transformação, foram obtidas três séries distintas, cujas quantidades de observações em cada uma é apresentada na Tabela V.1.

Tabela V.1 Número de observações das séries diárias, semanais e mensais

Série diária	2557
Série semanal	365
Série mensal	84

V.1.3 Transformação da série em estacionária

Um passo importante para a preparação do experimento é a transformação das séries temporais não-estacionárias em séries estacionárias (GUJARATI; PORTER, 2008). Neste experimento optou-se por aplicar a remoção de tendência (TSAY, 2001). Esta etapa é importante uma vez que os modelos adotados esperam como entrada séries estacionárias.

A transformação das séries diárias, semanais e mensais em séries estacionárias se deu em três estágios. Primeiramente, calculou-se a regressão linear simples para cada uma das séries de dados (diária, semanal e mensal). Em seguida, subtraiu-se o valor da regressão das respectivas séries, obtendo como resultado o chamado resíduo, que corresponde à série sem tendência. A previsão é feita em cima da série de resíduos. Por último, a tendência é reaplicada. Neste caso, o valor previsto de resíduos é somado a regressão linear, obtendo assim o valor da previsão de TSM.

É possível concluir, portanto, que a remoção de tendência da série representa um passo importante para ambos os modelos ARIMA e RNA, pois é a partir da série de resíduos que as previsões serão realizadas. O detalhamento dos procedimentos realizados após a preparação dos experimentos para cada um dos modelos é descrito nos próximos itens.

V.1.4 Particionamento dos dados

O particionamento dos dados em duas amostras, uma para treinamento e outra para teste, é uma etapa importante para a preparação do experimento e modelagem do problema. Enquanto a amostra de treinamento é utilizada para o ajustamento do modelo de previsão, a amostra de teste é utilizada para avaliar suas previsões.

Tendo isto em vista, os dados coletados pelo PIRATA e que alimentam ambas as técnicas de previsão foram segmentados em duas partes. A proporção dos valores destinados a cada uma das partes não exige uma regra para realizar a divisão (HAN; KAMBER; PEI, 2011). No entanto, conforme detalhado no Capítulo IV, a recomendação é que a maior parcela das observações seja utilizada para o treinamento do modelo, enquanto uma parcela menor deve ser utilizada para testá-lo, ou seja, para avaliar as previsões realizadas pelo modelo em questão.

Dessa forma, a partir do total de observações equivalente ao período de sete anos para cada uma das três séries temporais (diária, semanal e mensal), destinou-se uma quantidade de dados equivalente à seis anos para o conjunto de dados de treinamento e um ano para o conjunto de dados de teste. A Tabela V.2 apresenta o resumo da quantidade de dados destinada a cada um dos conjuntos para as séries diárias, semanais e mensais.

Tabela V.2 Particionamento dos dados das séries de dados diários, semanais e mensais entre os conjuntos de teste e de treinamento.

Amostra / Série	Treinamento	Teste	Total
Diária	2191	366	2557
Semanal	313	52	365
Mensal	72	12	84

Adicionalmente, a Tabela V.3 mostra informações estatísticas sobre os conjuntos de treinamento das séries diária, semanal e mensal. É possível perceber que a transformação dos dados não causou grandes alterações nas propriedades estatísticas dos dados. Outra observação importante é que as séries são não-estacionárias, isto é, em diferentes amostras, a média e variância não são constantes.

Tabela V.3 Estatística descritiva do conjunto de treinamento correspondente ao período de 6 anos para os dados diários, semanais e mensais.

Ano	Série	Mínimo (°C)	Máximo (°C)	Média (°C)	Desvio Padrão (°C)
1	Diário	23,50	28,82	26,26	1,73
	Semanal	23,59	28,68	26,26	1,74
	Mensal	23,81	28,49	26,27	1,78
2	Diário	23,52	28,92	26,05	1,56
	Semanal	23,63	28,80	26,05	1,57
	Mensal	23,80	28,26	26,06	1,60
3	Diário	23,92	28,80	26,27	1,48
	Semanal	24,09	28,72	26,28	1,49
	Mensal	24,34	28,51	26,28	1,50
4	Diário	24,46	28,88	26,52	1,23
	Semanal	24,69	28,66	26,52	1,23
	Mensal	25,02	28,54	26,53	1,25
5	Diário	23,56	29,43	26,79	1,91
	Semanal	23,87	29,36	26,77	1,91
	Mensal	24,01	29,15	26,91	1,84
6	Diário	23,67	28,62	26,31	1,46
	Semanal	23,90	28,39	26,31	1,47
	Mensal	24,15	28,14	26,32	1,49

V.2 Experimento comparativo entre os modelos RNA e ARIMA

Os modelos RNAs e ARIMA são conhecidos por suas capacidades de realizar boas previsões. Este primeiro experimento envolvendo a comparação entre as duas técnicas diferentes tem como objetivo observar o desempenho dos modelos ARIMA e RNA quando são utilizados dados transformados de TSM e, dessa forma, avaliar qual dos dois modelos consegue realizar melhores previsões com os dados diários, semanais e mensais.

Salienta-se que o objetivo deste experimento não é realizar previsões de longo prazo, mas ter uma prévia sobre o comportamento de cada um dos modelos com os dados transformados, a fim de que, em um segundo momento possam ser realizadas previsões com maiores horizontes.

Sendo assim, definiu-se que o horizonte de previsão deveria ser curto: para os modelos utilizando dados diários adotou-se o horizonte correspondente ao período de uma semana; para o caso dos dados semanais este horizonte adotado foi de um mês e, por fim, o horizonte previsto para modelos utilizando dados mensais adotado foi de três meses.

A seguir, os procedimentos realizados durante a modelagem das redes neurais e da modelagem do ARIMA são descritos e os resultados obtidos pelos modelos para as previsões das séries diárias, semanais e mensais são analisados e discutidos.

V.2.1 Modelo RNA

A modelagem das Redes Neurais Artificiais envolve uma série de definições que podem impactar nos resultados obtidos por esse método de previsão. Conforme já fora abordado no Capítulo III desta dissertação, não existe uma maneira mais eficiente para a definição das configurações da RNA, sendo muitas vezes o método de tentativa e erro a maneira mais indicada para fazê-lo (OGASAWARA *et al.*, 2009). Tendo isto em vista, com auxílio do *software* WEKA, foram testadas diferentes topologias para as RNAs e as configurações dos parâmetros do modelo que apresentaram os melhores resultados estão descritas na Tabela V.4.

Tabela V.4 Arquitetura das Redes Neurais Artificiais para os dados diários, semanais e mensais.

Dados	Nº nós da camada de entrada	Nº nós da camada escondida	Nº de nós da camada de saída
Diário	7	4	1
Semanal	4	2	1
Mensal	3	2	1

Devido ao horizonte previsto não ser muito extenso, adotou-se a prática comum de que a quantidade de nós da camada de entrada para cada um dos experimentos realizados deveria ser igual ao seu horizonte de previsão (HAYKIN, 2008).

Já o número de nós que compõem a camada escondida, bem como a quantidade de camadas que a compõem, foram estabelecidos através do método de tentativa e erro, de forma que as configurações que apresentaram resultados com os menores erros foram selecionadas. Além disso, todos os três modelos usam um nó de saída (ZHANG, 2000).

A quantidade de nós na camada de saída do modelo RNA foi definida com base no tipo de previsão adotado, conhecido como *one-step-ahead*. Neste tipo de previsão, os valores previstos são iterativamente utilizados para novas previsões, o dado previsto é incorporado à amostra de treinamento e utilizado para prever o dado seguinte. A Figura V.3 apresenta a arquitetura da rede para os dados diários, semanais e mensais.

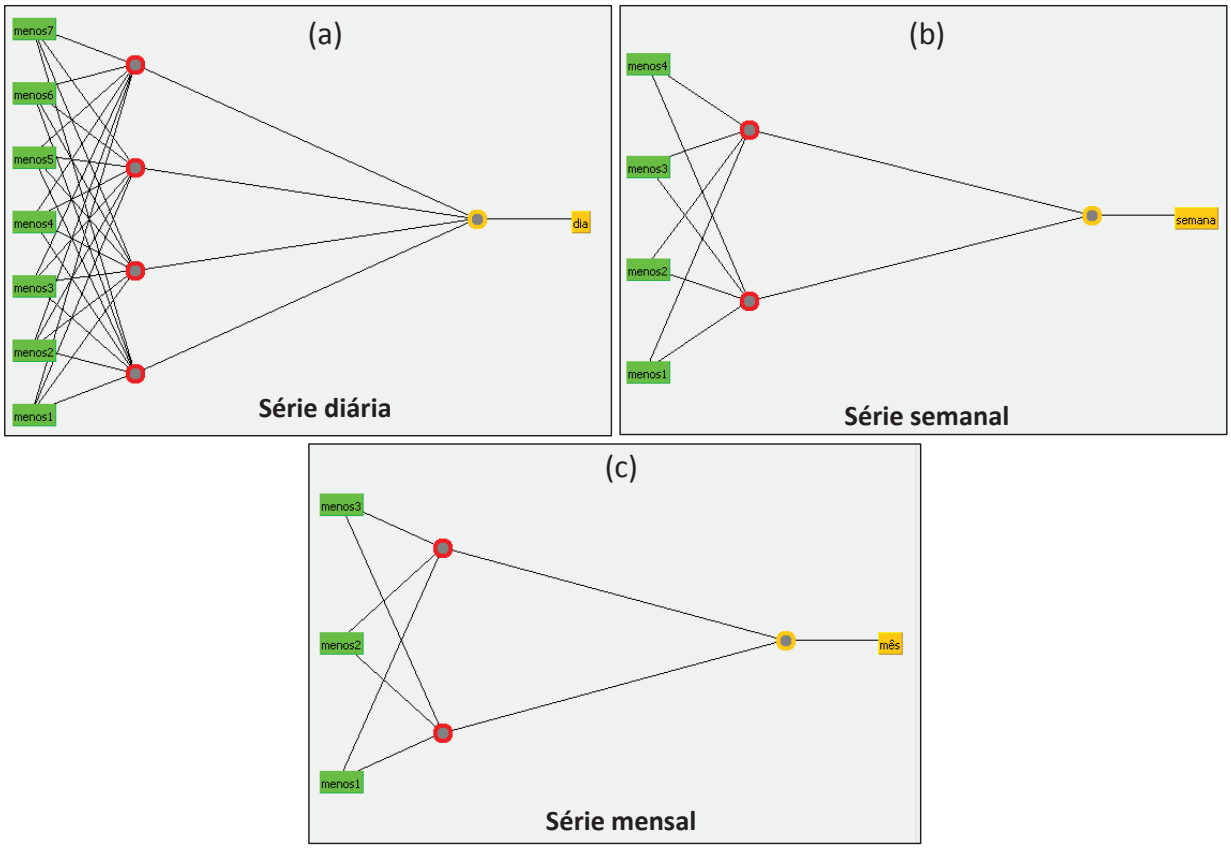


Figura V.3 Arquitetura da rede do modelos para as séries dos dados diários(a), semanais(b) e mensais(c).

