

**CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA
CELSO SUCKOW DA FONSECA**

**Avaliação de Agregação Temporal na Previsão da
Temperatura de Superfície do Mar do Oceano
Atlântico**

Rebecca Pontes Salles

Prof. Orientador:
Eduardo Soares Ogasawara, D.Sc.

**Rio de Janeiro,
Julho de 2016**

**CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA
CELSO SUCKOW DA FONSECA**

**Avaliação de Agregação Temporal na Previsão da
Temperatura de Superfície do Mar do Oceano
Atlântico**

Rebecca Pontes Salles

Projeto final apresentado em cumprimento às
normas do Departamento de Educação
Superior do Centro Federal de Educação
Tecnológica Celso Suckow da Fonseca,
CEFET/RJ, como parte dos requisitos para
obtenção do título de Bacharel em Ciência da
Computação.

Prof. Orientador:
Eduardo Soares Ogasawara, D.Sc.

**Rio de Janeiro,
Julho de 2016**

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

S168 Salles, Rebecca Pontes
Avaliação de agregação temporal na previsão da temperatura
de superfície do mar do Oceano Atlântico / Rebecca Pontes
Salles.—2016.
xiii, 58f. : il. (algumas color.) , grafs. , tabs. ; enc.

Projeto Final (Graduação) Centro Federal de Educação
Tecnológica Celso Suckow da Fonseca , 2016.
Bibliografia : f. 51-58
Orientador : Eduardo Soares Ogasawara

1. Computação. 2. Precipitação (Meteorologia) – Previsão. 3.
Análise de séries temporais. 4. Temperatura atmosférica. 5.
Atlântico, Oceano, I. Ogasawara, Eduardo Soares (Orient.). II.
Título.

CDD 004

DEDICATÓRIA

A Deus e à minha família amada que me ajudaram, apoiaram e guiaram ao longo de toda a minha vida.

AGRADECIMENTOS

Agradece-se ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento parcial desta pesquisa.

Agradece-se também as contribuições de Patricia Mattos que deu início a pesquisas sobre o tema abordado.

RESUMO

Eventos ambientais extremos como secas afetam milhões de pessoas em todo o mundo. Apesar de não ser possível evitar este tipo de evento, a sua previsão sob diferentes horizontes permite a mitigação de eventuais danos causados por sua ocorrência. Uma variável importante para a identificação de ocorrências de secas é a Temperatura da Superfície do Mar (TSM). No Oceano Atlântico Tropical, dados de TSM são coletados e fornecidos pelo projeto *Prediction and Research Moored Array in the Tropical Atlantic* (PIRATA), que é uma rede de observação composta por boias com sensores dispostas nesta região. Sensores deste tipo, e de forma mais geral sensores relacionados a Internet das Coisas (IoT), comumente levam a perda de dados que influencia a qualidade dos conjuntos de dados coletados para o ajustamento de modelos de previsão. Neste contexto, esta pesquisa explora a influência da agregação temporal na previsão de horizontes à frente de TSM considerando diferentes horizontes de previsão e diferentes tamanhos dos conjuntos de dados de treino. Vários experimentos foram conduzidos usando conjuntos de dados fornecidos pelo projeto PIRATA e os resultados obtidos apontam cenários para conjuntos de treino e horizontes de previsão indicando se séries temporais agregadas de TSM podem ser vantajosas para a previsão.

Palavras-chave: modelos de previsão; séries temporais; agregação temporal; temperatura da superfície do mar; Oceano Atlântico

ABSTRACT

Extreme environmental events such as droughts affect millions of people all around the world. Although it is not possible to prevent this type of event, its prediction under different time horizons enables the mitigation of eventual damages caused by its occurrence. An important variable for identifying occurrences of droughts is the Sea Surface Temperature (SST). In the Tropical Atlantic Ocean, SST data are collected and provided by the Prediction and Research Moored Array in the Tropical Atlantic (PIRATA) Project, which is an observation network composed of sensor buoys arranged in this region. Sensors of this type, and more generally Internet of Things (IoT) sensors, commonly lead to data losses that influence the quality of data sets collected for adjusting prediction models. In this paper, we explore the influence of temporal aggregation in predicting step-ahead SST considering different prediction horizons and different sizes for training data sets. We have conducted several experiments using data collected by PIRATA Project. Our results point out scenarios for training data sets and prediction horizons indicating whether or not temporal aggregated SST time series may be beneficial for prediction.

Keywords: prediction models; time series; temporal aggregation; sea surface temperature; Atlantic Ocean

SUMÁRIO

1	Introdução	1
2	As Secas e a Temperatura da Superfície do Mar	8
2.1	As Secas no Brasil	8
2.2	Temperatura de Superfície do Mar	10
2.3	Projeto PIRATA	11
3	Séries Temporais e a Agregação Temporal	13
3.1	Conceitos Gerais	13
3.2	Modelos Lineares para Séries Temporais	14
3.3	Modelos ARIMA	14
3.4	Modelo <i>Random Walk</i>	16
3.5	Encontrando Modelos Lineares Adequados	16
3.6	Agregação Temporal de Séries Temporais	17
3.7	Previsão com Agregação Temporal	18
3.8	Medidas de Erro	19
4	Trabalhos Relacionados	21
5	Metodologia de Previsão de TSM	24
6	Avaliação Experimental	28
6.1	Conjunto de Dados	28
6.2	Configurações Experimentais	29
6.2.1	Seleção de Dados	29
6.2.2	Análise de Dados	33
6.2.3	Particionamento de Dados	35
6.2.4	Geração de Erros de Previsão	36
6.3	Discussão de Resultados	37
7	Conclusões	47
	Referências Bibliográficas	51

LISTA DE FIGURAS

FIGURA 1:	Regiões que foram afetadas pelas secas no Brasil e em outros países da América do Sul durante o período de 1980 a 2001. Fonte: Adaptado de GLOBAL RISK DATA PLATAFORM [2016]	9
FIGURA 2:	Localização das boias do projeto PIRATA (Adaptado de GOOS-Brasil [2015])	12
FIGURA 3:	Diagrama para a agregação temporal de uma série temporal hipotética correspondendo a um ano de observações diárias	18
FIGURA 4:	Localização e exemplos da identificação das boias do projeto PIRATA (Adaptado de GOOS-Brasil [2015])	28
FIGURA 5:	Exemplo de dados de TSM coletados por boias do projeto PIRATA contendo observações faltantes	30
FIGURA 6:	Gráficos gerados com o pacote desenvolvido <i>TSPred</i> retratando previsões de modelos ARIMA e seus intervalos de confiança junto com os dados reais contidos nos respectivos conjuntos de dados de teste. As previsões são representadas pelas linhas sólidas e os dados reais são representados pelas linhas pontilhadas.	31
FIGURA 7:	Exemplo de particionamento de dados para as séries temporais da boia <i>On0e</i>	36
FIGURA 8:	Visão geral da produção de erros de previsão de agregações mensais até um ano à frente	37
FIGURA 9:	Visão geral do processo experimental aplicado para a geração dos dados de erro de previsão analisados	38
FIGURA 10:	Número relativo de vezes em que cada abordagem de previsão aplicada teve uma melhor performance estatisticamente significativa quando comparada à base de referência <i>Random Walk</i>	40
FIGURA 11:	Gráfico das vitórias de cada abordagem de previsão com relação às suas performances na geração de até doze previsões agregadas semanais	41

FIGURA 12: Gráfico das vitórias de cada abordagem de previsão com relação às suas performances na geração de até doze previsões agregadas mensais 44

LISTA DE TABELAS

TABELA 1:	Parâmetros e notações para a metodologia e avaliação experimental	25
TABELA 2:	Dados ininterruptos de TSM selecionados das boias PIRATA	32
TABELA 3:	Resultados e análise dos testes estatísticos	34
TABELA 4:	Definição dos horizontes de previsão considerados na avaliação experimental com relação à geração de previsões de agregações semanais e mensais	39
TABELA 5:	Resultados de comparação estatística (<i>p</i> -valores) entre <i>PrePred</i> com <i>Random Walk</i> e <i>PostPredW</i> com ARIMA, com relação às suas performances na geração de previsões agregadas semanais	41
TABELA 6:	Resultados de comparação estatística (<i>p</i> -valores) entre <i>PrePred</i> com <i>Random Walk</i> e <i>PrePred</i> , <i>PostPredW</i> e <i>PostPredM</i> com ARIMA, com relação às suas performances na geração de previsões agregadas mensais	42
TABELA 7:	Resultados de comparação estatística (<i>p</i> -valores) entre <i>PrePred</i> e <i>PostPredW</i> , com relação às suas performances na geração de previsões agregadas semanais	43
TABELA 8:	Resultados de comparação estatística (<i>p</i> -valores) entre <i>PrePred</i> , <i>PostPredW</i> e <i>PostPredM</i> , com relação às suas performances na geração de previsões agregadas mensais	45

LISTA DE ALGORITMOS

PrePred	Previsão de agregações temporais usando dados diários básicos	24
PostPred	Previsão de agregações temporais usando dados agregados	26
PredExp	Metodologia experimental	26
TSMPart	Particionamento de uma série temporal de TSM em conjuntos de treino e teste	35

LISTA DE ABREVIACES

ADF	<i>Augmented Dickey-Fuller</i>	33, 34
AICC	Critrio De Informao De Akaike Corrigido	17
AR	Modelos Autorregressivos	15
ARIMA	Modelos Autorregressivos Integrados De Mdias Mveis	5, 6, 14, 16, 24, 25, 26,
	27, 29, 31, 34, 38, 39, 42, 47, 48, 49	
ARMA	Modelos Autorregressivos De Mdias Mveis	15, 16
ATLAS	<i>Autonomous Temperature Line Acquisition System</i>	28
CFSV2	<i>The NCEP Climate Forecast System Verso 2</i>	21
CNPQ	Conselho Nacional De Desenvolvimento Cientfico E Tecnolgico	v
EM-DAT	<i>Emergency Events Database</i>	1
FUNCEME	Fundao Cearense De Meteorologia E Recursos Hdricos	9
H0	Hiptese Nula	34
HP	Horizonte De Previso	6, 42, 45
HYCOM	<i>HYbrid Coordinate Ocean Model</i>	21
IOT	Internet Das Coisas	vi, vii, 4, 47
KPSS	<i>Kwiatkowski-Phillips-Schmidt-Shin</i>	17, 33, 34
MA	Modelos De Mdias Mveis	15
MAPE	Erro Percentual Absoluto Mdio	19, 36
MSE	Erro Mdio Quadrtico	19, 36, 38
NCEP	<i>National Centers for Environmental Prediction</i>	21
NMSE	Erro Mdio Quadrtico Normalizado	19, 36
OSEN	Oscilao Sul Do El Nio	2, 8
PIRATA	<i>Prediction and Research Moored Array in the Tropical Atlantic</i>	vi, vii, 5, 6, 11, 12,
	28, 29, 47	
SMAPE	Erro Percentual Absoluto Mdio Simtrico	19, 36
SST	Sea Surface Temperature	vii
SVM	<i>Support Vector Machine</i>	14, 21
TSM	Temperatura Da Superfcie Do Mar	vi, xii, 2, 3, 4, 5, 6, 7, 10, 12, 17, 21, 23, 24,
	28, 29, 31, 33, 34, 35, 37, 38, 44, 47, 48, 49	
ZCIT	Zona De Convergncia Intertropical	2, 8, 9, 10

Capítulo 1

Introdução

Com relativa frequência os fenômenos naturais impactam de forma relevante na vida de bilhões de pessoas. Eventos desta magnitude são chamados de eventos extremos [Teixeira et al., 2015]. Embora ainda não haja um consenso quanto a sua definição, pode-se considerar um evento natural como extremo tendo como base três critérios: raridade, intensidade e severidade [Beniston et al., 2007]. A raridade está relacionada com a frequência em que um evento ocorre, já o critério intensidade refere-se a eventos que possuam magnitude relativamente alta ou baixa, ou seja, que possuam uma grande variação em relação ao valor normal. A severidade considera os impactos socioeconômicos causados por um evento, isto é, os prejuízos causados à sociedade [Beniston et al., 2007; Teixeira, 2014].

Dentre os eventos extremos que causam mais danos a sociedade, destacam-se as secas, os ciclones e as enchentes (ou inundações) [Teixeira, 2014]. Estes eventos causaram juntos um prejuízo estimado em 1,22 trilhões de dólares e afetaram cerca de 4,7 bilhões de pessoas, ocasionando 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 *Emergency Events Database* (EM-DAT) [EM-DAT, 2016].

Entre estes eventos, as secas se destacam por causarem impactos mais severos, já que ocasionaram um número de mortes maior do que os terremotos, ciclones tropicais e inundações causaram juntos, tendo ocasionado cerca de 11,7 milhões de fatalidades entre 1900 a 2013 [EM-DAT, 2016]. A seca pode ser definida como uma condição de umidade insuficiente gerada por uma diminuição da quantidade de precipitação recebida durante um determinado período de tempo [McKee et al., 1993; Mishra and Singh, 2010], que varia de acordo com cada caso, podendo ter duração de vários meses, estações do ano ou mesmo anos.

Além disso, o crescimento da população mundial e o resultante aumento da demanda por água - devido principalmente à expansão dos setores energético, agrícola e industrial - têm contribuído para que a seca se torne um problema ainda mais preocupante [Mishra and Singh, 2010], pois a escassez de água já é uma realidade para muitas regiões do mundo e afeta milhões de pessoas todos os anos [Teixeira, 2014].