Além de definir a chamada arquitetura da rede, existem outras configurações que precisam ser definidas para as RNAs, como é caso do algoritmo de treinamento e a taxa de aprendizagem.

De uma maneira geral, as configurações para os modelos RNA de dados diários, semanais e mensais foram parecidas. O algoritmo de aprendizagem definido para cada um deles foi o mesmo, *MPL* com *backpropagation*. Além disso, as funções de ativação foram a sigmoide, para os pesos dos nós das camadas escondidas e a função linear, para o peso dos das camadas de saída.

As maiores diferenças entre os modelos se deram nas definições das taxas de aprendizagem, do momento e da quantidade de épocas, ou seja, rodadas de treinamento do modelo. A taxa de aprendizagem refere-se à velocidade com que os pesos das ligações entre os nós são alterados de modo permitir que o mínimo global para o erro seja encontrado mais rapidamente, enquanto o momento complementa a função da taxa de aprendizagem, na medida em que procura evitar que a rede neural se torne instável com o aumento da taxa de aprendizagem. Em todos os casos, os valores foram obtidos a partir de análises de sensibilidade, onde foram testadas diferentes opções e observados os impactos das mudanças

nos resultados. A Tabela V.5 apresenta o resumo das principais configurações dos modelos RNA utilizando dados diários, semanais e mensais.

Tabela V.5 Configurações das Redes Neurais Artificiais para os dados diários, semanais e mensais.

Dados	Taxa de Aprendizagem	Momento	*Epochs*
Diário	0,1	0,05	2500
Semanal	0,01	0,01	2500
Mensal	0,05	0,005	2000

Com base nos resultados apresentados no Quadro 7, é possível observar que os valores assumidos pela taxa de aprendizagem e pelo momento foram maiores para os dados diários e menores para os dados mensais. Este resultado pode ser justificado pelo tamanho da amostra de treinamento de cada um dos conjuntos de dados, pois uma vez que a série mensal apresenta um conjunto consideravelmente menor do que o apresentado pela série diária, um valor maior assumido pela taxa de aprendizagem poderia levar o modelo mais facilmente ao mínimo local e não ao mínimo global, como é desejado que aconteça. Por outro lado, caso uma taxa de aprendizado muito baixa fosse assumida pelo modelo utilizado a série diária, o mínimo global poderia não ser obtido.

V.2.2 Modelo ARIMA

É possível afirmar que a modelagem do ARIMA é relativamente mais simples do que a modelagem de uma RNA. Diferentemente das RNAs, onde uma grande quantidade de variáveis deve ser definida, nos modelos ARIMA essas variáveis correspondem apenas aos parâmetros *p, d e q,* que dizem respeito, respectivamente, ao número de termos auto-regressivos, ao número de termos de médias móveis e ao número de vezes que a série temporal precisa sofrer diferenciação antes de tornar-se estacionária.

Devido aos horizontes previstos serem relativamente curtos, determinou-se que o valor assumido por *p* deveria ser igual ao horizonte de previsão definido para cada um dos experimentos realizados. Dessa forma, o parâmetro *p* do modelo ARIMA (p,d,q) assume valores iguais a 7, 4 e 3, respectivamente para as séries diária, semanal e mensal.

Além disso, por simplificação, estabeleceu-se que o parâmetro *d* assumiria o valor de zero e que, devido ao fato de as séries já passarem pelo procedimento de remoção de tendência durante a etapa de preparação do experimento, o valor assumido por *q* também deveria ser igual à zero. Sendo assim, apenas a parte correspondente ao AR do modelo ARIMA (p,d,q) foi explorada durante a sua modelagem nesta dissertação.

Outro ponto que se diferencia do experimento das RNAs é o software utilizado. Enquanto a modelagem e previsão das RNAs foram realizadas no WEKA, o modelo ARIMA foi implementado no software MATLAB, onde suas previsões também foram realizadas.

O tipo de previsão utilizado pelo modelo ARIMA, assim como fora definido para as RNAs, corresponde ao *one-step-ahead*, onde os valores previstos são iterativamente utilizados nas previsões seguintes, conforme o horizonte de previsão vai sendo aumentado.

V.2.3 Análise dos resultados

Os resultados das previsões obtidos pelos modelos RNA e ARIMA anteriormente descritos utilizando os dados diários, dados semanais e dados mensais são discutidos nos itens a seguir. Devido ao fato de existirem três séries com periodicidades diferentes, o resultado da previsão para cada uma das séries será abordada separadamente. É válido ressaltar, mais uma vez, que o foco da análise deste primeiro experimento é, principalmente, avaliar o desempenho das duas técnicas para cada série de dados de TSM utilizada.

V.2.3.1 Previsões utilizando dados diários

As previsões realizadas com dados diários corresponderam ao período de sete dias à frente. Nesta previsão, ambos os modelos ARIMA e RNA não se saíram bem, considerando a variação do valor da série real apresentada no período. A Tabela V.6 apresenta os resultados para cada horizonte previsto, em dias, utilizando as medidas de desempenho definidas no Capítulo III.

Tabela V.6 Resumo dos resultados das previsões utilizando dados diários.

Dias	Modelo	EMA	EMPA	SEQ	EMQ	REMQ
1	ARIMA	0,12	0,48%	0,01	0,01	0,12
	RNA	0,03	0,12%	0,00	0,00	0,03
2	ARIMA	0,20	0,84%	0,10	0,05	0,22
	RNA	0,15	0,61%	0,07	0,04	0,19
3	ARIMA	0,26	1,08%	0,24	0,08	0,28
	RNA	0,18	0,73%	0,12	0,04	0,20
4	ARIMA	0,29	1,19%	0,37	0,09	0,30
	RNA	0,21	0,89%	0,24	0,06	0,24
5	ARIMA	0,30	1,26%	0,51	0,10	0,32
	RNA	0,20	0,84%	0,26	0,05	0,23
6	ARIMA	0,29	1,21%	0,56	0,09	0,31
	RNA	0,21	0,85%	0,31	0,05	0,23
7	ARIMA	0,28	1,15%	0,60	0,09	0,29
	RNA	0,20	0,82%	0,33	0,05	0,22

A Figura V.4 mostra a série prevista utilizando os modelos ARIMA e RNA quando comparada aos valores reais de TSM. A partir da Figura V.4 é possível perceber um comportamento linear apresentado na previsão do ARIMA, enquanto para as RNAs os valores das previsões ficaram oscilando na tentativa de prever o comportamento da série real. Neste primeiro resultado, observa-se que as RNAs conseguiram prever com maior acurácia que o modelo ARIMA, entretanto, considerando que a série real apresenta uma variação de aproximadamente 0,3°C no período avaliado, mesmo as RNAs não apresentaram bons resultados..

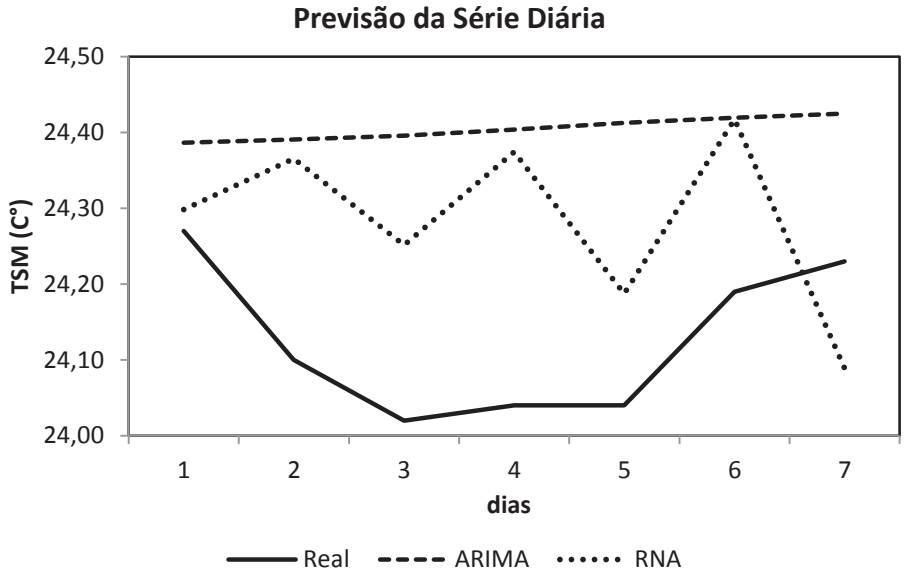


Figura V.4 Comparação entre as séries resultantes das previsões utilizando dados diários e a série real de TSM.

V.2.3.2 Previsões utilizando dados semanais

Nesta segunda previsão utilizando dados semanais, ao contrário do ocorrido no experimento anterior, o modelo ARIMA apresentou resultados mais acurados do que os obtidos com as RNAs. A Tabela V.7 apresenta os resultados para o horizonte previsto de quatro semanas. Pode-se observar que, conforme o horizonte vai aumentando, os erros das redes neurais e do ARIMA também aumentam.

Tabela V.7 Resumo dos resultados das previsões utilizando dados semanais.

Horizonte (semanas)	Modelo	EMA	EMPA	SEQ	EMQ	REMQ
1	ARIMA	0,12	0,48%	0,01	0,01	0,12
	RNA	0,45	1,86%	0,20	0,20	0,45
2	ARIMA	0,20	0,84%	0,10	0,05	0,22
	RNA	0,46	1,90%	0,42	0,21	0,46
3	ARIMA	0,26	1,08%	0,24	0,08	0,28
	RNA	0,55	2,26%	0,94	0,31	0,56
4	ARIMA	0,29	1,19%	0,37	0,09	0,30
	RNA	0,52	2,14%	1,12	0,28	0,53

A Figura V.5 abaixo mostra que a previsão realizada pelos modelos ARIMA e RNA foram parecidas, com uma pequena vantagem para o ARIMA. Neste caso, é possível observar que os modelos mostraram dificuldades para acompanhar o comportamento de crescimento da série real.