Um fato importante que deve ser levado em conta sobre as secas é a dificuldade de avaliar e prever tais eventos, pois segundo Svoboda et al. [2002], a previsão e avaliação de secas são eventualmente mais difíceis de se realizar do que os demais desastres naturais [Teixeira, 2014]. Contudo, a despeito da complexidade relativa ao processo de avaliação e previsão de secas, tendo em vista que a duração, extensão e impacto destas são oscilantes de acordo com a região, as estratégias e princípios concernentes ao gerenciamento do risco de tais eventos são similares e não dependem da realidade de cada país [Tadesse et al., 2004].

Atualmente, existe uma considerável quantidade de dados relacionados aos índices e variáveis climáticos e hidrológicos, como também uma grande variedade de escala de tempo e espaço que devem ser levados em consideração na modelagem de previsão de secas [Tadesse et al., 2004], tendo como componente principal a identificação de preditores efetivos para previsões de secas de longo prazo [Teixeira, 2014] e, nesse aspecto, a temperatura de superfície do mar (TSM) tornou-se uma grande aliada.

A TSM nada mais é que a temperatura da água do mar medida próximo a atmosfera, ou seja, medida em pouca profundidade [Teixeira, 2014]. A TSM é um dos fatores mais relevantes relativos ao condicionamento da variabilidade climática e de eventos extremos, incluindo as secas, onde sua importância para a elaboração de modelos de previsão é ressaltada por Mishra and Singh [2011]. Há evidências na literatura que apontam para o relacionamento entre a TSM do oceano Atlântico sudoeste tropical e a ocorrência de eventos climáticos extremos no nordeste do Brasil [Cho et al., 2010; Fu et al., 2001; Lins et al., 2013; Moura and Shukla, 1981; Ronchail et al., 2002; Sukov et al., 2008]. De maneira geral, a variabilidade da TSM do Oceano Atlântico Tropical tem uma forte influência sobre a distribuição de precipitações na América do Sul Tropical [Nobre and Shukla, 1996], incluindo mais especificamente o nordeste do Brasil [Hastenrath, 1984] e a porção sul-oeste da região Amazônica [Ronchail et al., 2002; Yoon and Zeng, 2009], uma vez que esta variabilidade modula os padrões de anomalias de precipitação, resultando nas secas nesta região [Lins et al., 2013].

De fato, as causas para as secas severas na região nordeste foram estudadas por um longo tempo e atualmente entende-se que os principais fatores que contribuem para a ocorrência deste desastre natural estão relacionados a TSM, incluindo a Oscilação Sul do El Niño (OSEN) e a Zona de Convergência Intertropical (ZCIT) [Durand et al., 2005; Hastenrath, 2011; Liu and Juárez, 2001; Moura and Shukla, 1981].

No Brasil, registros de secas e seus impactos socioeconômicos são datados desde o início da

colonização Portuguesa, mas foi no século XVII que os governos começaram a tomar iniciativas para mitigar seus efeitos [Hastenrath, 2011]. É estimado que 32.8 milhões de pessoas foram afetadas pela seca nos últimos trinta anos, o que levou a uma perda de aproximadamente R\$2.4 bilhões para o país [EM-DAT, 2016].

Há, também, fortes indicações de que, além das secas da região Nordeste do Brasil, a TSM tenha um papel de suma importância em outros fenômenos ocorridos no Oceano Atlântico Tropical [Teixeira, 2014], entre eles destacam-se: (i) o processo que dá origem aos ciclones tropicais no Oceano Atlântico [Sukov et al., 2008]; (ii) as precipitações pluviais na região Amazônica [Fu et al., 2001]; (iii) o volume vegetativo da Amazônia [Cho et al., 2010] e (iv) o sequestro de carbono¹ no oceano [Gruber et al., 2002].

Ainda neste contexto, exemplifica-se a importância da TSM ao se destacar outro evento extremo: o furacão Catarina [Sukov et al., 2008], no sul do Brasil em 2004, que chamou a atenção de pesquisadores sobre o papel desempenhado pela variação da TSM no Atlântico Sul Tropical. Catarina, que afetou 150 mil pessoas, foi o primeiro furacão já registrado no Atlântico Sul e ocorreu em condições muito diferentes daquelas normalmente observadas.

Esta pesquisa teve como motivação a relevância da TSM para o sistema oceano-atmosférico, assim como a dificuldade concernente ao processo de previsão de eventos extremos [Teixeira, 2014]. Assim, esta pesquisa tem por objetivo contribuir para o monitoramento das séries temporais da TSM do Atlântico Sul e a construção de um modelo de maior precisão de sua evolução (o que inclui horizontes de curto e longo prazo) com o intuito de colaborar para se obter uma redução, na medida do possível, da vulnerabilidade da sociedade diante dos eventos extremos, como secas ou enchentes no nordeste brasileiro e na bacia Amazônica [Ward and Folland, 1991], e seus lamentáveis efeitos [Teixeira et al., 2015]. Embora haja muitos trabalhos e modelos que focam o estudo e previsão de curto prazo da TSM [Aguilar-Martinez and Hsieh, 2009; Hertig and Jacobeit, 2010; Lins et al., 2013; Wu et al., 2006], o objetivo desta pesquisa é analisar as vantagens de diferentes técnicas para a previsão de TSM com (i) diferentes horizontes de previsão e (ii) diferentes tamanhos do conjunto de dados de treinamento.

Dada a grande diversidade de eventos extremos e sua natureza comumente instável, a análise da primeira variável (i) muitas vezes é um recurso muito importante para os estudos que incidem sobre a previsão de tais eventos, uma vez que o horizonte de previsão desejado depende

¹O sequestro de carbono é a absorção de grandes quantidades de gás carbônico (CO_2) encontradas na atmosfera. Processo pelo qual organismos capturam o carbono e lançam oxigênio na atmosfera através da fotossíntese. Esse processo natural ajuda a diminuir consideravelmente a quantidade de CO_2 na atmosfera.

fortemente das características de cada evento natural, e muitas vezes um pesquisador está interessado em obter informações futuras de curto a longo prazo sobre um evento. A última variável (ii) se torna especialmente importante devido os dados de TSM serem coletados de diferentes fontes, como sensores, por exemplo, que não são a prova de falhas. No caso de falhas do equipamento, um problema que deve ser lidado é a ausência de dados em algumas partes das séries temporais de TSM. Abordar esta falta de dados torna-se importante para permitir o desenvolvimento de modelos de previsão precisos [Salles et al., 2015]. Esta questão é particularmente importante no contexto da Internet das Coisas (IoT), uma rede de informação global dinâmica que consiste em objetos conectados à Internet, tais como identificações por frequências de rádio, sensores, e atuadores, bem como outros instrumentos e aparelhos inteligentes [Perera et al., 2014]. O paradigma da IoT está rapidamente ganhando terreno no cenário das telecomunicações sem fio modernas e é significativamente vulnerável a interferências de comunicação e mau funcionamento de instrumentos/sensores [Atzori et al., 2010; Chen et al., 2015, 2014; Perera et al., 2014; Tsai et al., 2014]. Muitas vezes, é possível aplicar técnicas de imputação [Yozgatligil et al., 2012] na série temporal incompleta a fim de se computar valores plausíveis para as observações faltantes. Entretanto, caso esta não seja uma opção viável, os dados podem ser particionados, de maneira a produzir conjuntos de dados de diferentes tamanhos utilizados para o treinamento de modelos de previsão, que podem ter que ser constantemente reajustados aos dados disponíveis.

Nesta pesquisa, 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 agregação temporal [Tiao, 1972]. Mais especificamente, esta pesquisa estuda a influência de agregações temporais nas previsões de séries temporais de TSM com o objetivo de identificar se a previsão de uma série agregada pode trazer benefícios frente à previsão direta usando-se a série sem nenhuma intervenção, principalmente na avaliação dos possíveis impactos que estas agregações podem causar em horizontes de previsão mais extensos [Teixeira, 2014]. Além disso, estamos interessados em analisar o contexto em que esses benefícios se mostram mais proeminentes dados diferentes horizontes de previsão e tamanhos de conjuntos de dados de treinamento. Apesar dos efeitos da agregação temporal na previsão de séries temporais serem extensivamente abordados na literatura [Rostami-Tabar et al., 2013, 2014; Wei, 1978; Stram and Wei, 1986; Englund et al., 1999; Abraham, 1982; Nelson and Plosser, 1982; Tiao, 1972; Wei, 2005; Silvestrini and Veredas, 2008; Kourentzes

et al., 2014; Petropoulos and Kourentzes, 2014; Kourentzes and Petropoulos, 2015; Athanasopoulos et al., 2015], as pesquisas mostraram que as vantagens e desvantagens da agregação temporal estão fortemente ligadas às propriedades estatísticas das séries temporais básicas (não-agregadas) disponíveis. Propriedades como estacionariedade, autocorrelação e sazonalidade se mostraram muito importantes para a análise da técnica de agregação temporal, entretanto análises estatísticas demonstraram considerável diversidade em tais propriedades observadas nas séries temporais de TSM. Com isso, os conhecimentos prévios quanto aos efeitos da agregação temporal não podem ser diretamente aplicados, e um maior estudo se faz necessário.

Os efeitos das agregações temporais foram avaliados utilizando-se modelos lineares já consolidados na literatura: os Modelos Autorregressivos Integrados de Médias Móveis (ARIMA) [Box et al., 2008] e o modelo *Random Walk*, utilizado como base de referência para se analisar os resultados experimentais obtidos. Ambos ARIMA e *Random Walk* são modelos estatísticos simples e bem-estabelecidos que podem produzir uma família de outros modelos lineares de previsão e quando bem ajustados, são capazes de lidar com não-estacionariedade e sazonalidade em séries temporais. Estes modelos são também são capazes de inspirar maior confiança em cientistas de dados já que são modelos interpretáveis, ao contrário de alguns métodos de aprendizado de máquina que são normalmente considerados "caixas-pretas". Estas características indicam o potencial destes modelos para se executar um estudo em que uma maior análise do comportamento das séries temporais é um fator de interesse.

Esses modelos foram implementados na linguagem e ambiente para computação estatística R [R Development Core Team, 2008]. Particularmente, foram desenvolvidas diversas funções para automatizar o treinamento e previsão de séries temporais com ARIMA. Estas funções selecionam um modelo ARIMA otimizado, enquanto também consideram sazonalidade, e estão publicamente disponíveis como um pacote-R (*TSPred*) [Salles and Ogasawara, 2015] que permite novos estudos e garante a reprodutibilidade aos experimentos. O pacote desenvolvido *TSPred* foi projetado para se explorar não só os modelos ARIMA, mas também diversos outros modelos lineares a fim de prover o modelo linear mais adequado aos dados passados como entrada, a ser usado como modelo referencial (*baseline*) para previsão de séries temporais com métodos de aprendizado de máquina.

Conjuntos de dados contendo séries temporais diárias de TSM obtidas *online* a partir do sítio do projeto *Prediction and Research Moored Array in the Tropical Atlantic* (PIRATA) [GOOS-Brasil, 2015] foram utilizados como entrada para os modelos de previsão e para a análise e

execução dos algoritmos. O projeto PIRATA 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 [Teixeira, 2014]. Os resultados experimentais indicam que a previsão usando séries temporais da TSM agregadas temporalmente leva a previsões mais precisas em horizontes de longo prazo de acordo com várias medidas de precisão. Além disso, os experimentos mostram que os benefícios da aplicação de agregação temporal são influenciados pelo Horizonte de Previsão (HP) e também pelo tamanho do conjunto de dados de treinamento.

Em trabalho anterior [Teixeira, 2014] foi realizada a comparação da previsão de dados de TSM em três frequências distintas (diários, semanais e mensais) utilizando-se técnicas de previsão que também faziam uso dos modelos ARIMA, assim como do método de aprendizado de máquina conhecido como Redes Neurais [Hippert et al., 2001], a fim de se avaliar os impactos da transformação de dados nos horizontes de previsão de TSM. Os resultados obtidos apontaram que a utilização da série em baixa frequência trazia uma melhora de 71% na previsão de longo prazo. Entretanto, através da análise destes resultados observou-se a necessidade de uma avaliação mais aprofundada das condições em que tais transformações de dados, i.e., agregações temporais, trazem benefícios à produção de melhores performances de previsão.

Sendo assim, esta pesquisa estende os estudos iniciados na dissertação de mestrado de Patrícia Mattos Teixeira, formada pelo Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), visando avaliar extensivamente os benefícios da agregação temporal para a previsão de TSM com diferentes horizontes de previsão, diferentes quantidades de dados disponíveis para treinamento de métodos de previsão, e com dados de TSM de diferentes fontes e apresentando propriedades estatísticas diversas. Para isso, nesta pesquisa foram utilizados dados de TSM, provenientes do projeto PIRATA, coletados por diversas boias ao invés de apenas uma, como feito por Teixeira [2014]. Além disso, a avaliação experimental consistiu em diversas análises das propriedades estatísticas das diversas séries temporais de TSM selecionadas, além de vários experimentos de análise comparativa e estatística dos resultados de previsão de TSM com dados agregados temporalmente ou não.

Além desta introdução (Capítulo 1), a pesquisa está organizada em mais cinco capítulos. O Capítulo 2 apresenta informações relevantes sobre os eventos climáticos extremos e a sua relação com a variável de TSM, além de introduzir o projeto PIRATA. O Capítulo 3 apresenta uma base teórica de previsão de séries temporais e de agregação temporal. O Capítulo 4 descreve tra-

balhos relacionados. O Capítulo 5 apresenta a metodologia de previsão de séries temporais de TSM aplicada nesta pesquisa. O Capítulo 6 apresenta as configurações e resultados da avaliação experimental. E finalmente, o Capítulo 7 apresenta as conclusões gerais da pesquisa.

Capítulo 2

As Secas e a Temperatura da Superfície do Mar

As secas ocorrem em diversas regiões do mundo e trazem com elas consequências econômicas e sociais de grande escala para milhões de pessoas [EM-DAT, 2016]. A magnitude desse 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 and Singh, 2010].

2.1 As Secas no Brasil

No Brasil, o registro de secas e suas consequências socioeconômicas existem desde o início da colonização portuguesa, entretanto, somente no século XVII os governos começaram a tomar iniciativas para tentar atenuar seus efeitos [Hastenrath, 2011; Teixeira, 2014]. 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 causado ao país [EM-DAT, 2016]. Historicamente, a região Nordeste do Brasil tem sido a mais afetada [Hastenrath, 2011; Liu and Juárez, 2001], como mostra a região alaranjada na Figura 1, com consequências sociais e econômicas significativas para o país, pois devido a frequência das secas, uma parcela relevante da população daquela região migrou para outras regiões, principalmente a Sudeste, à procura de trabalho e melhores condições de vida [Hastenrath, 2011], reduzindo assim a capacidade econômica da região nordestina, causando impactos também em outras regiões [Teixeira, 2014].

As causas para que a região Nordeste seja tão severamente atingida pelas secas vêm sendo estudadas há bastante tempo, mas só recentemente chegou-se a um consenso a respeito dos principais fatores que contribuem para a ocorrência desse desastre natural naquela região [Teixeira, 2014]. Os fenômenos conhecidos por OSEN e ZCIT foram associados à ocorrência de secas no Nordeste por diversos estudos realizados nessa região [Durand et al., 2005; Hastenrath, 2011; Liu and Juárez, 2001; Moura and Shukla, 1981].

O El Niño é causado por uma corrente oceânica que ocorre ao longo da costa oeste da Amé-

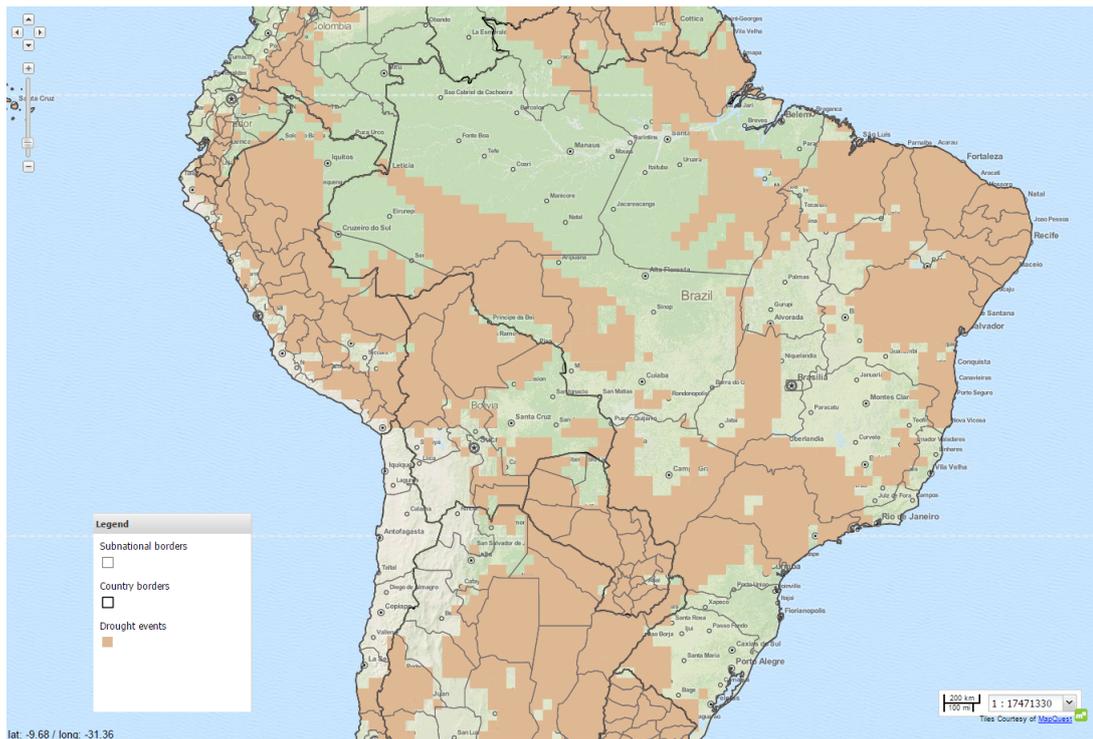


Figura 1: Regiões que foram afetadas pelas secas no Brasil e em outros países da América do Sul durante o período de 1980 a 2001. Fonte: Adaptado de GLOBAL RISK DATA PLATFORM [2016]

rica do Sul, desde o Equador até o Peru [Liu and Juárez, 2001] cuja temperatura eventualmente torna-se excepcionalmente quente [NOAA, 2016]. Alguns estudos procuram mostrar que nos 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 and Juárez [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 ZCIT [Teixeira, 2014].

A ZCIT é o sistema meteorológico mais importante no que concerne à indicação do volume de chuvas no Nordeste [Teixeira, 2014]. A ZCIT, segundo a Fundação Cearense de Meteorologia e Recursos Hídricos (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, 2011; Teixeira, 2014]. Os fenômenos climáticos no Nordeste ocorrem 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, 2011; Teixeira, 2014].

Como a ZCIT ocorre significativamente sobre os oceanos, suas movimentações e sua localização são fortemente afetadas pela temperatura de superfície do mar [Teixeira, 2014], que exerce um papel importante na dinâmica oceano-atmosférica, influenciando, por consequência, na ocorrência ou não de chuvas na região Nordeste [Hastenrath, 2011; Moura and Shukla, 1981; Teixeira, 2014].

2.2 Temperatura de Superfície do Mar

A temperatura de superfície do mar (TSM) é considerada um indicador de acúmulo de calor do oceano e refere-se a temperatura da água do oceano em sua superfície, isto é, os dados de TSM correspondem 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 [Lins et al., 2013; Teixeira, 2014].

A TSM nos oceanos tropicais Pacífico e Atlântico é a principal variável física influenciadora das condições climáticas em várias áreas do globo terrestre [Alves et al., 2003; Teixeira, 2014]. 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 [Teixeira, 2014]. 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 and Zeng, 2009; Teixeira, 2014]. 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 and Singh, 2011; Moura and Shukla, 1981; Vianna et al., 2010; Teixeira, 2014]. Além disso, a TSM tem sido associada também à processos relacionados mais diretamente com as mudanças climáticas, como nos processos de sequestro de carbono (fotossíntese) [Gruber et al., 2002; Teixeira, 2014].

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 [Teixeira, 2014].

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 atenuar seus efeitos [Teixeira, 2014]. Tendo isto em vista, foi desenvolvido um projeto de monitoramento do oceano Atlântico Tropical chamado PIRATA, que tem como um dos principais objetivos descrever e entender a evolução da temperatura de superfície do mar [Grotsky et al., 2005].

2.3 Projeto PIRATA

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 devido aos eventos ocorridos nesta área serem relativamente mais comuns e mais devastadores [Teixeira, 2014]. 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 PIRATA [Bourlès et al., 2008; Teixeira, 2014].

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; Teixeira, 2014].

O PIRATA teve seu início em 1997, a partir de uma parceria entre o Brasil, Estados Unidos e França. No início o programa contava com doze boias e compunham a fase piloto do programa, que iria durar até 2001, quando estas boias foram completamente instaladas, passando então para fase de consolidação [Bourlès et al., 2008; Servain et al., 1998; Teixeira, 2014]. 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 [Teixeira, 2014]. 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 [Teixeira, 2014].

A definição da localização das boias, que pode ser observada pelos círculos amarelos na Figura 2, 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 et al., 2003; Teixeira, 2014].

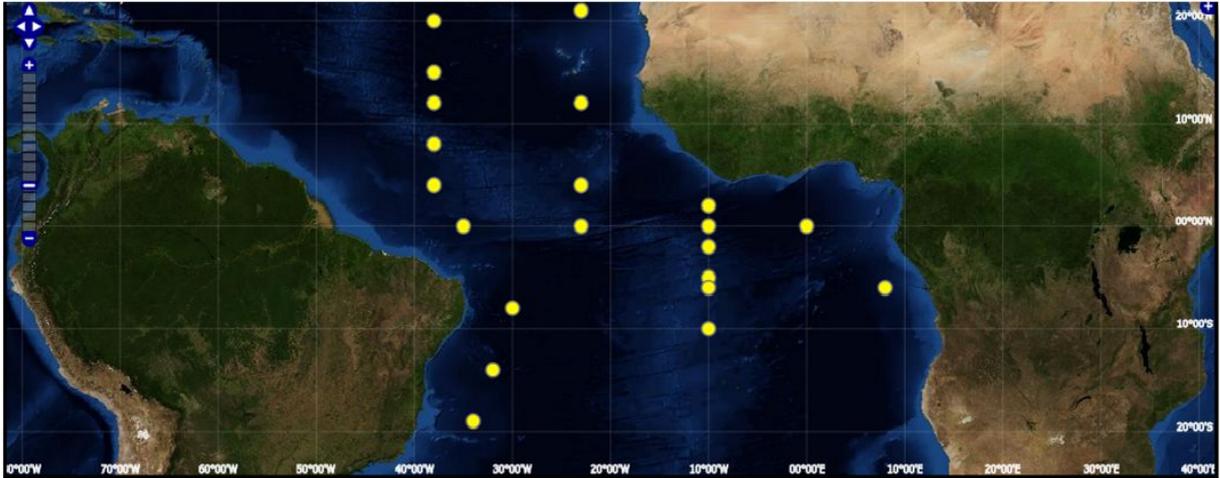


Figura 2: Localização das boias do projeto PIRATA (Adaptado de GOOS-Brasil [2015])

Dessa forma, os dados das diferentes variáveis obtidas e disponibilizados graças ao projeto PIRATA têm sido de grande utilidade para diversas pesquisas [Teixeira, 2014]. O objetivo dessa pesquisa é utilizar dados históricos de TSM disponibilizados pelo projeto PIRATA para realizar novas previsões de TSM com diferentes horizontes de previsão, diferentes tamanhos de conjuntos de dados de treinamento e utilizando diferentes frequências obtidas a partir da transformação dos dados originalmente diários do PIRATA em dados semanais e mensais através da aplicação da técnica de agregação temporal.

Capítulo 3

Séries Temporais e a Agregação Temporal

3.1 Conceitos Gerais

Uma série temporal x_t é uma sequência $\langle x_1, x_2, x_3, \dots, x_n \rangle$ de observações de um fenômeno de interesse coletadas ao longo do tempo, tal que x_1 corresponde ao valor da primeira (mais antiga) observação, enquanto x_n é a última (mais recente) delas. O comprimento n de uma série temporal x_t é representado como $|x_t| = n$.

A maioria dos trabalhos empíricos com séries temporais assume um processo estacionário implícito [Gujarati, 2002]. Uma série temporal estacionária x é um processo estocástico tal que: (i) sua função média \bar{x}_t é constante e não depende do tempo t e (ii) a função de auto covariância entre x_t e uma série temporal deslocada temporalmente x_s depende somente da diferença entre elas $|s - t|$ [Gujarati, 2002; Shumway and Stoffer, 2010]. Se essas restrições não são atendidas para x_t , esta série temporal é classificada como não-estacionária. Sob essas circunstâncias, uma preocupação frequente é a busca de maneiras de se transformar uma série temporal em estacionária.

Muito frequentemente encontra-se séries temporais não-estacionárias que apresentam comportamento estacionário em torno de uma tendência. É possível remover uma tendência (e transformar tal série temporal em estacionária) através de diferenciação. A primeira diferenciação é denotada como $\nabla x_t = x_t - x_{t-1}$ e é capaz de eliminar uma tendência linear. A segunda diferenciação pode eliminar uma tendência quadrática, e assim por diante. Para diferenciações de ordem mais alta aplica-se o operador *backshift* [Shumway and Stoffer, 2010] (Equação 3.1), onde ∇^d é a diferenciação de ordem d e o operador $(1 - B)^d$ pode ser algebricamente expandido para valores inteiros mais altos de d [Shumway and Stoffer, 2010].

$$\nabla^d = (1 - B)^d, B^k x_t = x_{t-k} \quad (3.1)$$

3.2 Modelos Lineares para Séries Temporais

Geralmente pode-se considerar um método para estimar um modelo de séries temporais como sendo *rígido* ou *flexível*. Métodos rígidos, tais como regressões lineares, fazem uma suposição inicial sobre a série temporal analisada e impõem um modelo específico sobre ela. O processo de modelagem, em seguida, pode ser entendido basicamente como um problema de estimativa de um conjunto de coeficientes. No entanto, o modelo adotado pode não se ajustar adequadamente a toda a série temporal [James et al., 2013].

Em contraste, os métodos flexíveis têm o seu foco no ajustamento de um modelo de modo que as observações da série temporal disponível sejam aproximadas, tanto quanto possível, dado um certo grau de suavização [James et al., 2013]. Alguns exemplos de métodos flexíveis são as *Splines* de regressão [James et al., 2013], *Support Vector Machine* (SVM) [Sapankevych and Sankar, 2009] e as Redes Neurais [Hippert et al., 2001]. Quando o objetivo é previsão, embora uma conclusão simples seria a de que métodos flexíveis são mais adequados do que os métodos rígidos, isso não é garantido, devido à possibilidade de ajustamento excessivo dos dados durante o treinamento, o que pode resultar em um desempenho inferior durante a previsão [James et al., 2013]. Assim, métodos rígidos não devem ser negligenciados quando se trata de previsão de séries temporais.