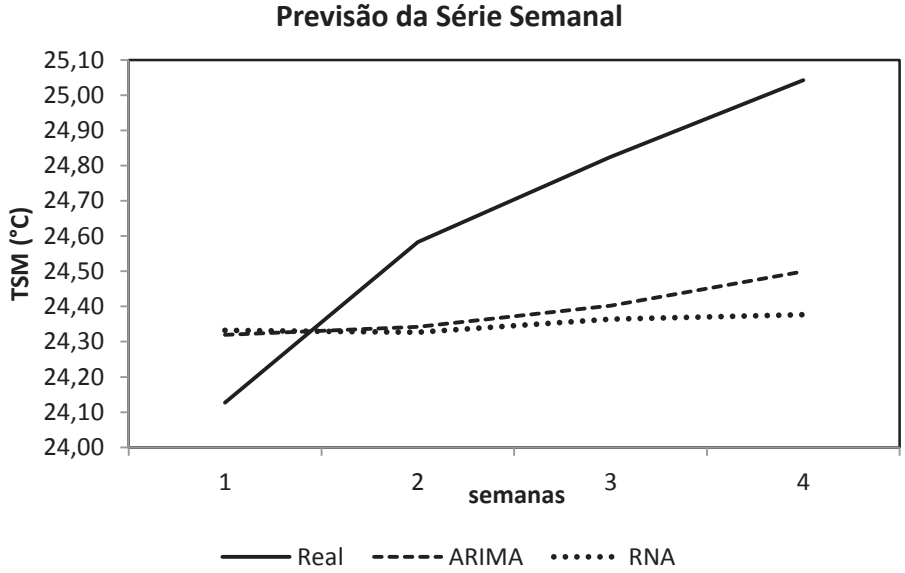


Figura V.5 Comparação entre as séries resultantes das previsões utilizando dados semanais e a série real de TSM.

V.2.3.2 Previsões utilizando dados mensais

Os resultados obtidos utilizando os dados mensais mostraram-se mais conclusivos que os anteriormente demonstrados para os dados diários e semanais. Neste experimento, a RNA apresentou seu pior desempenho quando comparado com os experimentos anteriores. Já o ARIMA, em contraposição, apresentou seu melhor resultado, sendo mais uma vez superior à RNA.

Embora não tenha havido muita diferença entre os resultados obtidos pelas técnicas tanto para os dados diários quanto para os dados semanais, no caso dos dados mensais houve uma significativa vantagem do ARIMA em relação às RNAs. A Tabela V.8 apresenta um resumo dos resultados das previsões obtidas pelos modelos ARIMA e RNA para os horizontes de um a três meses de previsão.

A partir da Tabela V.8 é possível observar também que o desempenho do modelo ARIMA apresenta um comportamento de melhoria com o incremento no horizonte de previsão. Este é um resultado muito favorável, uma vez que é interessante do ponto de vista de previsão ter horizontes mais longos.

Tabela V.8 Resumo dos resultados das previsões utilizando dados mensais.

Horizonte (meses)	Modelo	EMA	EMPA	SEQ	EMQ	REMQ
1	ARIMA	0,37	1,49%	0,14	0,14	0,37
	RNA	0,26	1,06%	0,07	0,07	0,26
2	ARIMA	0,27	1,08%	0,17	0,08	0,29
	RNA	0,41	1,60%	0,19	0,37	0,61
3	ARIMA	0,21	0,83%	0,17	0,06	0,24
	RNA	0,70	2,66%	0,67	2,02	1,42

A Figura V.6 reafirma a superioridade da previsão realizada com o modelo ARIMA. É possível perceber que, ao contrário dos resultados anteriores, o ARIMA conseguiu reconhecer o comportamento da série real e realizar assim uma boa previsão. Na Figura V.6 é bastante evidente a aproximação cada vez maior da previsão do ARIMA com a série real e o afastamento da RNA em relação também a série real.

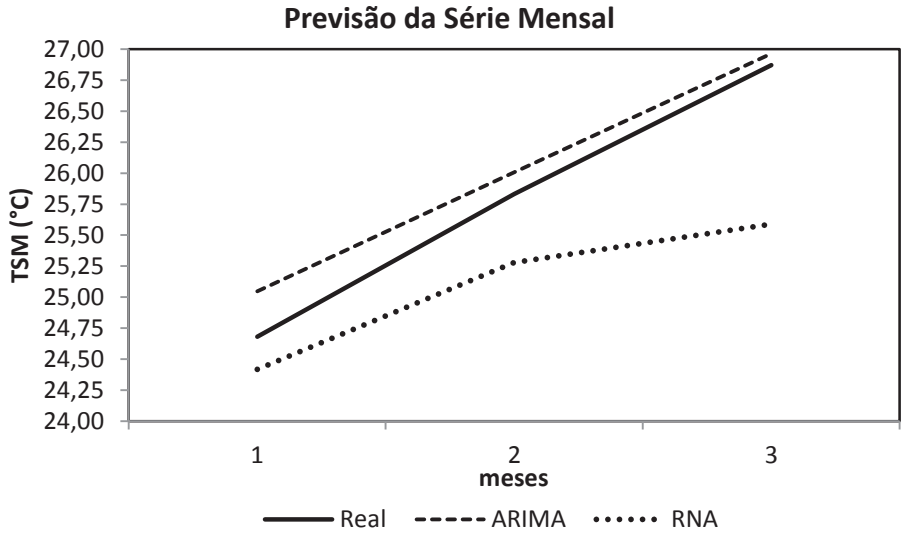


Figura V.6 Comparação entre as séries resultantes das previsões utilizando dados mensais e a série real de TSM.

A partir da análise dos resultados utilizando dados diários, semanais e mensais é possível concluir que o ARIMA foi superior às RNAs nos experimentos realizados, uma vez que apresentou melhores resultados em dois dos três casos analisados.

Dessa forma, como o principal objetivo deste primeiro experimento era avaliar o comportamento dos modelos em relação aos dados transformados, o ARIMA apresentou-se como o mais bem adaptado à mudança de periodicidade dos dados. Além disso, outro fator positivo do ARIMA é que se trata de um modelo relativamente simples, proporcionando maiores facilidades para o entendimento do modelo. Tendo definido o modelo a ser utilizado, na próxima seção deste capítulo são realizadas novas previsões utilizando os dados transformados, mas com horizontes de previsão maiores.

V.3 Modelos ARIMA

Uma vez tendo definido utilizar a técnica ARIMA, um segundo experimento foi desenvolvido. Neste novo experimento o objetivo não é mais comparar técnicas diferentes, mas sim avaliar a influência dos dados utilizados para o resultado da previsão.

A metodologia utilizada neste experimento é semelhante à apresentada no experimento anterior, com apenas alguns ajustes. A alteração mais relevante foi a de que, ao contrário do experimento anterior, onde a quantidade de entradas para o modelo era definida com base no horizonte de previsão, neste segundo experimento devido ao horizonte de previsão ser consideravelmente superior aos definidos anteriormente, foi necessário alterar o critério para a definição da quantidade de entradas do modelo.

Sendo assim, neste segundo experimento, envolvendo a análise de previsões utilizando dados com periodicidade diferente, a quantidade de entradas foi definida com base na análise de autocorrelação da amostra de treinamento de cada um dos conjuntos de dados (diário, semanal e mensal). O objetivo da autocorrelação é obter os *lags*, ou seja, o índice dos autoregressores com módulo de autocorrelação superior a um determinado patamar (Ogasawara *et al.*, 2013), que neste estudo foi definido como sendo 0,7. A Figura V.7 mostra a função de autocorrelação da amostra de dados de treinamento diários, semanais e mensais.

É possível perceber através da Figura V.7 (a) que a amostra de dados diários apresenta a maior autocorrelação entre as três amostras analisada. A Figura V.7 (c) mostra que autocorrelação da amostra de dado mensais apresentou oscilação com o aumento dos *lags*, enquanto Figura V.7 (b) mostra que a autocorrelaçao dos dados semanais também apresentaram uma tendência à oscilação, porém aparentemente mais lenta do que a apresentada pelos dados mensais.

Considerando o critério para a definição da quantidade de entradas como sendo o determinado patamar de 0,7 para o módulo de autocorrelação, a partir da Figura V.7 é possível observar que a quantidade de entradas definida para os modelos ARIMA utilizando dados diários, semanais e mensais corresponde a 20, 6 e 12, respectivamente. Sendo assim, estes são os valores assumidos pelo parâmetro *p* do modelo ARIMA (p,q,d) para cada uma das previsões realizadas para as respectivas séries de dados.