Outra questão importante diz respeito à facilidade de interpretação de um modelo de séries temporais. Gestores de negócios e pesquisadores tendem a se sentir mais seguros quando eles podem analisar e entender os modelos. Os modelos desenvolvidos por métodos rígidos tendem a ser consideravelmente mais fáceis de se interpretar, portanto, eles podem ser preferíveis caso a vantagem de métodos flexíveis não seja estatisticamente significativa ou seja apenas ligeiramente melhor [Salles et al., 2015]. Um dos métodos rígidos mais importantes para a previsão de séries temporais é o ARIMA [Gujarati, 2002].

3.3 Modelos ARIMA

Um objetivo comum para se ajustar/treinar modelos matemáticos para que reflitam uma série temporal x_t é a obtenção de previsões precisas de valores futuros de x_t baseadas em dados coletados previamente [James et al., 2013; Shumway and Stoffer, 2010]. Como citado anteriormente, um dos métodos lineares mais importantes de previsão de séries temporais é a classe sistemática de modelos chamada ARIMA [Box et al., 2008]. Modelos ARIMA(p, d, q) são

derivados de uma composição dos processos de modelagem autorregressivo (Modelos Autorregressivos (AR)) e de média móvel (Modelos de Médias Móveis (MA)) (respectivamente representados por p e q) com um processo de diferenciação preliminar adicional (I) (representado por d), desenvolvidos para lidar com não-estacionariedade [Gujarati, 2002].

Modelos AR são baseados na premissa de que o valor x_i de uma série temporal x_t pode ser descrito como uma função dos seus p valores anteriores, $x_{i-1}, x_{i-2}, \dots, x_{i-p}$. Um modelo autorregressivo de ordem p , $AR(p)$, é definido na Equação 3.2, onde x_t é estacionária, p é o número de passos em direção ao passado precisos para se prever o valor atual, $\phi_1, \phi_2, \dots, \phi_p$ são contantes ($\phi_p \neq 0$), e ω_t é uma série de ruídos brancos Gaussianos com média zero e variância σ_ω^2 [Box et al., 2008].

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + \omega_t \quad (3.2)$$

Em contrapartida, os modelos MA assumem que um valor x_i de uma série temporal pode ser descrito como uma função de seus últimos q valores de ruído branco $\omega_{i-1}, \omega_{i-2}, \dots, \omega_{i-q}$. O modelo de média móvel de ordem q , $MA(q)$, é definido na Equação 3.3, onde q é o número de termos deslocados temporalmente na média móvel da série de ruídos brancos e $\theta_1, \theta_2, \dots, \theta_q$ ($\theta_q \neq 0$) são parâmetros [Shumway and Stoffer, 2010]. O modelo MA também difere do modelo AR já que é estacionário independentemente dos valores assumidos pelos parâmetros $\theta_1, \dots, \theta_q$ [Shumway and Stoffer, 2010].

$$x_t = \omega_t + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \dots + \theta_q \omega_{t-q} \quad (3.3)$$

Ambos os processos AR e MA podem ser combinados resultando nos Modelos Autorregressivos de Médias Móveis (ARMA). O modelo $ARMA(p, q)$ é definido na Equação 3.4, onde $\phi(B)$ e $\theta(B)$ são os operadores AR e MA, definidos nas Equações 3.5 e 3.6, respectivamente [Box et al., 2008].

$$\phi(B)x_t = \theta(B)\omega_t \quad (3.4)$$

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p, \quad (3.5)$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q, \quad (3.6)$$

Em última análise, o modelo ARMA integrado, ou modelo ARIMA, pode ser formulado através da inclusão de um processo prévio de diferenciação, sendo portanto concebido para lidar com séries temporais não-estacionárias. Assim, uma série temporal x_t é definida como um modelo autorregressivos integrado de médias móveis de ordem (p, d, q) , isto é, $\text{ARIMA}(p, d, q)$, tal como apresentado na Equação 3.7, baseada nos operadores definidos nas Equações 3.5 e 3.6 [Shumway and Stoffer, 2010].

$$\phi(B)(1-B)^d x_t = \theta(B)\omega_t. \quad (3.7)$$

3.4 Modelo *Random Walk*

Em muitas situações, pode-se encontrar séries temporais que parecem se comportar de uma forma completamente aleatória ao longo do tempo. Neste caso, um modelo muito útil para se estimar tal série temporal é o *Random Walk* com desvio (*drift*), no qual o valor de uma série temporal x_t em um instante t é basicamente explicado pelo valor da série no instante $t - 1$ mais um movimento completamente aleatório representado por ω_t . Com isso, o modelo é formulado como na Equação 3.8 para $t = 1, 2, \dots$, onde ω_t é o ruído branco e $x_0 = 0$ a condição inicial. Assim, o modelo *Random Walk* pode também ser expressado como um modelo $\text{ARIMA}(0, 1, 0)$, [Shumway and Stoffer, 2010]:

$$x_t = \delta + x_{t-1} + \omega_t. \quad (3.8)$$

3.5 Encontrando Modelos Lineares Adequados

O passo mais geral para se ajustar um modelo ARIMA é a identificação da ordem autorregressiva p , da ordem de diferenciação d e da ordem de médias móveis q [Box et al., 2008]. A escolha dos parâmetros corretos para produzir um modelo ARIMA não é uma tarefa trivial. Com o objetivo de facilitar o processo de ajustamento de um modelo ARIMA, alguns algoritmos foram desenvolvidos [Hyndman and Khandakar, 2008] para se computar os parâmetros p , d e q . O número de diferenciações d é determinado pela repetição de testes *Kwiatkowski-Phillips-*

Schmidt-Shin (KPSS), que determinam se uma série temporal é estacionária. Os valores de p e q são então escolhidos pela minimização do valor do critério Critério de Informação de Akaike Corrigido (AICc) de segunda ordem, que mede a qualidade do ajustamento do modelo, depois da diferenciação do conjunto de dados de treino d vezes [Hyndman and Athanasopoulos, 2013].

3.6 Agregação Temporal de Séries Temporais

Seja x_t uma série temporal cujas observações são coletada em intervalos de tempo equidistantes, por exemplo, a TSM diária do Atlântico. Neste caso, é frequentemente interessante a previsão não das observações diárias subsequentes mas de um próximo período semanal ou mensal de temperaturas médias, que são agregações temporais dos dados diários [Tiao, 1972]. Considerando x_t como uma entrada, uma série temporal agregada não-sobreposta de período m , X_T , é definida na Equação 3.9, onde X_T é a série temporal agregada e m corresponde às agregações de x_t de período m , *i.e.*, a ordem de agregação. Assim, no exemplo anterior, t representa a unidade de tempo de dia e se m é 7 ou 30, então X_T é respectivamente uma série temporal de médias aritméticas semanais ou mensais de x_t [Wei, 2005]. A Figura 3 representa possíveis artefatos do processo de agregação temporal aplicado a uma série temporal hipotética com um ano de observações diárias x_t que deriva séries temporais agregadas semanais (X_T) e mensais (X_S).

$$X_T = \sum_{t=m(T-1)+1}^{mT} \frac{x_t}{m} = \frac{(1 + B + \dots + B^{m-1})x_{mT}}{m} \quad (3.9)$$

Se x_t é uma série temporal de frequência mais alta e X_T é a série agregada temporalmente (de frequência mais baixa), a previsão das próximas observações de X_T pode seguir duas diferentes abordagens: (i) agrega-se as previsões de x_t ou (ii) prevê-se diretamente com os dados de X_T . Na primeira abordagem, quando em posse dos dados de x_t (diários), aplica-se a solução mais simples e direta, *i.e.*, constrói-se um modelo de previsão que é utilizado para se prever as próximas observações diárias e então produz-se os dados agregados desejados (observações semanais ou mensais).

Entretanto, na prática, esta pode não ser sempre a abordagem mais vantajosa [Tiao, 1972]. Com isso, torna-se importante o estudo da eficiência relativa de ambas as abordagens, já que se sabe que a agregação temporal tende a amplificar a variabilidade de baixa frequência em séries temporais [Nelson and Plosser, 1982]. Pesquisas anteriores mostraram que quando a

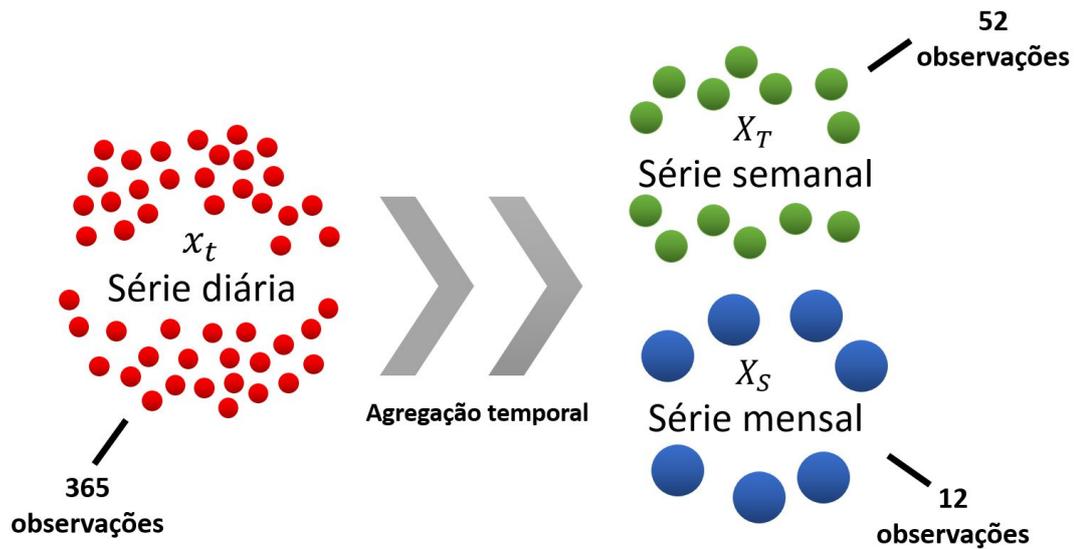


Figura 3: Diagrama para a agregação temporal de uma série temporal hipotética correspondendo a um ano de observações diárias

série básica (de menor unidade de tempo) x_t é estacionária ($d = 0$), a previsão baseada nos dados agregados derivados X_T pode ser eficiente para valores altos de m , enquanto que quando x_t é não-estacionária ($d > 0$), pode ser vantajoso usar dados de mais alta frequência, especialmente para previsões de curto prazo [Tiao, 1972].

3.7 Previsão com Agregação Temporal

Considera-se que duas séries temporais estão disponíveis: (i) a série temporal não-agregada de alta frequência x_t ($x_1, x_2, \dots, x_{mT_0}, x_{mT_0+1}, \dots, x_{m(T_0+N)}$) que possui $m(T_0 + N)$ observações e (ii) a série agregada de baixa frequência correspondente X_T ($X_1, X_2, \dots, X_{T_0}, X_{T_0+1}, \dots, X_{T_0+N}$) com $(T_0 + N)$ observações. As primeiras mT_0 e T_0 observações de (i) e (ii), respectivamente, são usadas para se desenvolver um modelo de previsão, *i.e.*, são usadas como o conjunto de dados de treino. As observações remanescentes de ambas as séries são usadas para avaliação de erros de previsão a fim de se analisar a performance de previsão do modelo, *i.e.*, são usadas como o conjunto de dados de teste [Abraham, 1982].

Existem duas abordagens gerais para a previsão de agregações temporais, *i.e.*, para a previsão de X_{T_0+j} , onde $j = 1, \dots, N$. Na primeira abordagem, constrói-se um modelo usando-se as primeiras mT_0 observações de x_t . O modelo é usado para sucessivamente gerar as previsões $\hat{x}_{m(T_0+j-1)+l}$ para as seguintes observações $x_{m(T_0+j-1)+l}$ ($l = 1, 2, \dots, m; j = 1, 2, \dots, N$). A partir destas previsões, a previsão para X_{T_0+j} é feita por agregação temporal, como é indicado na

Equação 3.10, onde a soma sobre $l = 1, \dots, m$ pode ser gerada como [Abraham, 1982] :

$$\hat{X}_{T_0+j}^{(1)} = \sum_{l=1}^m \frac{\hat{x}_{m(T_0+j-1)+l}}{m} \quad (j = 1, \dots, N). \quad (3.10)$$

Na segunda abordagem, a tarefa principal é também o desenvolvimento de um modelo, entretanto usando as primeiras T_0 observações da série temporal agregada disponível X_T . Este modelo é então usado para sucessivamente produzir as previsões observadas na Equação 3.11 para as agregações X_{T_0+j} ($j = 1, \dots, N$) [Abraham, 1982].

$$\hat{X}_{T_0+j}^{(2)} = E(X_{T_0+j} | X_{T_0+j-1}, X_{T_0+j-2}, \dots) \quad (3.11)$$

Os erros das previsões $\hat{X}_{T_0+j}^{(1)}$ e $\hat{X}_{T_0+j}^{(2)}$ são computados através da sua comparação às N observações agregadas de teste X_{T_0+j} ($j = 1, \dots, N$).

3.8 Medidas de Erro

Uma grande variedade de medidas de acurácia foram criadas e adotadas em vários experimentos práticos [Chen and Yang, 2004]. Embora seja comumente aceito que não existe uma única melhor medida de acurácia de previsão, algumas das mais frequentemente adotadas são a medida de Erro Médio Quadrático (MSE) (Equação 3.12) [Shumway and Stoffer, 2010; James et al., 2013] e sua variação, Erro Médio Quadrático Normalizado (NMSE) (Equação 3.13) [Chen and Yang, 2004], assim como a medida de Erro Percentual Absoluto Médio (MAPE) (Equação 3.14) [Chen and Yang, 2004] e sua variação, a medida de Erro Percentual Absoluto Médio Simétrico (SMAPE) (Equação 3.15) [Chen and Yang, 2004]. Todas estas medidas são computadas tendo como entrada a série temporal original $x_t = \{x_1, \dots, x_n\}$, o número de observações previstas, m , e o preditor de x_{n+i} , com $i = 1, 2, \dots, m$, denotado por \hat{x}_{n+i} .

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^m (\hat{x}_{n+i} - x_{n+i})^2, \quad (3.12)$$

$$\text{NMSE} = \sqrt{\frac{\sum_{i=1}^m (\hat{x}_{n+i} - x_{n+i})^2}{\sum_{i=1}^m (x_{n+i} - \bar{x})^2}}, \quad \text{onde } \bar{x} = \frac{1}{n} \sum_{k=1}^n x_k, \quad (3.13)$$

$$\text{MAPE} = \frac{1}{m} \sum_{i=1}^m \frac{|\hat{x}_{n+i} - x_{n+i}|}{|x_{n+i}|} \quad (3.14)$$

$$\text{SMAPE} = \frac{1}{m} \sum_{i=1}^m \frac{2|\hat{x}_{n+i} - x_{n+i}|}{(|x_{n+i}| + |\hat{x}_{n+i}|)}. \quad (3.15)$$

Capítulo 4

Trabalhos Relacionados

Devido a grande influência da TSM nas condições climáticas de muitas regiões em todo o globo, esta variável ambiental tem sido extensivamente estudada e tem sido objeto de muitas pesquisas principalmente relacionadas a questão de se prever a TSM no oceano Atlântico tropical.

A maioria das pesquisas anteriores são centradas em previsões de TSM a curto prazo baseadas em variáveis meteorológicas e modelos climáticos em geral, como o *HYbrid Coordinate Ocean Model* (HYCOM) e *The NCEP Climate Forecast System Versão 2* (CFSv2) do *National Centers for Environmental Prediction* (NCEP). Algumas pesquisas que se enquadram nesta categoria incluem Wang et al. [2014], Cho et al. [2010], Zheng and Zhu [2010], Barth et al. [2006], Skliris et al. [2011], Stockdale et al. [2011], Muller et al. [2007], Feng et al. [2010], Collins et al. [2004], Hu et al. [2012], Kara et al. [2009], e Liu et al. [2010]. Alguns destes autores também estudam métodos de aprendizado de máquina como uma análise complementar [Collins et al., 2004]. Outros hibridizam seus métodos baseados em sistemas de previsão climática com a adição de modelos estatísticos usados como previsores de TSM [Barth et al., 2006].

Não obstante, há muitas pesquisas baseadas em modelos de séries temporais estabelecidos construídos sobre observações de TSM passadas [Aguilar-Martinez and Hsieh, 2009; Hertig and Jacobeit, 2010; Lins et al., 2013; Wu et al., 2006]. Hertig and Jacobeit [2010] adotaram modelos estatísticos para investigar a previsibilidade de TSM mensais na área do Mediterrâneo, enquanto Wu et al. [2006] e Aguilar-Martinez and Hsieh [2009] realizam a previsão de TSM do oceano Pacífico tropical usando redes neurais. Aguilar-Martinez and Hsieh [2009] também adotam SVM para TSM. Similarmente, Lins et al. [2013] aplicam SVM para prever a TSM do oceano Atlântico tropical um ano à frente (como nesta pesquisa) usando séries temporais diárias. Lins et al. [2013] aplicaram modelos que assumem a correlação entre cada observação de TSM prevista e os dados do dia anterior ao longo de todos os anos anteriores disponíveis. Reciprocamente, Chattopadhyay [2007] fez uso de redes neurais artificiais e métodos estatísticos para previsão de TSM.

Os benefícios e limitações da agregação temporal em relação a previsão de séries temporais têm sido extensivamente estudados em várias áreas de domínio, especialmente ao longo das últimas décadas [Abraham, 1982; Englund et al., 1999; Nelson and Plosser, 1982; Rostami-Tabar et al., 2013, 2014; Silvestrini and Veredas, 2008; Stram and Wei, 1986; Tiao, 1972; Wei, 1978, 2005] e seu uso está se tornando cada vez mais comum [Kara et al., 2009; Skliris et al., 2011; Wang et al., 2014]. A maioria dos trabalhos que apresentam resultados sobre a influência da agregação temporal para previsão com diferentes horizontes observa que esta depende fortemente das propriedades estatísticas das séries temporais não-agregadas disponíveis, entre eles destacam-se os trabalhos de Rostami-Tabar et al. [2013], que afirmam que a agregação de séries altamente positivamente correlacionadas não melhora a precisão de previsões; e de Tiao [1972] e Rostami-Tabar et al. [2014] que mostram que as vantagens de previsão oferecidas pela agregação temporal de séries temporais são altamente influenciadas pela estacionariedade das séries não-agregadas. Wei [1978] apresenta análise similar focando no fator de sazonalidade das séries disponíveis. Destacam-se também os trabalhos de Nelson and Plosser [1982] que apontam a amplificação da variabilidade de baixa frequência em séries temporais temporalmente agregadas, e Stram and Wei [1986] que discutem o quanto o ajustamento dos componentes autorregressivos, de média móvel e de diferenciação dos modelos ARIMA é afetado pela agregação em séries temporais. Finalmente, os resultados de Englund et al. [1999] foram particularmente interessantes, já que mostram que para o problema de estimação dos preços de habitação e suas volatilidades a aplicação de desagregação temporal foi a abordagem mais vantajosa. Estes trabalhos dão exemplos de como os efeitos da agregação temporal são altamente correlacionados às características do conjunto de dados estudado, especialmente quando há uma considerável disparidade nas propriedades estatísticas das séries temporais não-agregadas disponíveis, que é o caso dos dados de TSM.

Kourentzes et al. [2014] e Petropoulos and Kourentzes [2014] estudam o melhoramento da previsão de séries temporais através de múltiplos níveis de agregação temporal usando um algoritmo *ensemble* para previsão de horizontes de curto a longo prazo. Similarmente, Kourentzes and Petropoulos [2015] e Athanasopoulos et al. [2015] aplicaram este mesmo algoritmo *ensemble*. Kourentzes and Petropoulos [2015] se focaram na previsão de séries temporais multivariadas. Athanasopoulos et al. [2015] fazem uso de uma estratégia mais dinâmica para ponderação dos pesos de diferentes modelos e níveis de agregação temporal para previsão de séries temporais. Estes trabalhos podem potencialmente se beneficiar dos resultados desta pesquisa já que

a análise dos efeitos da agregação temporal dado o número de observações de séries temporais disponíveis e os diferentes tamanhos possíveis de conjuntos de dados de treinamento pode adaptativamente influenciar os pesos atribuídos ao algoritmo *ensemble* aplicado.

Esta pesquisa é complementar às pesquisas relacionadas apresentadas. Não apenas os resultados desta pesquisa podem ser usados como critério base para o planejamento de experimentos de previsão de TSM agregadas, mas também estudos futuros sobre o valor prático de diferentes estratégias de agregação temporal para a previsão de diferentes conjuntos de dados de TSM podem ser baseados nesta metodologia. Além disso, de acordo com o conhecimento adquirido, não há outra pesquisa que realiza uma avaliação comparativa de diferentes abordagens de agregação temporal para previsão de séries temporais de TSM com diferentes horizontes de previsão e diferentes tamanhos de conjuntos de dados de treino.

Capítulo 5

Metodologia de Previsão de TSM

A metodologia aqui apresentada possibilita a avaliação da técnica de agregação temporal para a previsão de TSM do Oceano Atlântico. Assume-se a disponibilidade de várias instâncias de conjuntos de observações de séries temporais de treino e de teste de alta frequência (séries temporais diárias). Esta pesquisa se baseia na previsão de X_{T_0+j} ($j = 1, \dots, N$), para uma série temporal agregada correspondente (semanal e mensal), para permitir a previsão de ambos curto e longo prazo. Mais especificamente, pretende-se obter 52 e 12 passos a frente de séries temporais agregadas semanalmente e mensalmente, respectivamente, prevendo-se assim 12 meses a frente. Com este propósito, define-se $N = 365/m_w$ com $m_w = 7$ e $N = 365/m_m$ com $m_m = 30$, respectivamente para séries temporais agregadas semanais e mensais.

Daqui por diante adotam-se duas abordagens diferentes para a previsão das agregações temporais desejadas (semanais e mensais), referenciadas como *PrePred* e *PostPred*. Na Tabela 1 são descritos os parâmetros e notações da metodologia desta pesquisa. No Algoritmo PrePred e no Algoritmo PostPred os conceitos gerais de ambas abordagens são apresentadas.

Algoritmo PrePred: Previsão de agregações temporais usando dados diários básicos

```

1 Procedimento PrePred( $x_{tr}, x_{ts}, r, m, k$ )
2    $p \leftarrow \text{prev}(x_{tr}, r, k)$ 
3    $pa \leftarrow \text{agregTemp}(p, m)$ 
4    $x_{tsa} \leftarrow \text{agregTemp}(x_{ts}, m)$ 
5    $ea \leftarrow \text{erro}(pa, x_{tsa})$ 
6   retorna  $ea$ 

```

Na primeira abordagem, *PrePred*, representada pelo Algoritmo PrePred, o parâmetro $k = 1$ indica que o modelo ARIMA será ativado e $k = 0$ indica que o *Random Walk* será usado. O conjunto de treinamento de mT_0 observações diárias é representado por x_{tr} . O modelo é então usado para gerar $r = mN$ previsões diárias, que são representadas por p no passo 2 do Algoritmo PrePred. A partir destas previsões, a previsão para as agregações futuras desejadas são geradas pela Equação 3.10, e são representadas por pa . Aqui a notação a significa *agregada*.

A segunda abordagem, *PostPred*, é representada pelo Algoritmo PostPred. No passo 2,

Tabela 1: Parâmetros e notações para a metodologia e avaliação experimental

Variável	Valores Padrão	Descrição
x_t	N/A	série temporal diária básica
X_T	N/A	série temporal agregada correspondente a x_t
x_{tr}	N/A	subconjunto de x_t usado para treinamento
x_{ts}	N/A	subconjunto de x_t usado para teste
m	7 ou 30	parâmetro de unidade de tempo para agregação temporal
m_i	7 ou 30	parâmetro de unidade de tempo para agregação temporal realizada antes da previsão
m_f	7 ou 30	parâmetro de unidade de tempo para agregação temporal realizada após a previsão
N	$ x_{ts} /m = 365/m$	número de observações agregadas de série temporal a serem previstas
r	$ x_{ts} = mN = 365$	número de observações básicas de série temporal a serem previstas
k	ARIMA ou <i>Random Walk</i>	modelo de série temporal a ser aplicado para previsão
<i>PrePred</i>	$\text{PrePred}(x_{tr}, x_{ts}, r, m, k)$	previsão de agregações futuras baseada em dados diários básicos
<i>PostPredW</i>	$\text{PostPred}(x_{tr}, x_{ts}, r, 7, m, \text{ARIMA})$	previsão de agregações futuras baseada em dados semanais agregados
<i>PostPredM</i>	$\text{PostPred}(x_{tr}, x_{ts}, r, 30, 30, \text{ARIMA})$	previsão de agregações futuras baseada em dados mensais agregados

começa-se aplicando a Equação 3.9 a x_{tr} com $m = m_i$ para construir a série agregada X_1, X_2, \dots, X_{T_0} , representada por x_{tra} . Então constrói-se um modelo para x_{tra} , que é usado para gerar as previsões desejadas como na Equação 3.11, representadas como p . Vale a pena mencionar que o passo 5 do Algoritmo PostPred é incluído para considerar o caso em que haja interesse em se prever agregações futuras baseadas em dados previamente agregados (previsão de dados mensais a partir de dados semanais). No caso de $m_i = m_f$, $pa = p$, do contrário pa é gerado analogamente à Equação 3.10 baseado em p .

Prepara-se o conjunto de teste agregado x_{tsa} aplicando-se a Equação 3.9 a x_{ts} com m ou $m = m_f$. Em seguida, os erros das previsões geradas pa são computados nos passos 5 e 7 dos Algoritmos PrePred e PostPred, respectivamente. Tais previsões são comparadas às N observações de teste agregadas em x_{tsa} .