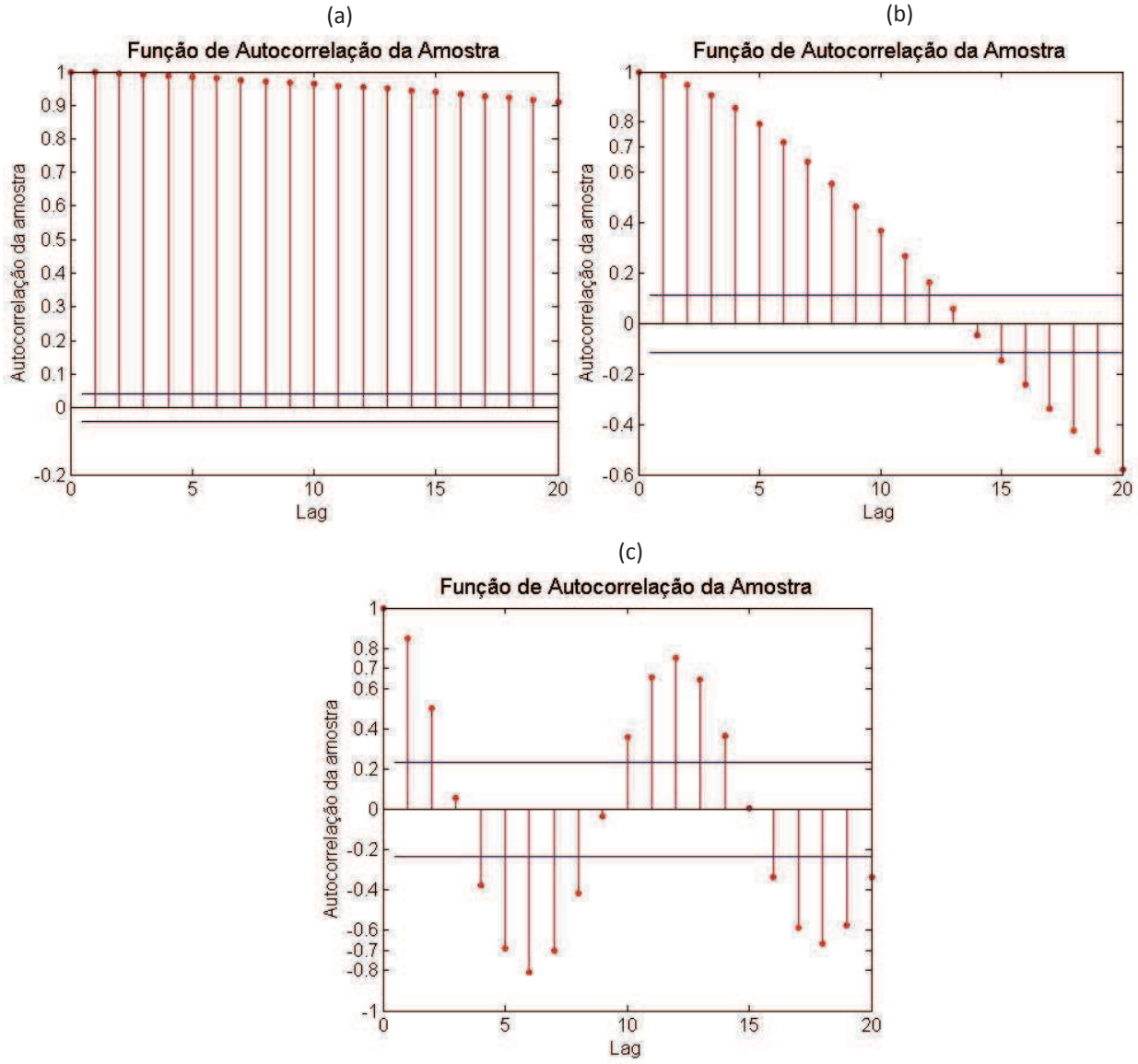


Figura V.7 Função de autocorrelação da amostra de dados de treinamento dos dados diários(a), semanais(b) e mensais(c).

V.3.2 Previsão utilizando dados diários

A previsão utilizando dados diários corresponde àquela que faz uso dos dados obtidos orginalmente no banco de dados do projeto PIRATA, isto é, são os valores de temperatura de superfície do mar registrados pelas boias.

Dessa forma as previsões foram realizadas utilizando a metodologia definida no item IV.2.2 deste capítulo, porém utilizando as configurações apresentadas na Tabela V.9 abaixo. O conjunto de treinamento equivale ao período de seis anos, a quantidade de entradas fora definida com base na função de autocorrelação do conjunto de treinamento e o horizonte de previsão equivale ao período de um ano à frente.

Tabela V.9 Configurações do modelo ARIMA para os dados diários

Conjunto de Treinamento	2191
Quantidade de entradas	20
Horizonte de previsão (dias)	366

A Figura V.8 abaixo mostra a comparação entre os resultados de TSM obtidos utilizando o modelo ARIMA com os dados mensais e a série real de TSM.

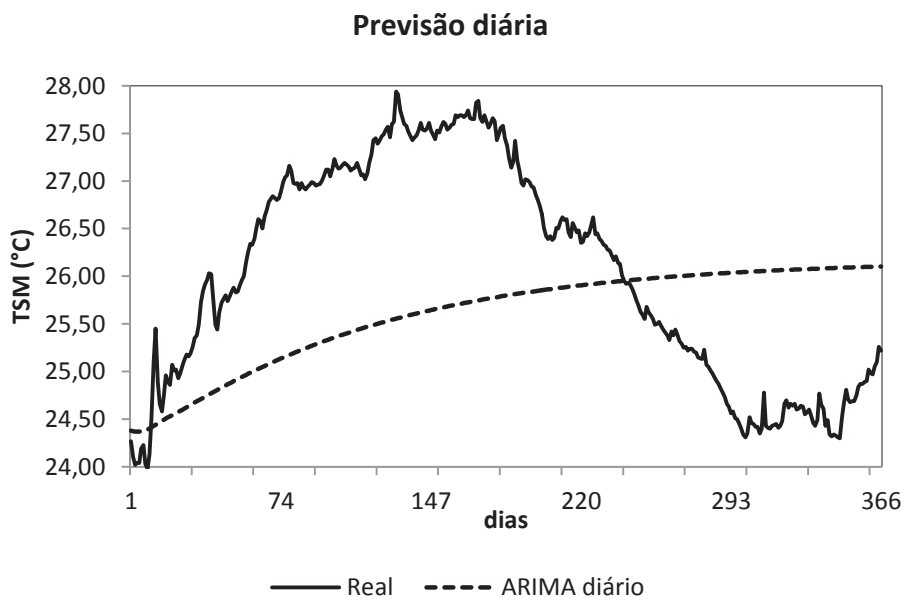


Figura V.8 Comparação entre a série resultante da previsão utilizando o modelo ARIMA com dados diários e a série real de TSM.

Como pode ser observado, os resultados obtidos por este modelo não foram satisfatórios, uma vez que a previsão não conseguiu acompanhar os valores da série real. No entanto, o erro médio percentual absoluto (EMPA) para o horizonte de um ano à frente, por exemplo, foi igual a 4,69%, o que pode ser considerado um resultado razoável. A Tabela V.10 apresenta os resultados das medidas de desempenho para 92, 182, 274 e 366 dias, o que

corresponde aproximadamente a uma previsão diária realizada durante o período de três, seis, nove e doze meses.

Tabela V.10 Resultados do modelo ARIMA utilizando dados diários.

Horizonte (dias)	EMA	EMPA	SEQ	EMQ	REMQ
92	1,04	3,94%	128,63	1,40	1,18
182	1,45	5,37%	447,03	2,46	1,57
274	1,18	4,37%	496,23	1,81	1,35
366	1,23	4,69%	681,61	1,86	1,36

V.3.3 Previsão utilizando dados semanais

A previsão utilizando os dados transformados em semanais foi realizada de forma semelhante à descrita para a previsão de dados diários, entretanto com as configurações personalizadas para este caso, que estão apresentadas A Tabela V.11.

Tabela V.11 Configurações do modelo ARIMA utilizando dados semanais.

Conjunto de Treinamento	313
Quantidade de entradas	6
Horizonte de previsão (semanas)	52

O horizonte de previsão, assim como no modelo que utiliza dados diários, é equivalente a um ano, entretanto, devido à mudança de periodicidade realizada durante a preparação do experimento, isto equivale à 52 passos à frente, ao invés dos 366 passos do modelo diário.

A Figura V.9 apresenta a comparação entre o resultado obtido através da previsão realizada pelo modelo ARIMA com dados semanais e os valores da série semanal de TSM.

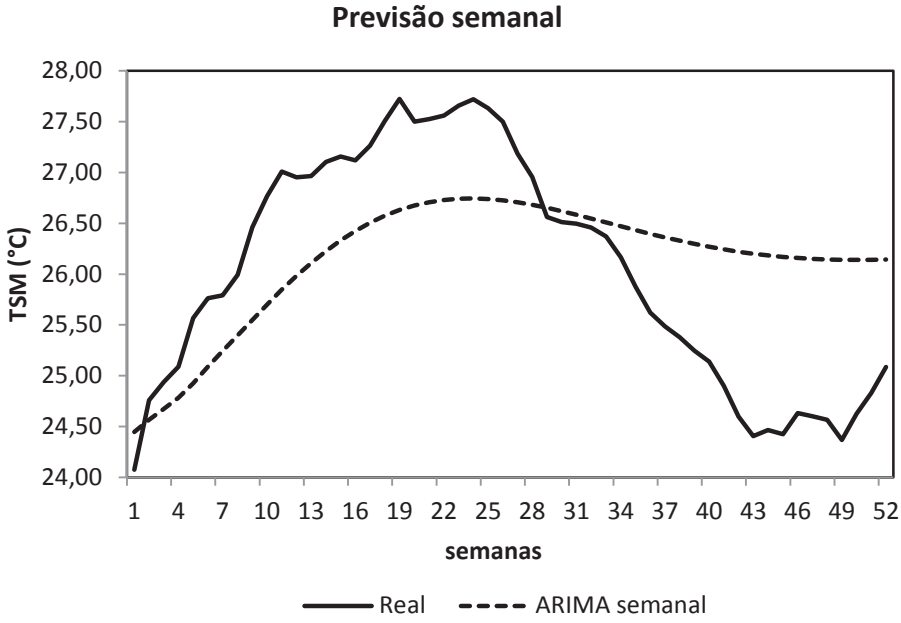


Figura V.9 Comparação entre a série resultante da previsão utilizando o modelo ARIMA com dados semanais e a série real de TSM.

É possível observar através da Figura V.9 que o modelo ARIMA conseguiu perceber a tendência inicial de crescimento da série, no entanto, não descreveu bem a subsequente queda dos valores de TSM da série. Ainda assim, conforme pode ser observado na Tabela V.12, os resultados obtidos por este modelo apresentaram bons resultados de uma maneira geral. A Tabela V.12 apresenta os resultados das medidas de desempenho para o período de 13, 26, 39 e 52 semanas, que correspondem à aproximadamente três, seis, nove e doze meses respectivamente.

Tabela V.12 Resultados do modelo ARIMA utilizando dados semanais.

Horizonte (semanas)	EMA	EMPA	SEQ	EMQ	REMQ
13	0,66	2,51%	6,85	0,53	0,73
26	0,76	2,83%	16,67	0,64	0,80
39	0,66	2,46%	20,81	0,53	0,73
52	0,87	3,38%	51,27	0,99	0,99

V.3.4 Previsão utilizando dados mensais

O horizonte previsto para a previsão utilizando dados mensais, assim como nas previsões diária e semanal, equivale ao período de um ano à frente. Entretanto, é interessante observar que isso equivale a prever 12 passos à frente, ou seja, a quantidade de dados

previstos corresponde a menos de 4% e 25% dos valores previstos para o caso da previsão com dados diários e semanais, respectivamente, e, no entanto, equivale ao mesmo período de previsão de um ano. Ainda que os valores da previsão mensal estejam mais consolidados do que os obtidos com a previsão diária, por exemplo, a mudança de periodicidade e a consequente redução dos dados previstos pode ser interessante para previsões de longo prazo ou pesquisas que não desejem ter a precisão de um valor diário, mas sim o comportamento de uma dada variável no período. A Tabela V.13 apresenta as configurações do modelo ARIMA quando é utilizada a série de dados mensais.

Tabela V.13 Configurações do modelo ARIMA utilizando dados mensais.