O conceito geral da metodologia experimental desta pesquisa para cada série temporal disponível é apresentado no Algoritmo PredExp. O algoritmo gera previsões baseadas em modelos ARIMA adequadamente treinados e otimizados. Nesta pesquisa também foram construídos mo-

Algoritmo PostPred: Previsão de agregações temporais usando dados agregados

```

1 Procedimento PostPred( $x_{tr}, x_{ts}, r, m_i, m_f, k$ )
2    $x_{tra} \leftarrow \text{agregTemp}(x_{tr}, m_i)$ 
3    $ra \leftarrow r/m_i$ 
4    $p \leftarrow \text{prev}(x_{tra}, ra, k)$ 
5    $pa \leftarrow \text{agregTemp}(p, m_f)$ 
6    $x_{tsa} \leftarrow \text{agregTemp}(x_{ts}, m_f)$ 
7    $ea \leftarrow \text{erro}(pa, x_{tsa})$ 
8   retorna  $ea$ 

```

delos *Random Walk* para os dados básicos (diários), que são usados como base de referência (*baseline*) para propósitos de comparação de performance. Ambos ARIMA e *Random Walk* são modelos estatísticos consideravelmente simples e bem-estabelecidos. De acordo com o ajustamento dos parâmetros desses modelos é possível produzir uma família de outros modelos de previsão como o modelo autorregressivo e de médias móveis. Quando bem ajustados, estes modelos são também capazes de lidar com não-estacionariedade e sazonalidade em séries temporais. Além disso, suas previsões são capazes de inspirar confiança em cientistas de dados já que modelos lineares são interpretáveis, ao contrário de alguns métodos de aprendizado de máquina que são normalmente considerados modelos "caixa-preta".

Algoritmo PredExp: Metodologia experimental

```

1 Procedimento PredExp( $x_{tr}, x_{ts}$ )
2   para cada modelo  $k$  em {ARIMA, Random Walk} faça
3      $edw \leftarrow \text{PrePred}(x_{tr}, x_{ts}, |x_{ts}|, 7, k)$ 
4      $edm \leftarrow \text{PrePred}(x_{tr}, x_{ts}, |x_{ts}|, 30, k)$ 
5   fim
6    $eww \leftarrow \text{PostPred}(x_{tr}, x_{ts}, |x_{ts}|, 7, 7, \text{ARIMA})$ 
7    $ewm \leftarrow \text{PostPred}(x_{tr}, x_{ts}, |x_{ts}|, 7, 30, \text{ARIMA})$ 
8    $emm \leftarrow \text{PostPred}(x_{tr}, x_{ts}, |x_{ts}|, 30, 30, \text{ARIMA})$ 
9   retorna a lista de todos os erros gerados

```

Os Algoritmos PrePred, PostPred e PredExp foram implementados utilizando-se a linguagem e ambiente para computação estatística R [R Development Core Team, 2008]. R foi escolhido devido ao seu alto desempenho, praticidade e flexibilidade, além de sua facilidade de programação e configuração. Como uma ferramenta e ambiente, R é notavelmente adequado para problemas de computação estatística e elaboração de gráficos, e como uma linguagem de programação interpretada, seguindo o paradigma funcional, é muito portátil e altamente polimórfica, o que são características extremamente úteis quando se lida com problemas de ciência de dados. R oferece uma grande variedade de técnicas estatísticas, entre elas a modelagem

linear e não-linear, testes estatísticos clássicos, análise de séries temporais, etc.

Particularmente, já que o processo de obtenção de modelos ARIMA adequados não é usualmente uma tarefa trivial, foi desenvolvido um pacote de funções R chamado *TSPred* [Salles and Ogasawara, 2015] para automatizar o treinamento e previsão de séries temporais com modelos lineares, entre eles, os modelos ARIMA. Estas funções selecionam um modelo ARIMA otimizado, enquanto também consideram sazonalidade, usando a função `auto.arima` do pacote-R *forecast* [Hyndman and Khandakar, 2008]. A função `auto.arima` seleciona um modelo ARIMA otimizado através da definição dos valores de seus parâmetros, p , d e q , computados através de uma variação do algoritmo de Hyndman e Khandakar [Hyndman and Khandakar, 2008; Hyndman and Athanasopoulos, 2013].

Dado um conjunto de dados, *TSPred* é capaz de computar o modelo linear mais adequado dentre os modelos ARIMA, regressão polinomial, entre outros. O objetivo do pacote é a geração de modelos lineares a serem utilizados como bases de referência (*baselines*) para previsão de séries temporais com métodos de aprendizado de máquina. Com isso, ao se utilizar o pacote de funções desenvolvido, o risco de ajustamento inadequado de modelos lineares é minimizado, e conseqüentemente, análises tendenciosas de performance de métodos de previsão podem ser evitadas. O pacote de funções R desenvolvido nesta pesquisa estão publicamente disponíveis [Salles and Ogasawara, 2015] a fim de se permitir novos estudos e aplicações em previsões de séries temporais.

Para se fazer uma análise mais completa do valor prático de cada abordagem adotada, decidiu-se aplicá-las utilizando-se diferentes unidades de tempo. Além disso, um dos objetivos desta pesquisa é estudar as vantagens e limitações de cada unidade de tempo adotada na geração de previsões agregadas. Com isso, PrePred e PostPred são aplicados para gerar tanto previsões agregadas semanais (passos 3 e 6) quanto mensais (passos 4, 7 e 8) pelo Algoritmo PredExp, respectivamente. No passo 7, previsões de agregações mensais são geradas baseadas em dados agregados semanais. Os erros de previsão calculados para agregações semanais e mensais são então analisados estatisticamente para avaliação da eficiência de cada abordagem, usando o teste t ou o teste Wilcox, dependendo da normalidade da distribuição dos erros.

Capítulo 6

Avaliação Experimental

6.1 Conjunto de Dados

Foram realizados diversos experimentos utilizando-se séries temporais diárias coletadas pelo projeto PIRATA. Como citado antes, este projeto corresponde a uma rede de observação composta por vinte e uma boias distribuídas pelo oceano Atlântico tropical, e foi planejada para monitorar uma série de variáveis dos processos de interação oceano-atmosféricas [Bourlès et al., 2008; Servain et al., 1998]. As boias adotadas pelo projeto, conhecidas como *Autonomous Temperature Line Acquisition System* (ATLAS), foram projetadas para medir variáveis meteorológicas com sensores hidrológicos entre a superfície da água e quinhentos metros de profundidade. A localização das vinte e uma boias é apresentada na Figura 2 na Seção 2.3, e alguns exemplos da identificação das boias, referenciadas pelas suas coordenadas, podem ser visualizados na Figura 4.



Figura 4: Localização e exemplos da identificação das boias do projeto PIRATA (Adaptado de GOOS-Brasil [2015])

Uma das variáveis observadas pelo PIRATA é a TSM, medida entre um metro e um metro

e meio de profundidade, cujos dados são utilizados nos experimentos desta pesquisa. As observações diárias de TSM são constantemente atualizadas e estão disponíveis para referência nas páginas da Internet de [GOOS-Brasil \[2015\]](#). O intervalo geral dos dados de medições de TSM disponíveis até a realização desta pesquisa compreende o período de janeiro de 1998 até março de 2015, e estes dados têm sido de grande utilidade para inúmeros outros pesquisadores [[Lins et al., 2013](#)].

6.2 Configurações Experimentais

6.2.1 Seleção de Dados

No conjunto de dados disponível pelo PIRATA, assim como em qualquer campanha de coleta de dados ambientais, dados faltantes são comumente observados como resultado do mau funcionamento de equipamentos de monitoramento ou falhas na aquisição de dados, entre outros fatores. No caso desta pesquisa, nenhuma das boias produziu registros absolutamente completos. As interrupções nas medições podem abranger muitos meses, o que representa um obstáculo para a análise de dados e a aplicação de técnicas de previsão [[Salles et al., 2015](#)]. A Figura 5 apresenta dados de TSM coletados por algumas boias e as interrupções de aquisição destes dados ao longo de anos.

Devido às dificuldades resultantes da ausência de dados, muitas vezes pesquisadores optam por aplicar técnicas de imputação [[Yozgatligil et al., 2012](#)] na série temporal a fim de se computar valores plausíveis para as observações faltantes. Entretanto, esta pode não ser uma opção viável. Um problema bem conhecido na literatura é a acumulação de erros referente à propagação de erros de previsão passados em previsões futuras, o que é uma característica comumente observada na previsão de séries temporais de múltiplos passos a frente (*multistep-ahead prediction*), especialmente quando se adota uma abordagem de previsão de múltiplos estágios (*multi-stage prediction*) [[Cheng et al., 2006](#)]. Exemplos deste fenômeno são apresentados nas Figuras 6a e 6b, que mostram gráficos gerados com o pacote desenvolvido *TSPred* retratando previsões de modelos ARIMA e seus intervalos de confiança junto com os dados reais contidos nos respectivos conjuntos de dados de teste. Nestes gráficos pode-se facilmente observar o estiramento progressivo do intervalo de confiança e perda de representatividade das previsões do modelo ARIMA conforme o horizonte de previsão aumenta.

Sendo assim, caso a quantidade de dados faltantes seja grande o suficiente para que a propa-

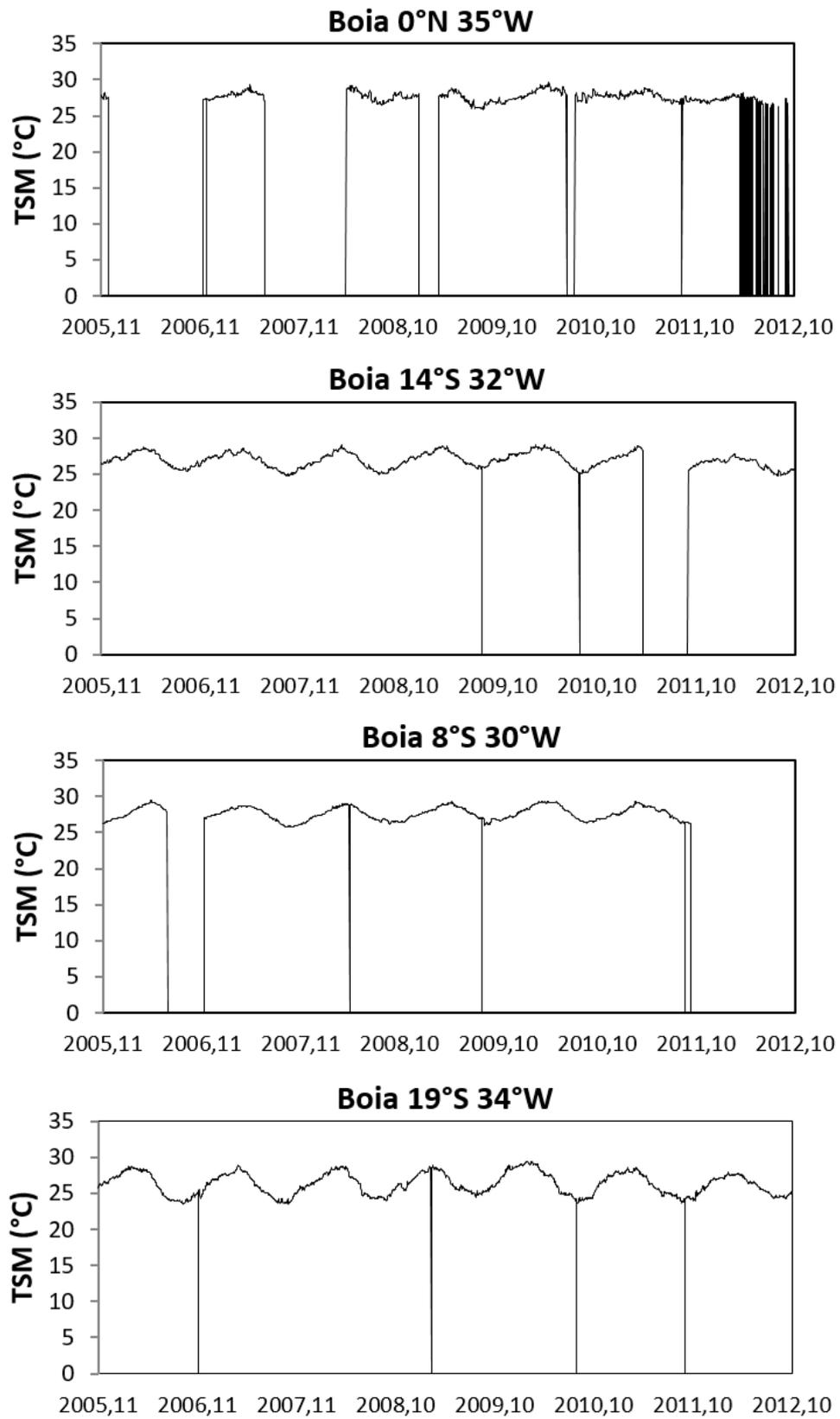
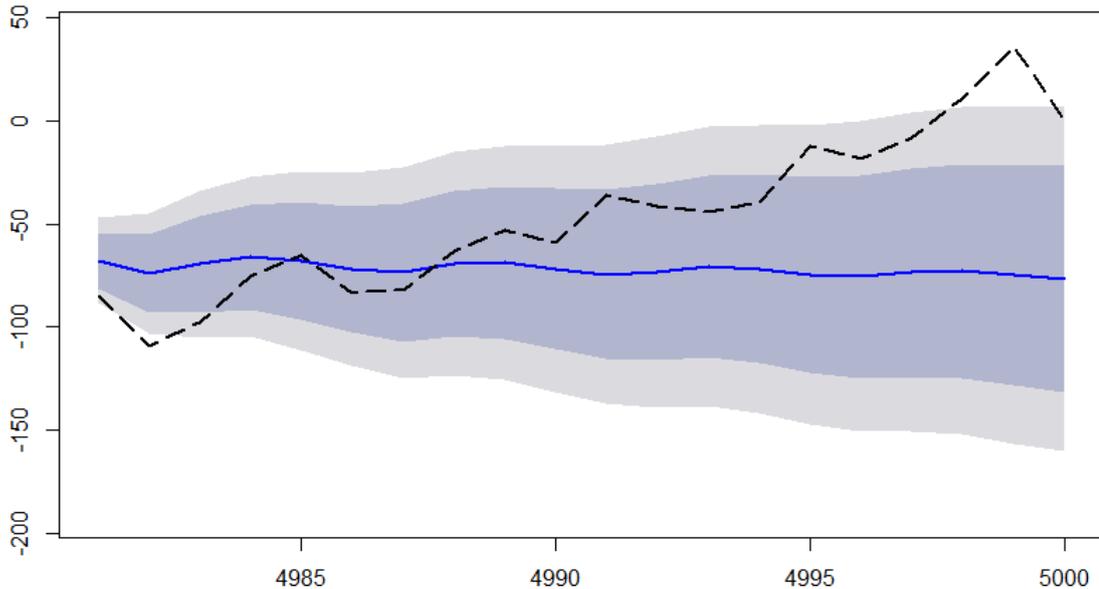
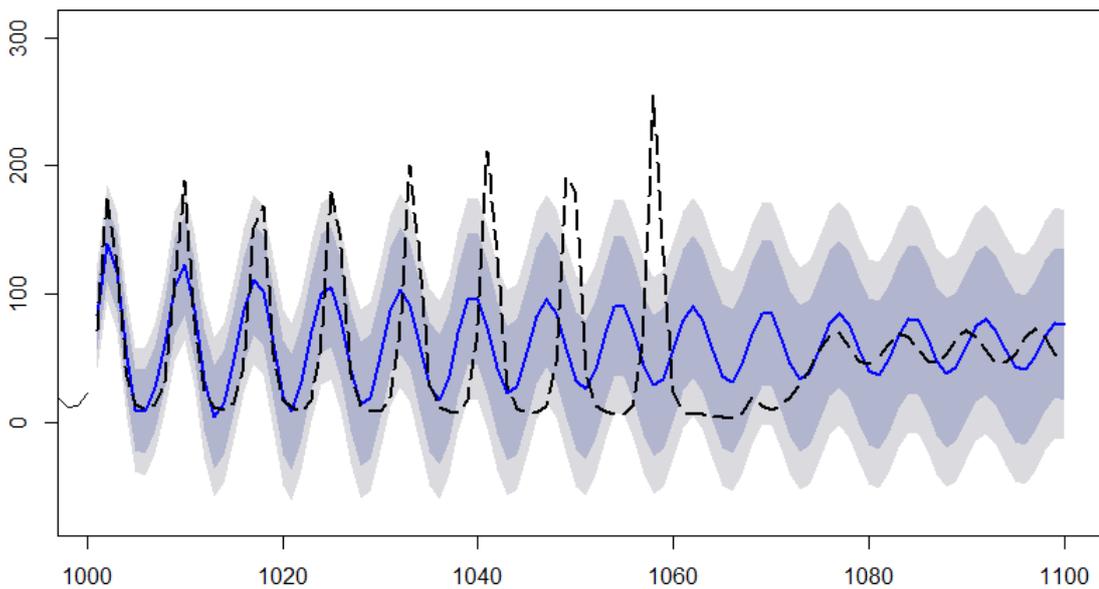


Figura 5: Exemplo de dados de TSM coletados por boias do projeto PIRATA contendo observações faltantes



(a) Previsões ARIMA para o último dos 5 blocos de dados faltantes da série temporal da competição CATS [Lendasse et al., 2004].



(b) Previsões ARIMA para a série temporal A da competição Santa Fe [Weigend, 1993].

Figura 6: Gráficos gerados com o pacote desenvolvido *TSPred* retratando previsões de modelos ARIMA e seus intervalos de confiança junto com os dados reais contidos nos respectivos conjuntos de dados de teste. As previsões são representadas pelas linhas sólidas e os dados reais são representados pelas linhas pontilhadas.

gação de erros de previsão venha a interferir significativamente na análise das performances de métodos de previsão que venham a ser aplicados, o pesquisador pode decidir por particionar e selecionar os dados de TSM, de maneira a produzir conjuntos de dados de diferentes tamanhos utilizados para o treinamento de modelos de previsão e análise de suas performances.

Em consequência, o primeiro passo experimental desta pesquisa foi a seleção de dados de medição ininterruptos para gerar um conjunto de dados que maximize a quantidade de obser-

vações ininterruptas. Para este propósito, foi considerado uma interrupção de dados quando ocorria uma ausência de dados consecutivos correspondentes a um mês de medição. Além disso, foram selecionados apenas períodos com pelo menos dois anos de observações ininterruptas, o que permite a formação de um conjunto dados de treino com no mínimo um ano de observações utilizado para a previsão de um ano à frente. A tabela 2 apresenta para cada boia, referenciada pelas suas coordenadas, os intervalos de tempo dos dados das séries temporais selecionadas para análise e o número de observações em cada série temporal

Tabela 2: Dados ininterruptos de TSM selecionados das boias PIRATA

Boia	Série(s)	No. de dados	Total
0n0e	06/2005 - 05/2008	1070	2159
	10/2010 - 09/2013	1089	
0n10w	07/2006 - 06/2014	2914	2914
	03/1999 - 02/2005	2178	
0n23w	06/2005 - 05/2008	1087	4708
	11/2010 - 10/2014	1443	
0n35w	02/1998 - 01/2004	2183	3247
	04/2009 - 03/2012	1064	
10s10w	01/1999 - 12/2011	4740	4740
12n23w	01/2009 - 12/2013	1799	1799
12n38w	01/2004 - 12/2012	3231	3231
14s32w	01/2006 - 12/2010	1824	1824
15n38w	03/1998 - 02/2004	2184	4737
	12/2005 - 11/2012	2553	
19s34w	11/2005 - 10/2012	2553	2553
20n38w	12/2008 - 11/2014	2183	2183
21n23w	11/2008 - 10/2011	1084	1776
	01/2013 - 12/2014	692	
4n23w	01/2007 - 12/2011	1823	1823
4n38w	08/2000 - 07/2007	2550	2550
6s10w	01/2003 - 12/2014	4342	4342
8n38w	09/2000 - 08/2007	2550	3270
	08/2010 - 07/2012	720	
8s30w	01/2007 - 12/2010	1459	2144
	03/2013 - 02/2015	685	
Total da quantidade de dados			50000

6.2.2 Análise de Dados

Com o objetivo de se desenvolver modelos de previsão adequados e se conseguir uma parametrização e configuração ótimas do método de previsão escolhido, deve-se ter especial atenção às características inerentes dos dados disponíveis. Por isso, foi feita uma análise inicial dos dados no sentido de se obter previsões mais precisas. Neste contexto, as séries temporais de TSM selecionadas foram sujeitas a análises estatísticas nas quais foram executados oito diferentes testes estatísticos para autocorrelação, aleatoriedade e independência, heteroscedasticidade, linearidade e estacionariedade. O conhecimento dessas características estatísticas das séries temporais é muito útil para orientar o processo de previsão subsequente já que pode-se perceber a necessidade de diversas atividades de pré-processamento de dados, tais como transformações de dados e remoção de tendência, e fazer escolhas de modelagem bem informadas.

Para testar a autocorrelação nas séries temporais, ou seja, para investigar se os resíduos de um modelo de regressão linear construído com base nos dados de cada série temporal são independentes, foram realizados os testes de Durbin-Watson e Breusch-Godfrey. Para testar a aleatoriedade e independência nas séries temporais, isto é, para testar se a sequência de observações é aleatoriamente distribuída, foi realizado o teste de Box-Pierce. Para testar a heteroscedasticidade nas séries temporais, isto é, para investigar se a variância estimada dos resíduos de um modelo de regressão linear construído com base nos dados de cada série temporal são dependentes dos valores das variáveis independentes, foi realizado o teste Goldfeld-Quandt. Para testar a linearidade nas séries temporais, isto é, para testar se um modelo de regressão linear construído com base nos dados de cada série temporal seria corretamente especificado como linear, foi realizado o teste *White Neural Network* a fim de se investigar, mais especificamente, se a amostra de observações de séries temporais é linear em relação à média. Finalmente, para testar a estacionariedade nas séries temporais, ou seja, para testar se as séries temporais de observações contêm uma raiz unitária e são, portanto, não-estacionárias, foram realizados dois testes diferentes: o *Augmented Dickey-Fuller* (ADF) e o KPSS.

O teste KPSS foi realizado uma vez para se testar estacionariedade de nível e uma segunda vez para se testar estacionariedade de tendência. Além disso, todos os experimentos foram realizados, com uma definição de número máximo de termos deslocados temporalmente nas séries temporais, parâmetro exigido por quase todas as funções estatísticas aplicadas, proposta pelo manual-R da função *adf.test* do pacote *tseries* [Trapletti et al., 2015].

Nos experimentos de análise estatística executados foram usadas funções dos seguintes pacotes-R: *car* [Fox et al., 2015], *lmtest* [Hothorn et al., 2015], *tseries* [Trapletti et al., 2015], *urca* [Pfaff and Stigler, 2013] e *stats*, que é originalmente encontrado na instalação base do R.

Um resumo dos resultados dos experimentos estatísticos é apresentado na Tabela 3, onde se pode observar a porcentagem dos resultados dos testes que confirmam a Hipótese Nula (H0) de cada teste.

Tabela 3: Resultados e análise dos testes estatísticos

Testes Estatísticos		H0	Análise
Testes de autocorrelação	Durbin-Watson	Resíduos não-correlacionados	0%
	Breusch-Godfrey		0%
Teste de aleatoriedade	Box-Pierce	Aleatoriedade	0%
Teste de heteroscedasticidade	Goldfeld-Quandt	Homoscedasticidade	68%
Teste de linearidade	White Neural Network	Linearidade na média	60%
Testes de estacionariedade	ADF	Não-estacionariedade	44%
	KPSS	Estacionariedade de tendência	64%
	KPSS	Estacionariedade de nível	56%

Pela análise dos resultados, pode-se concluir que nenhuma das séries temporais confirmou as hipóteses nulas de resíduos não-correlacionados e aleatoriedade. Este resultado não foi uma surpresa, já que além da correlação entre medidas consecutivas de temperatura diária, é esperado que os dados contenham sazonalidade devido ao relacionamento entre a TSM e as quatro estações do ano.

A análise dos testes de heteroscedasticidade e linearidade mostra que uma quantidade substancial das séries temporais são de fato homocedásticas com um comportamento linear em torno da média. Estes resultados são muito favoráveis para a obtenção de previsões de séries temporais precisas, já que os dados apresentam um certo nível de estabilidade na sua variância.

Finalmente, com os testes de estacionariedade, pode-se concluir que uma parcela razoável das séries temporais são não-estacionárias, enquanto uma porção também substancial delas são ou estacionárias ao redor de uma tendência, ou ao redor de um nível que muda ao longo do tempo (ou muda a inclinação). Estes resultados motivam a aplicação de atividades de transformação de dados, tais como remoção de tendências, a fim de se conseguir propriedades estacionárias antes de se realizar quaisquer métodos de previsão, que, em geral, assumem o comportamento estacionário de uma série temporal. Neste sentido, os modelos ARIMA apre-

sentam o benefício de apresentarem componentes de diferenciação, que possibilitam a remoção de possíveis tendências nas séries temporais às quais são ajustados.

Sobretudo, devido à existência de considerável disparidade quanto à estacionariedade das séries temporais no conjunto de dados analisado, conhecimentos prévios sobre as vantagens da agregação temporal para previsão [Tiao, 1972] podem não ser diretamente aplicáveis. Com isso, observa-se a necessidade de um estudo experimental sobre a validade prática de se realizar previsões de TSM utilizando-se dados agregados.

6.2.3 Particionamento de Dados

Cada série temporal de dados de TSM selecionada foi particionada em conjuntos de treino e conjuntos de teste para a aplicação da metodologia experimental desta pesquisa. A fim de se promover uma melhor discussão sobre o efeito do número de observações dos conjuntos de treino nas tarefas de construção de modelos (x_{tr}) e previsão (x_{ts}), foi aplicado o Algoritmo TSMPart para o particionamento dos dados.

Algoritmo TSMPart: Particionamento de uma série temporal de TSM em conjuntos de treino e teste

```

1 Procedimento TSMPart ( $x_t$ )
2    $n \leftarrow 365$ 
3    $nAnos \leftarrow |x_t|/n$ 
4   para  $i = 1, \dots, (nAnos - 1)$  faça
5     para  $j = 0, \dots, (nAnos - i + 1)$  faça
6        $x_{tr} \leftarrow x_{nj}, \dots, x_{nj+ni}$ 
7        $x_{ts} \leftarrow x_{nj+ni+1}, \dots, x_{nj+ni+n}$ 
8     fim
9   fim
10  retorna a lista de todos os  $x_{tr}$  e  $x_{ts}$  gerados

```

Um exemplo de aplicação do Algoritmo TSMPart para o particionamento de séries temporais selecionadas da bóia *On0e* pode ser melhor visualizado na Figura 7 onde pode-se observar os diferentes conjuntos de dados de treino e teste usados nestes experimentos, derivados da mesma série temporal original selecionada. O número de observações nos conjuntos de teste é fixo e equivalente ao número de previsões de um ano à frente, enquanto que o tamanho dos conjuntos de treino crescem gradualmente para $|x_t| - |x_{ts}| = |x_t| - n$ observações.