Conjunto de Treinamento	72
Quantidade de entradas	12
Horizonte de previsão (meses)	12

O modelo ARIMA utilizando dados mensais apresentou ótimos resultados em suas previsões, como pode ser observado na Figura V.10. Em todos os horizontes previstos a série prevista de TSM apresentou um comportamento bastante semelhante à apresentada pela série dos dados reais, havendo momentos em que os valores foram praticamente coincidentes, como é o caso da previsão para o 12º mês, onde a diferença entre o valor previsto e o valor real foi de apenas 0,015°C.

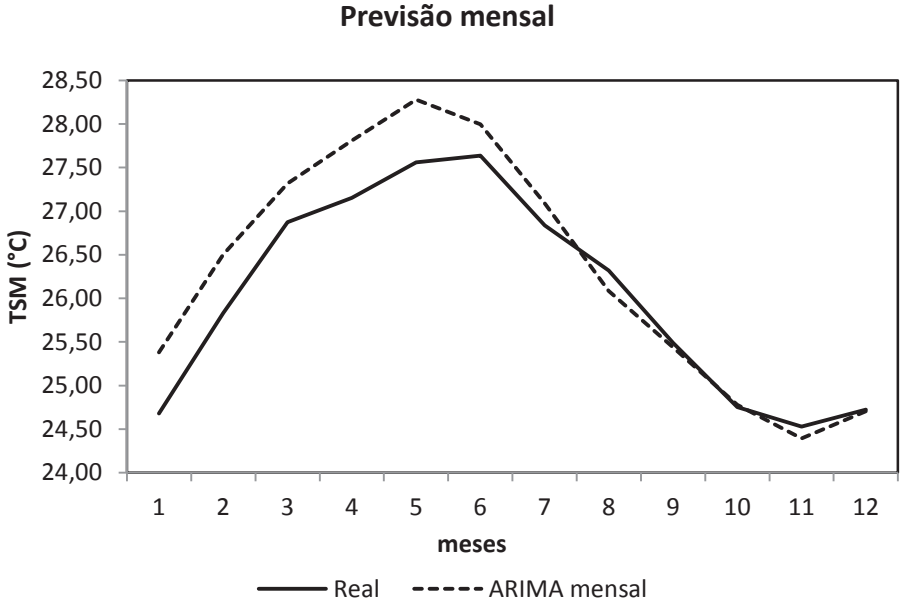


Figura V.10 Comparação entre a série resultante da previsão utilizando o modelo ARIMA com dados mensais e a série real de TSM.

O resultado obtido neste experimento utilizando o modelo ARIMA com dados mensais reforça o resultado do experimento envolvendo a comparação entre os modelos ARIMA e as RNA, onde os resultados utilizando esta série de dados mensais destacaram-se. Além disso, chama-se atenção ao fato de que, ao contrário do que se espera com o aumento do horizonte de previsão, o erro da previsão realizada pelo modelo ARIMA utilizando dados mensais decresceu, conforme mostra a Tabela V.14.

Tabela V.14 Resultados do modelo ARIMA utilizando dados mensais.

Horizonte (meses)	EMA	EMPA	SEQ	EMQ	REMQ
3	0,61	2,36%	1,14	0,38	0,62
6	0,59	2,24%	2,21	0,37	0,61
9	0,45	1,72%	2,34	0,26	0,51
12	0,36	1,35%	2,36	0,20	0,44

V.3.5 Análise dos resultados

Para que seja possível realizar a comparação entre os resultados do modelo ARIMA utilizando as três séries de dados distintas (diária, semanal e mensal), foi necessário realizar um segundo passo neste experimento a fim de transformar os resultados de previsão diária obtidos pelo modelo ARIMA de dados diários em dados semanais e mensais. Para tanto, a mesma metodologia para mudança de periodicidade definida no item IV.1.4 deste capítulo foi aplicada aos dados previstos pelo modelo ARIMA de dados diários.

A transformação dos dados semanais em dados mensais poderia ocasionar perdas na verossimilhança dos resultados, uma vez que não é possível estabelecer uma regra onde fosse definido que um determinado número de semanas equivalesse a um mês. Dessa forma, optou-se por comparar apenas resultados obtidos utilizando a série diária com os resultados dos modelos ARIMA de dados semanais e ARIMA de dados mensais.

Após a transformação dos dados diários previstos em dados semanais, a nova série obtida foi comparada com a série de dados reais e com a série resultante da previsão do modelo ARIMA utilizando dados semanais. A Figura V.11 representa graficamente a comparação destes resultados.

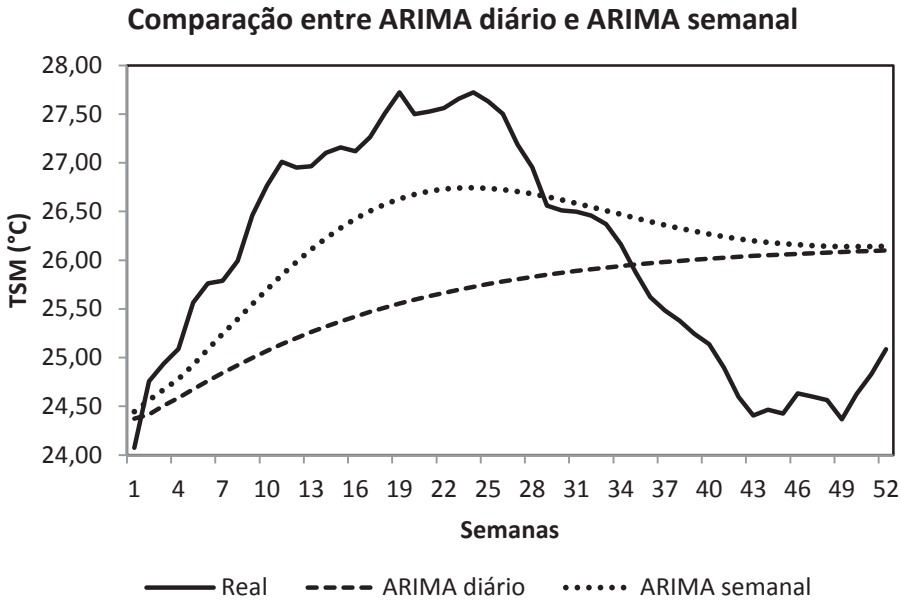


Figura V.11 Comparação entre a série real de TSM e as séries resultantes das previsões utilizando o modelo ARIMA com dados diários e dados semanais.

Percebe-se, a partir da Figura V.11, que os resultados do modelo ARIMA semanal aproximaram-se mais à realidade dos resultados obtidos através do modelo ARIMA diário. A vantagem do modelo ARIMA de dados semanais fica mais evidente quando observados os valores dos erros presentes na Tabela V.15.

Ao observar os resultados apresentados na Tabela V.15 e aqueles apresentados anteriormente na Tabela V.10, é interessante perceber que a transformação dos resultados das previsões diárias em previsões semanais causou pouco impacto nos resultados obtidos pelas medidas desempenho acima, com exceção da soma dos erros quadráticos (SEQ), que por se tratar de um valor cumulativo pode ser justificado pela maior quantidade de passos à frente previstos na previsão diária.

Tabela V.15 Comparação entre os resultados do modelo ARIMA utilizando dados diários e dados semanais.

Horizonte (semanas)	Modelo	EMA	EMPA	SEQ	EMQ	REMQ
13	ARIMA diário	1,08	4,09%	18,99	1,46	1,21
	ARIMA semanal	0,66	2,51%	6,85	0,53	0,73
26	ARIMA diário	1,48	5,47%	65,28	2,51	1,58
	ARIMA semanal	0,76	2,83%	16,67	0,64	0,80
39	ARIMA diário	1,19	4,42%	71,62	1,84	1,36
	ARIMA semanal	0,66	2,46%	20,81	0,53	0,73
52	ARIMA diário	1,24	4,74%	97,84	1,88	1,37
	ARIMA semanal	0,87	3,38%	51,27	0,99	0,99

Assim como realizado para os dados semanais, os dados diários previstos pelo modelo ARIMA foram transformados em dados mensais e comparados com os resultados da previsão utilizando dados mensais e a série de dados reais de TSM. A Figura V.12 representa graficamente a comparação destes resultados.

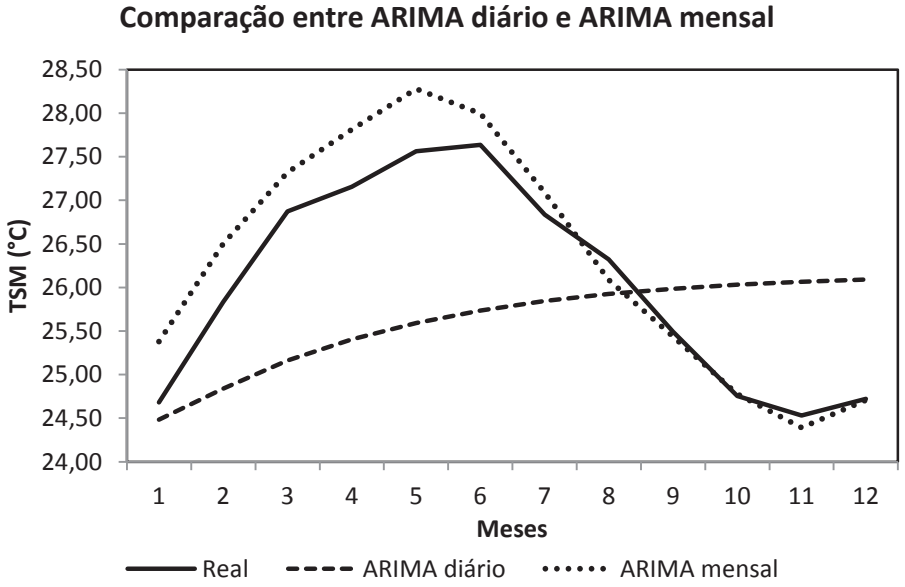


Figura V.12 Comparação entre a série real de TSM e as séries resultantes das previsões utilizando o modelo ARIMA com dados diários e dados mensais.

A Figura V.12, em conjunto com os resultados apresentados na Tabela V.16, ressalta mais uma vez a superioridade das previsões realizadas pelo modelo ARIMA utilizando os dados mensais, que assim como no primeiro experimento envolvendo a comparação entre os modelos ARIMA e RNA, proporcionou os resultados de previsão mais acurados. Salienta-se também o fato de que o único modelo que apresentou tendência de melhoria na previsão para horizontes maiores de previsão foi ARIMA utilizando dados mensais, cujos erros mostram-se decrescentes para as medidas de desempenho EMA, EMPA, EMQ e REMQ.

Tabela V.16 Comparação entre os resultados do modelo ARIMA utilizando dados diários e mensais.

Horizonte (meses)	Modelo	EMA	EMPA	SEQ	EMQ	REMQ
3	ARIMA diário	0,97	3,67%	3,95	1,32	1,15
	ARIMA mensal	0,61	2,36%	1,14	0,38	0,62
6	ARIMA diário	1,42	5,25%	14,50	2,42	1,55
	ARIMA mensal	0,59	2,24%	2,21	0,37	0,61
9	ARIMA diário	1,16	4,29%	15,89	1,77	1,33
	ARIMA mensal	0,45	1,72%	2,34	0,26	0,51
12	ARIMA diário	1,22	4,63%	21,77	1,81	1,35
	ARIMA mensal	0,36	1,35%	2,36	0,20	0,44

Mais uma vez tomando como referência os resultados do modelo ARIMA utilizando dados diários apresentado na Tabela V.16 observa-se que a transformação dos dados após a previsão causaram impactos pouco significativos nos resultados.

Posto que originalmente todos os dados de TSM eram diários, a mudança de periodicidade dos dados diários previstos permite avaliar o impacto nos resultados das previsões da ordem em que a mudança de periodicidade é realizada. Sendo assim, com base nos resultados apresentados, é possível perceber que o modelo ARIMA ao utilizar os dados transformados antes que fosse realizada a previsão apresentou resultados mais satisfatórios.

Os resultados obtidos nos experimentos realizados nesta dissertação mostraram que a transformação da periodicidade dos dados permite que as previsões realizadas para maiores horizontes sejam mais acuradas. No entanto, o momento que esta transformação deve ser realizada é um fator importante, sendo desejável que a transformação ocorra antes da previsão para que melhores resultados sejam obtidos.

Observa-se também que as séries mensais apresentaram um desempenho melhor. Muito possivelmente esta melhoria de desempenho esteja relacionada a consideração de aspectos como sazonalidade que podem ser vistos nos diagramas de autocorrelação da Figura V.7. (c), onde podem ser observados *lags* de 12 termos. Esse resultado é interessante, visto que em previsões de eventos extremos é necessária uma previsão com um horizonte de tempo maior para que as comunidades e autoridades governamentais possam se preparar melhor para os possíveis danos causados por estes eventos.

Capítulo VI - Conclusão

A ocorrência de eventos extremos pode causar sérios impactos na vida da população e, por isso, embora em muitos casos não seja possível evitá-los, é muito importante estar preparado para quando estes vierem a ocorrer, a fim de mitigar seus efeitos muitas vezes catastróficos. E é nesse ponto que a comunidade científica vem se empenhando em dar sua contribuição. Apesar de muito estar sendo estudado sobre os eventos extremos e investido para melhor prevê-los, as fatalidades que continuam ocorrendo até hoje são prova de que ainda há muito a ser feito e que é possível continuar melhorando. O Brasil é um bom exemplo disto. Apesar de as secas serem um evento recorrente na região Nordeste do Brasil, ano após ano, centenas de famílias sofrem com seus impactos, precisando se deslocar para outras cidades e deixar seus bens para trás ou sofrerem com a escassez de água.

Alguns eventos extremos estão relacionados a ocorrência de mudanças em indicadores climáticos no oceano. Parâmetros como a temperatura da superfície do mar (TSM) são estudados e relacionados com a ocorrência de eventos extremos próximo ao litoral, e em particular ao litoral norte e sul do Brasil que é banhado pelo Oceano Atlântico. O projeto *Prediction and Research Moored Array in the Tropical Atlantic* (PIRATA), fruto de uma parceria entre Brasil, França e EUA, é composto por um conjunto de boias localizadas no Oceano Atlântico que foram colocadas com a proposta de gerar dados relacionados a diversos parâmetros do oceano. O objetivo é que estes dados possam contribuir com as pesquisas climáticas sobre a parte tropical do Oceano Atlântico, que é relativamente pouco estudado quando comparado com os Oceanos Índico e Pacífico. Além disso, ao obter informações sobre o Oceano Atlântico Tropical, o PIRATA proporciona uma visão mais integrada do que acontece no sistema oceano-atmosférico de todo planeta, uma vez que preenche este espaço de estudo ainda pouco desenvolvido.

Um dos assuntos largamente estudados no que concerne aos eventos extremos são as técnicas utilizadas para prevê-los com cada vez maior antecedência, visto que as previsões permitem que sejam tomadas medidas mitigadoras para os impactos dos eventos extremos. Neste sentido, o uso de métodos de previsão aplicados aos principais parâmetros relacionados aos eventos extremos são úteis para prever a ocorrência destes eventos. Dentre as técnicas de previsão mais utilizadas estão as Redes Neurais Artificiais, os modelos Auto-regressivos de Médias Móveis (ARIMA) e o *Support Vector Machine* (SVM), entre outros.

Neste trabalho, o foco foi realizar a previsão da temperatura de superfície do mar, uma vez que este parâmetro tem sido citado na literatura como um dos mais influentes para a ocorrência de eventos extremos. Os dados utilizados foram obtidos através do projeto PIRATA e correspondem a uma série temporal com dados relativos ao período de 2005 a 2012. Uma

comparação entre os resultados de previsão dos modelos ARIMA e RNAs e os dados reais de TSM foi realizada.

A atenção durante os estudos relacionados às previsões é comumente voltada para a técnica de previsão em si, isto é, suas configurações e modelagem, o que muitas vezes faz com que a preparação para o experimento seja menos explorada. Os procedimentos de pré-processamento dos dados, quando bem executados, podem influenciar positivamente nos resultados das previsões, uma vez que buscam tornar a informação mais clara para que as técnicas de previsão possam fazer uso dela. A contribuição desse trabalho está, principalmente, no pré-processamento dos dados, que ocorre antes da aplicação dos modelos de previsão. Este pré-processamento se deu, entre outas técnicas, com a mudança da periodicidade dos dados diários de temperatura de superfície do mar (TSM) em dados semanais e mensais. Esta transformação possibilitou que fosse possível analisar o desempenho de cada uma das técnicas utilizadas para os dados com diferentes frequências.

No primeiro experimento realizado, foram comparadas as técnicas ARIMA e RNA a partir da análise de suas previsões para horizontes mais curtos. Neste primeiro momento, o principal objetivo era averiguar o desempenho de cada uma das técnicas quando utilizadas séries transformadas. Dessa forma, tanto o ARIMA quanto as RNAs realizaram previsões de curto prazo para dados diários, semanais e mensais. Para os dados semanais e mensais, o ARIMA apresentou um resultado mais acurado, enquanto as RNAs obtiveram um melhor desempenho para os dados diários.

Com os resultados mais favoráveis ao ARIMA, optou-se pelo uso desta técnica para o experimento subsequente. Neste segundo experimento, utilizou-se o modelo ARIMA para realizar previsões com um horizonte de até um ano à frente para cada uma das séries de dados (diárias, semanais e mensais). O objetivo deste experimento, no entanto, mais do que avaliar o ARIMA como ferramenta previsão, era avaliar a influência das séries transformadas para que esta técnica obtivesse bons resultados.

Os resultados obtidos nos experimentos computacionais com os dados de TSM evidenciaram que ao utilizar dados transformados semanais e mensais, o modelo ARIMA obteve previsões mais acuradas do que quando foi utilizada a própria série diária de TSM. Destacando-se os resultados das previsões obtidas utilizando os dados mensais, que além de terem sido mais satisfatórios em praticamente todos os experimentos realizados, apresentaram seus melhores resultados com horizontes mais extensos de previsão, que neste estudo corresponde a um ano à frente.

Observou-se também que a transformação dos dados realizada posteriormente à previsão não proporcionou o mesmo efeito que a transformação realizada como uma etapa de preparação para o experimento, levando à conclusão que a ordem em que os dados são tratados é um fator muito importante para o sucesso da previsão.

Este estudo procura contribuir não apenas ao comparar duas técnicas de previsão (ARIMA versus RNAs), mas também ao avaliar o impacto da utilização de dados cujas frequências sejam diferentes em cada uma das técnicas. Sendo uma proposta para estudos futuros a realização de uma análise mais aprofundada sobre cada um dos modelos de previsão ao utilizarem dados transformados, com configurações diferentes das adotadas nesta dissertação e para maiores horizontes de previsão, além de explorar mais aprofundadamente as possíveis aplicações de técnicas de pré-processamento de dados, como por exemplo, a utilização da diferenciação para o tratamento de séries não estacionárias e o tratamento da sazonalidade das séries.

Além disso, observou-se que o uso dos dados de TSM provenientes apenas uma boia do projeto PIRATA mostrou-se um ponto que poderia ser melhorado, de forma a aproveitar melhor os dados disponibilizados por esta rede de observação oceânica. Portanto, uma proposta para trabalhos futuros seria a utilização de várias boias, caracterizando assim séries espaço-temporais ao invés de séries temporais como fora realizado neste estudo.

Portanto, é possível concluir que ainda há muito ser explorado e a ser mais bem aprofundado nos estudos e previsões relacionados aos eventos extremos, a fim de que seja possível deter cada vez mais conhecimento sobre a natureza e, assim, também conseguir intervir com maior antecedência e precisão, contribuindo para a redução dos impactos causados pelos eventos extremos na sociedade.

Referências Bibliográficas

AGUILAR-MARTINEZ, S.; HSIEH, W. Forecasts of Tropical Pacific Sea Surface Temperatures by Neural Networks and Support Vector Regression. **International Journal of Oceanography**, v. 2009, p. 1–13, 2009.

ALVES, J. M. B.; DE SOUZA, R. O.; CAMPOS, J. N. B. Previsão da anomalia de temperatura da superfície do mar (TSM) no atlântico tropical, com a equação da difusão de temperatura. **Revista Climanálise**, v. 3, p. 163–172, 2006.

BARUA, S.; NG, A.; PERERA, B. Artificial Neural Network–Based Drought Forecasting Using a Nonlinear Aggregated Drought Index. **Journal of Hydrologic Engineering**, v. 17, n. 12, p. 1408–1413, 2012.

BENISTON, M. et al. Future extreme events in European climate: an exploration of regional climate model projections. **Climatic Change**, v. 81, n. 1, p. 71–95, 2007.

BLAIR, T. A. **Meteorologia**. Ao Livro Técnico, 1964.

BOURLÈS, B. et al. The Pirata Program: History, Accomplishments, and Future Directions. **Bulletin of the American Meteorological Society**, v. 89, n. 8, p. 1111–1125, 2008.

BOX, G. E. .; JENKINS, G. M.; REINSEL, G. C. **Time series analysis: forecasting and control**. 4 ed. Wiley, 2008.

BROCKWELL, P. J.; DAVIS, R. A. **Introduction to Time Series and Forecasting**. 2 ed. New York: Springer, 2010.

CHATTOPADHYAY, S. Feed forward Artificial Neural Network model to predict the average summer-monsoon rainfall in India. **Acta Geophysica**, v. 55, n. 3, p. 369–382, 2007.

CHO, J. et al. A study on the relationship between Atlantic sea surface temperature and Amazonian greenness. **Ecological Informatics**, Special Issue on Advances of Ecological Remote Sensing Under Global Change. v. 5, n. 5, p. 367–378, 2010.

DHANYA, C. T.; KUMAR, D. N. Data mining for evolution of association rules for droughts and floods in India using climate inputs. **Journal of Geophysical Research: Atmospheres**, v. 114, n. D2, p. D02102, 2009.

DURAND, B. et al. Tropical Atlantic Moisture Flux, Convection over Northeastern Brazil, and Pertinence of the PIRATA Network*. **Journal of Climate**, 2005.

EM-DAT. **The OFDA/CRED International Disaster Data Base. Centre for Research on the Epidemiology of Disasters**. Brussels (Belgium): Université Catholique de Louvain, 2013. Disponível em: <www.emdat.be>. Acesso em: maio 2013.

FAROKHNIA, A.; MORID, S.; BYUN, H.-R. Application of global SST and SLP data for drought forecasting on Tehran plain using data mining and ANFIS techniques. **Theoretical and Applied Climatology**, v. 104, n. 1-2, p. 71–81, 2011.

FIELD, C. B. **Managing the risks of extreme events and disasters to advance climate change adaption: special report of the Intergovernmental Panel on Climate Change**. New York: Cambridge University Press, 2012.

FU, R. et al. How Do Tropical Sea Surface Temperatures Influence the Seasonal Distribution of Precipitation in the Equatorial Amazon? **Journal of Climate**, v. 14, n. 20, p. 4003–4026, 2001.

FUNCEME. **Sistemas atuantes sobre o Nordeste**. Fundação Cearense de Meteorologia e Recursos Hídricos, 2013. Disponível em: <www.funceme.br/>. Acesso em: abril 2013.

GOOS-BRASIL. **Dados do PIRATA**. GOOS-Brasil, 2013. Disponível em: <http://www.goosbrasil.org>.

GRODSKY, S. A. et al. Tropical instability waves at 0°N, 23°W in the Atlantic: A case study using Pilot Research Moored Array in the Tropical Atlantic (PIRATA) mooring data. **Journal of Geophysical Research: Oceans**, v. 110, n. C8, p. C08010, 2005.

GRUBER, N.; KEELING, C. D.; BATES, N. R. Interannual Variability in the North Atlantic Ocean Carbon Sink. **Science**, v. 298, n. 5602, p. 2374–2378, 2002.

GUJARATI, D. N.; PORTER, D. C. **Basic econometrics**. 5 ed. McGraw-Hill New York, 2008.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**, 3 ed. Morgan Kaufmann, 2011.

HASTENRATH, S. Exploring the climate problems of Brazil's Nordeste: a review. **Climatic Change**, v. 112, n. 2, p. 243–251, 2012.

HAYKIN, S. **Neural Networks and Learning Machines**. 3. ed. Prentice Hall, 2008.

HO, S. L.; XIE, M.; GOH, T. N. A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction. **Computers & Industrial Engineering**, v. 42, n. 2–4, p. 371–375, 2002.

IPEADATA. **Ipeadata database**. http://www.ipeadata.gov.br, 2013.

LINS, I. D. et al. Prediction of sea surface temperature in the tropical Atlantic by support vector machines. **Computational Statistics & Data Analysis**, v. 61, p. 187–198, 2013.

LIU, W. T.; JUÁREZ, R. I. N. ENSO drought onset prediction in northeast Brazil using NDVI. **International Journal of Remote Sensing**, v. 22, n. 17, p. 3483–3501, 2001.

MAIER, H. R.; DANDY, G. C. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. **Environmental Modelling & Software**, v. 15, n. 1, p. 101–124, 2000.

MIGUENS, A. P. **Manual de Navegação.** Navegação: A Ciência e A Arte, v. 3 - Navegação Eletrônica e em Condições Especiais, DHN, 2000.

MISHRA, A. K.; SINGH, V. P. A review of drought concepts. **Journal of Hydrology**, v. 391, n. 1–2, p. 202–216, 2010.

MISHRA, A. K.; SINGH, V. P. Drought modeling – A review. **Journal of Hydrology**, v. 403, n. 1–2, p. 157–175, 2011.

MORETTIN, P. A.; TOLOI, C. M. DE C. **Modelos para previsão de séries temporais**. Instituto de Matemática Pura e Aplicada, 1981.

MORID, S.; SMAKHTIN, V.; BAGHERZADEH, K. Drought forecasting using artificial neural networks and time series of drought indices. **International Journal of Climatology**, v. 27, n. 15, p. 2103–2111, 2007.

MOURA, A. D.; SHUKLA, J. On the Dynamics of Droughts in Northeast Brazil: Observations, Theory and Numerical Experiments with a General Circulation Model. **Journal of the Atmospheric Sciences**, v. 38, n. 12, p. 2653–2675, 1981.

NOAA. **NOAA's El Niño Page**. National Oceanic and Atmospheric Administration, 2013a. Disponível em: <http://www.pmel.noaa.gov/tao/>. Acesso em: mar. 2013.

NOAA. **Dados do PIRATA**. National Oceanic and Atmospheric Administration, 2013b. Disponível em: <http://www.pmel.noaa.gov/tao/>. Acesso em: fev. 2013.

OGASAWARA, E. et al. **Neural networks cartridges for data mining on time series**International Joint Conference on Neural Networks. IJCNN 2009. Atlanta, GA, USA: IEEE Computer Society, 2009

PARISI, F.; SLIVA, A.; SUBRAHMANIAN, V. S. A temporal database forecasting algebra. **International Journal of Approximate Reasoning**, v. 54, n. 7, p. 827–860, 2013.

PREVIEW GLOBAL RISK DATA PLATAFORM. **Global Assessment Reports on Disaster Risk Reduction.** United Nations Strategy for Disaster Reduction (UNISDR), 2013. Disponível em: <http://preview.grid.unep.ch/>. Acesso em: out. 2013.

SERVAIN, J. et al. A Pilot Research Moored Array in the Tropical Atlantic (PIRATA). **Bulletin of the American Meteorological Society**, v. 79, n. 10, p. 2019–2031, 1998.

SERVAIN, J.; CLAUZET, G.; WAINER, I. C. Modes of tropical Atlantic climate variability observed by PIRATA. **Geophysical Research Letters**, v. 30, n. 5, 2003.

SLINGO, J. **Statistical models and the global temperature record**. Met Office, 28 maio 2013. Disponível em: <http://www.metoffice.gov.uk/research/news/statistical-models-and-temperature>. Acesso em: 9 set. 2013.

SUKOV, A. I. et al. A Sequential Analysis Method for the Prediction of Tropical Hurricanes. **International Journal of Remote Sensing**, v. 29, n. 9, p. 2787–2798, 2008.

SVOBODA, M. et al. The Drought Monitor. **Bulletin of the American Meteorological Society**, v. 83, n. 8, p. 1181–1190, 2002.

TADESSE, T. et al. Drought Monitoring Using Data Mining Techniques: A Case Study for Nebraska, USA. **Natural Hazards**, v. 33, n. 1, p. 137–159, 2004.

TSAY, R. S. **Analysis of Financial Time Series**. 1 ed. Wiley-Interscience, 2001.

VIANNA, M. L. et al. Interactions between Hurricane Catarina (2004) and warm core rings in the South Atlantic Ocean. **Journal of Geophysical Research: Oceans**, v. 115, n. C7, 2010.

WU, A.; HSIEH, W. W.; TANG, B. Neural Network forecasts of the tropical Pacific sea surface temperatures. **NEURAL NETWORKS**, v. 19, n. 2, p. 145–154, 2006.

YOON, J.-H.; ZENG, N. An Atlantic influence on Amazon rainfall. **Climate Dynamics**, v. 34, n. 2-3, p. 249–264, 2010.

ZHANG, G. P. Neural networks for classification: a survey. **IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews**, v. 30, n. 4, p. 451–462, 2000.

ZHANG, G. P. Time series forecasting using a hybrid ARIMA and neural network model. **Neurocomputing**, v. 50, p. 159–175, 2003.

ZHANG, G.; PATUWO, B. E.; HU, M. Y. Forecasting with artificial neural networks: The state of the art. **International Journal of Forecasting**, v. 14, n. 1, p. 35–62, 1998.

Apêndice I - Mecanismo de treinamento por *backpropagation*

Considerando os conceitos básicos sobre as redes neurais anteriormente apresentados, torna-se mais fácil conseguir entender o processo que envolve o seu funcionamento. O primeiro passo é entender como funciona uma unidade de processamento da rede, isto é, cada um dos neurônios de uma RNA. A Figura Apêndice I.1 mostra o funcionamento de um neurônio.

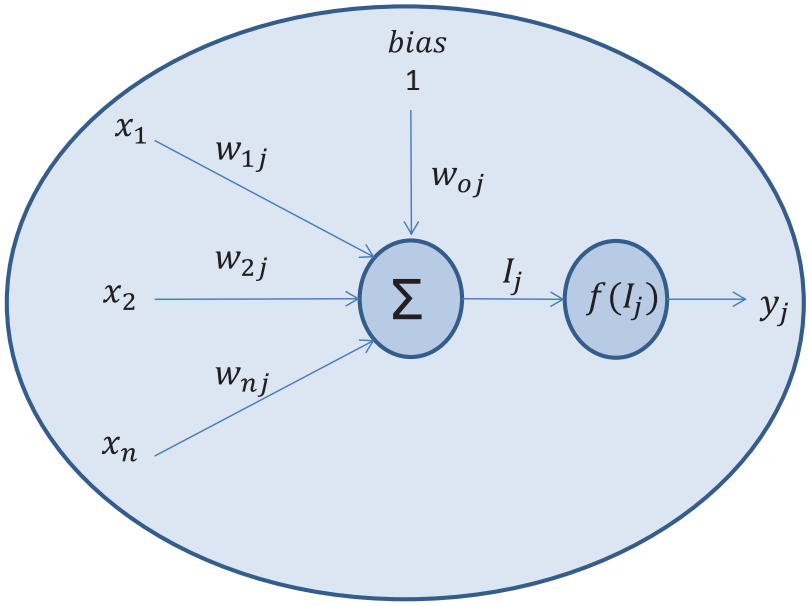


Figura Apêndice I.1: Funcionamento de um neurônio j

Como fora abordado previamente, todas as ligações ou arestas de uma rede neural possuem um peso associado. Dessa forma, cada entrada recebida por um determinado neurônio tem um peso associado, que corresponde à ligação pela qual chegou ao neurônio. Além dos pesos, é somado um componente chamado de *bias*, que tem sempre valor unitário de entrada. Após receber todos os sinais de entrada, com seus respectivos pesos, é realizado um somatório, a fim de contabilizar o total de sinais de entrada para aquele determinado neurônio, compilando todos os sinais em uma só entrada.

$$I_j = \sum w_{ij} O_i + b_j \qquad (18)$$

Onde I_j é a entrada do neurônio j; w_{ij} é o peso da ligação entre os neurônios *i* e *j*; O_i é a saída do neurônio *i*; e b_j é o *bias* ou limiar do neurônio *j*.

O resultado desse somatório é utilizado pela função de transferência do neurônio, que por sua vez gerará uma saída para os demais neurônios. No caso da camada de entrada,

$$X_i = O_i \quad (19)$$

Para as demais camadas,

$$f(I_j) = (1 + \exp(-I_j))^{-1} \quad (20)$$

$$O_j = f(I_j) \quad (21)$$

Onde $f(I_j)$é a função de transferência do neurônio *j*, neste caso, é uma função sigmoidal; I_j é a entrada do neurônio j; e O_j é a saída do neurônio *j.*

Após compreender a dinâmica de em cada unidade de processamento individualmente, é preciso levar em conta que esta dinâmica vai se repetindo ao longo da rede, em um processo conhecido como *feed forward*, isto é, um fluxo de propagação das entradas em direção à saída da rede utilizada. No caso da rede neural do tipo *perceptron* de múltiplas camadas, que utilize o algoritmo de aprendizagem supervisionada *backpropagation*, depois que esse fluxo chega à saída da rede, calcula-se o erro em relação saída esperada e esse erro é retropropagado ao longo da rede e, dessa forma, o fluxo se inverte, em um processo chamado de *feed backward*. Tanto o processo de *feed forward* quanto o *feed backward* serão mais bem detalhados a seguir. A Figura Apêndice I.2 exemplifica uma rede cujos processamentos ocorrem a partir da camada de entrada até a camada de saída.

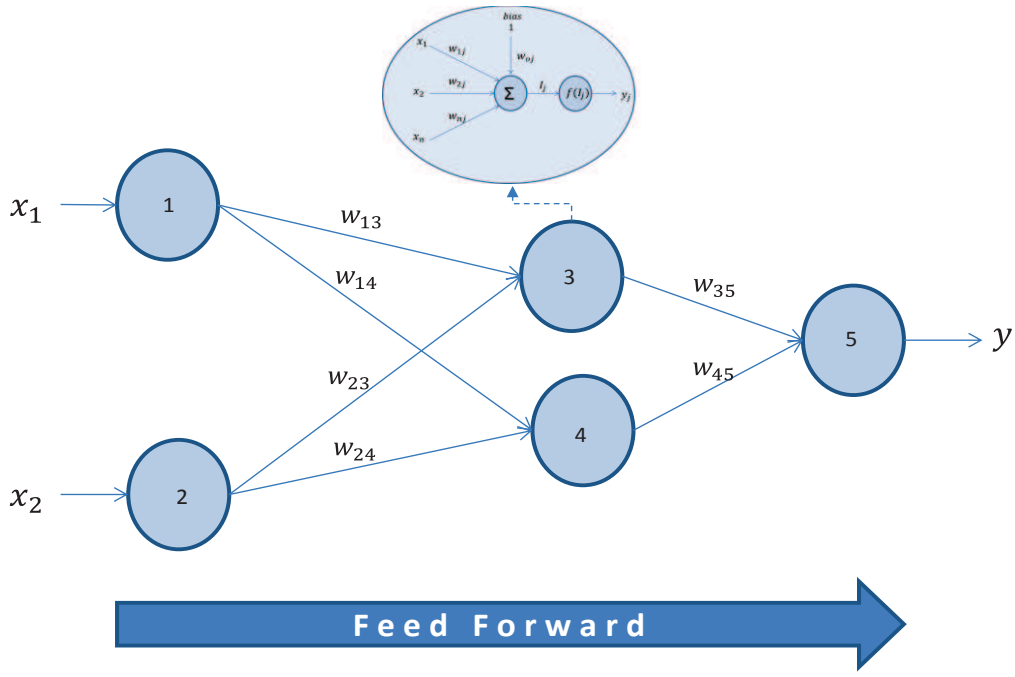


Figura Apêndice I.2: Exemplo de uma rede com fluxo *feed forward*

Em uma rede *perceptron* multicamadas, os dados de entrada do modelo são recebidos pela camada de entrada, que transmite esses dados para a camada posterior, a camada escondida. Na camada escondida, cada nó recebe esses dados e os processa, gerando uma saída para a próxima camada, a camada de saída. A partir da camada de saída, então, a resposta da rede, ou seja, a previsão realizada pelo modelo é obtida. No entanto, durante a fase de treinamento da rede, o que se deseja é que a resposta obtida por ela seja o mais próximo possível da realidade. Dessa forma, quando uma RNA usa o tipo de aprendizagem supervisionada, mais especificamente o algoritmo de aprendizagem *backpropagation*, é dada a rede uma resposta esperada, para que o resultado efetivamente obtido como resposta seja comparado ao resultado que é esperado como resposta e, assim, seja possível calcular o erro entre o que foi obtido e o que era esperado. No caso da camada de saída o erro pode ser calculado como sendo:

$$E = y - z \quad (22)$$

Onde E é o erro da rede; y é a saída obtida pela rede; e z é a saída que é esperada pela rede.

Tendo calculado o erro da rede, o mesmo é retropropagado ao longo das camadas, partindo da camada de saída com sentido à camada inicial. Sendo assim, nas camadas intermediárias, o erro do neurônio é calculado levando também em consideração o erro dos neurônios das camadas anteriores.

$$e_i = \sum w_{ij} e_j \quad (23)$$

Onde e_i é o erro do neurônio *i*; e_j é o erro do neurônio *j*; e w_{ij} é o peso da ligação do neurônio *i* com o neurônio *j*.

A retropropagação dos erros tem o objetivo de tentar ajustar os pesos anteriormente usados, de forma que, com uma nova configuração aperfeiçoada, se consiga obter um resultado cada vez mais próximo do desejado e, assim, o erro seja minimizado o máximo possível. O ajuste dos pesos é calculado pelas seguintes fórmulas:

$$\Delta w_{ij} = \eta e_j O_i \quad (24)$$

$$w_{ij} = w_{ij} + \Delta w_{ij} \quad (25)$$

Onde Δw_{ij} é o ajuste dos pesos; η é a taxa de aprendizado; e_j é o erro do neurônio *j*; O_i é a saída do neurônio i; e w_{ij} é o peso da ligação do neurônio *i* com o neurônio *j*,

Esse fluxo de retropropagação dos erros para o ajuste dos pesos é realizado no sentido oposto ao fluxo inicial da rede, ou seja, os erros partem da camada de saída no sentido da camada inicial, a de entrada. Por isso, esse fluxo é conhecido como *feed backward*, que em uma tradução literal seria retroalimentação. A Figura Apêndice I.3 ilustra o fluxo de propagação dos erros, contrário ao sentido inicial da rede.

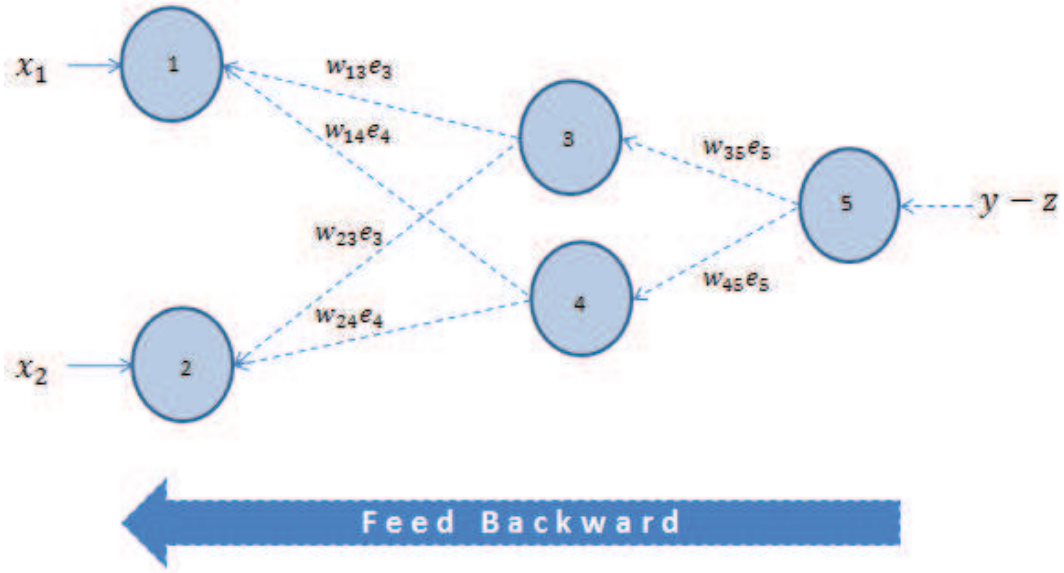


Figura Apêndice I.3: Exemplo de uma rede com fluxo *feed backward*

Portanto, a partir do conhecimento adquirido sobre o funcionamento de uma rede é possível criar modelos adaptados para necessidade de cada problema específico. No entanto, é preciso ressaltar que, apesar da representação de seu funcionamento permitir um melhor entendimento sobre as redes neurais artificiais, a maior desvantagem desse método é justamente o fato dessa representação não ser fiel à realidade do que ocorre dentro do modelo, pois, segundo apontam Han *et al.* (2011), as redes neurais são como uma caixa preta e, sendo assim, interpretar a maneira como o conhecimento é adquirido pelas RNAs e representá-la é complicado para o entendimento humano.