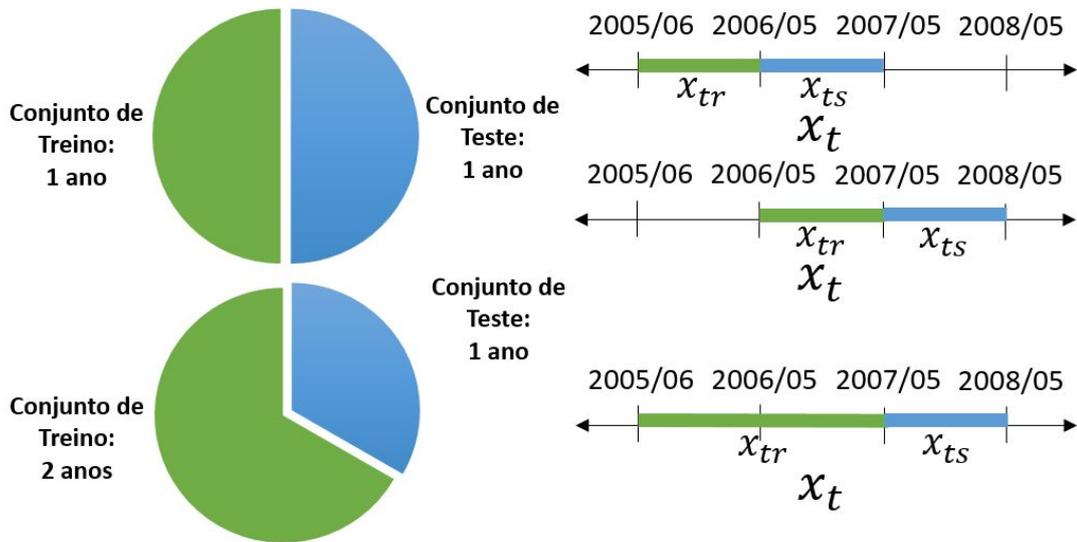


Figura 7: Exemplo de particionamento de dados para as séries temporais da boia *On0e*

6.2.4 Geração de Erros de Previsão

Foram computados doze valores de erro de previsão para cada horizonte de previsões agregadas. Estes erros são cumulativamente produzidos através do aumento do número de passos à frente sob análise, de tal modo que cada um destes e_i erros é computado como descrito na Equação 6.1, onde \hat{X}_{T_0+j} denota tanto $\hat{X}_{T_0+j}^{(1)}$ quanto $\hat{X}_{T_0+j}^{(2)}$. Esta equação pode ser considerada tanto para previsões agregadas semanais (12 semanas à frente) quanto mensais (12 meses à frente).

$$e_i = \frac{\sum_{j=1}^i (X_{T_0+j} - \hat{X}_{T_0+j})^2}{i} \quad (i = 1, \dots, 12) \quad (6.1)$$

A Figura 8 ilustra o conceito geral da produção de erros de previsão de um ano à frente para o caso de agregações mensais, onde se pode observar a sua característica cumulativa.

Embora a Equação 6.1 apresentada aplique a medida MSE, foram adotadas quatro medidas de erro diferentes para análise, incluindo MSE, NMSE, MAPE e SMAPE. Além disso, a restrição de doze semanas à frente para a análise de previsões de agregações semanais foi vantajosa a fim de se padronizar as definições dos horizontes de previsão considerados para discussão nesta avaliação experimental (apresentadas na tabela 4), e também a fim de evitar comparações tendenciosas entre os erros de previsão de agregações semanais e mensais, dado que o número máximo de erros de previsão gerados pela metodologia utilizada nesta pesquisa para cada previsão de um ano à frente de agregações mensais é doze.

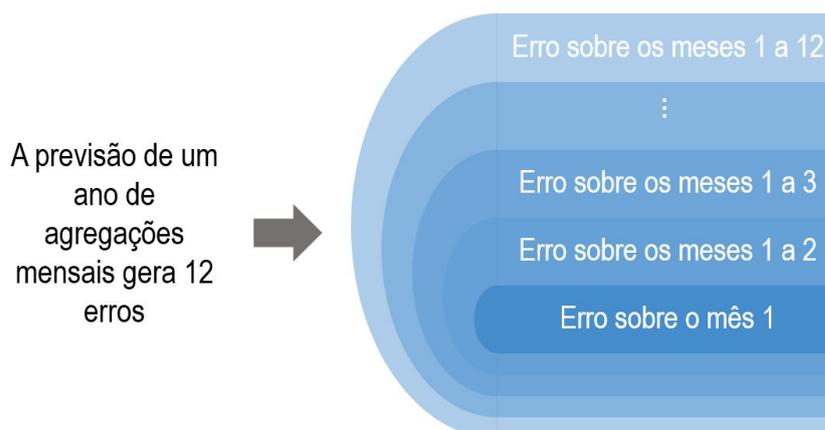


Figura 8: Visão geral da produção de erros de previsão de agregações mensais até um ano à frente

Todos os erros de previsão gerados nesta avaliação experimental estão disponíveis no website criado para esta pesquisa [Salles, 2015]. Outros resultados que podem também ser encontrados neste website [Salles, 2015] são: (i) os resultados concernentes a normalidade dos erros de previsão gerados; (ii) Os modelos ARIMA produzidos na avaliação experimental; (iii) os dados de TSM e funções R usados para a realização da avaliação experimental incluídos em arquivos RData; (iv) códigos para geração de arquivos csv com os erros de previsão de TSM computados e os p-valores de suas análises estatísticas; (v) os resultados dos testes estatísticos realizados para análise das características das séries temporais de TSM no conjunto de dados experimentais; e (vi) os resultados dos testes estatísticos realizados para análise e comparação dos erros de previsão gerados na avaliação experimental.

6.3 Discussão de Resultados

Nos experimentos realizados foram gerados vinte e quatro mil, setecentos e vinte erros de previsão para cada uma das quatro medidas de erro adotadas. Todos estes resultados foram estatisticamente analisados, o que permitiu a comparação do comportamento de *PrePred* e *PostPred*. Além disso, foi possível observar como as performances de previsão foram afetadas pela agregação temporal com diferentes horizontes à frente e tamanhos de conjuntos de treino.

A metodologia de particionamento de dados aplicada nesta pesquisa resultou em conjuntos de treino contendo entre um e onze anos de observações. Contudo, os conjuntos de treino com mais de seis anos de observações vêm de poucas boias (menos de cinco). Com o intuito de evitar conclusões tendenciosas, foram escolhidos para discussão somente os erros provenientes de previsões baseadas em conjuntos de treino com seis anos de observações ou menos, já que

se sabe que eles se originam de pelo menos oito boias diferentes. Além disso, notou-se que não existia diferença estatística entre os erros de previsão computados em medidas de acurácia diferentes, por isso a análise foi limitada a apresentar erros MSE.

Uma visão mais geral do processo experimental aplicado nesta pesquisa para a geração dos dados de erro de previsão analisados e discutidos nesta seção pode ser mais facilmente obtida pela análise da Figura 9. Além disso, na Tabela 4 é apresentada uma definição dos horizontes de previsão considerados para discussão nesta avaliação experimental com relação a geração de erros de previsão de agregações semanais e mensais.

Um critério adotado para esta definição é o fato de que, no contexto de previsão de séries temporais de TSM de alta frequência (diárias), a tarefa de prever observações de até um ano a frente (365 observações a frente), usando uma abordagem de previsão de múltiplos estágios (*multi-stage prediction*), que é o caso nesta pesquisa, implicaria em uma quantidade crescente de acumulação de erros e perda de representatividade das estimativas de temperatura geradas pelos modelos ARIMA [Cheng et al., 2006]. Uma estratégia similar para a definição e discussão dos horizontes de previsão foi adotada por Kourentzes et al. [2014] e por Kourentzes and Petropoulos [2015].

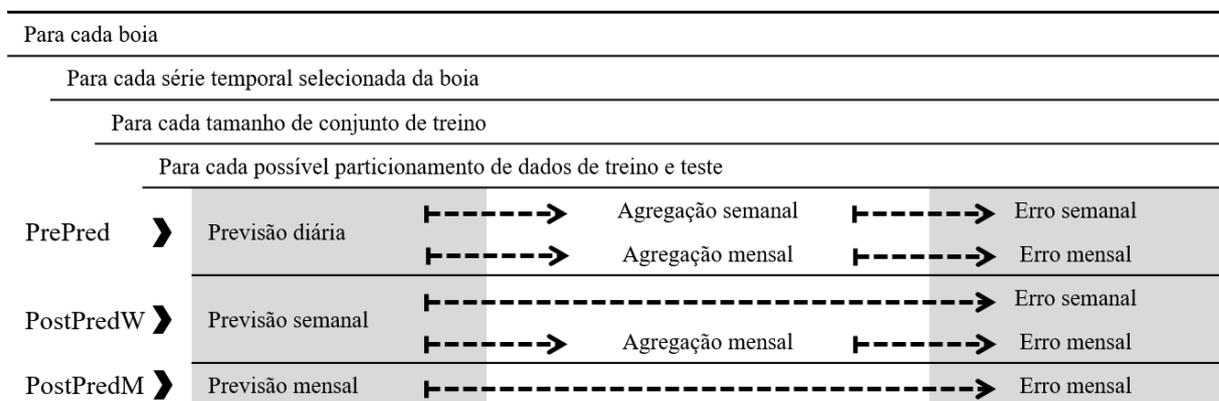


Figura 9: Visão geral do processo experimental aplicado para a geração dos dados de erro de previsão analisados

O modelo *Random Walk* usado na abordagem *PrePred* foi adotado como base de referência (*baseline*). Com isso, foi estudada a validade do uso dos modelos ARIMA através da sua comparação com esta base. A Figura 10 apresenta gráficos com o número relativo de vezes em que cada abordagem de previsão aplicada (*PrePred* e *PostPred*) teve uma melhor performance estatisticamente significativa quando comparada à base de referência *Random Walk*. A Figura 10a refere-se aos erros de previsão para a previsão de 12 semanas à frente, enquanto a Figura

Tabela 4: Definição dos horizontes de previsão considerados na avaliação experimental com relação à geração de previsões de agregações semanais e mensais

Agregações Previstas	Horizonte de Previsão		
	Curto	Médio	Longo
Semanais	1 - 4	5 - 8	9 - 12
Mensais	1 - 4	5 - 8	9 - 12

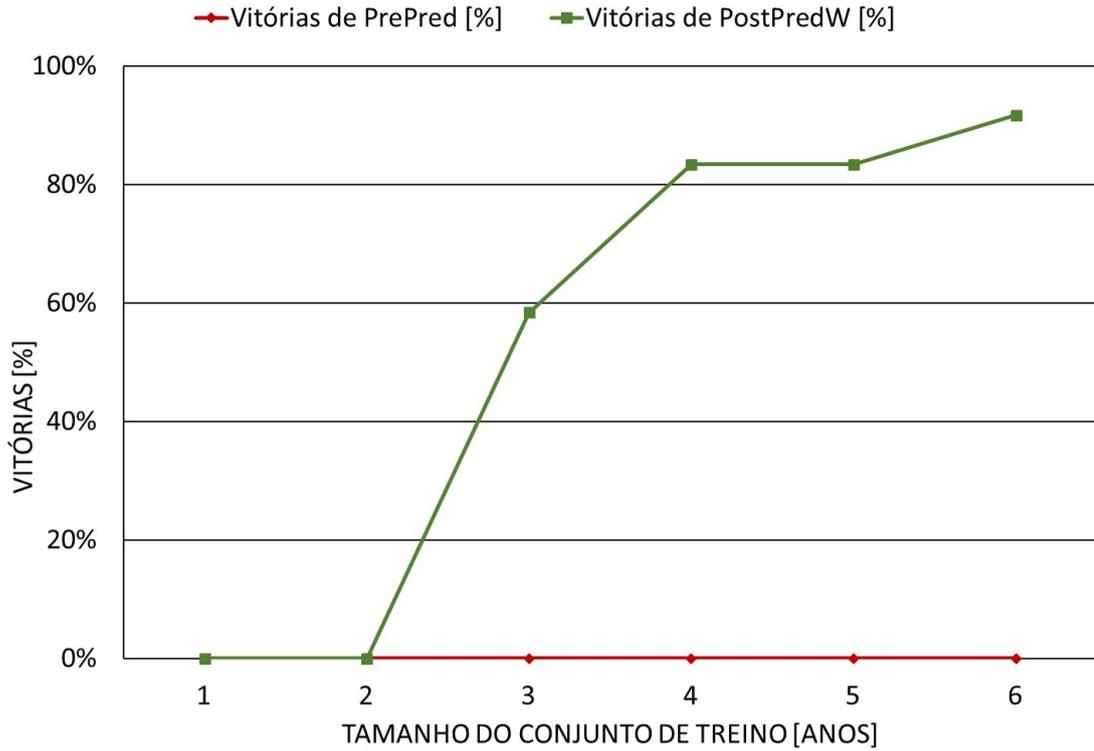
10b refere-se aos erros de previsão para a previsão de 12 meses à frente. Os resultados estão agrupados de acordo com o tamanho (em anos) dos conjuntos de treino modelados.

Pela Figura 10, pode-se notar que não há diferença estatística entre a performance de previsão da abordagem *PrePred* com o ARIMA e *Random Walk*. Isto é verdade particularmente para a geração de futuras agregações semanais. Este resultado é relevante pois leva à conclusão de que a performance do *PrePred* com ARIMA não foi estatisticamente melhor do que usar o *Random Walk* para previsão. Por outro lado, *PostPredW* e *PostPredM* tiveram performances notavelmente melhores conforme os tamanhos dos conjuntos de treino aumentavam. Também foi notado que *PrePred*, *PostPredW* e *PostPredM* em geral pareceram ter performance melhor que a base de referência para previsões de longo prazo, conforme pode ser observado na Tabela 5 e na Tabela 6.

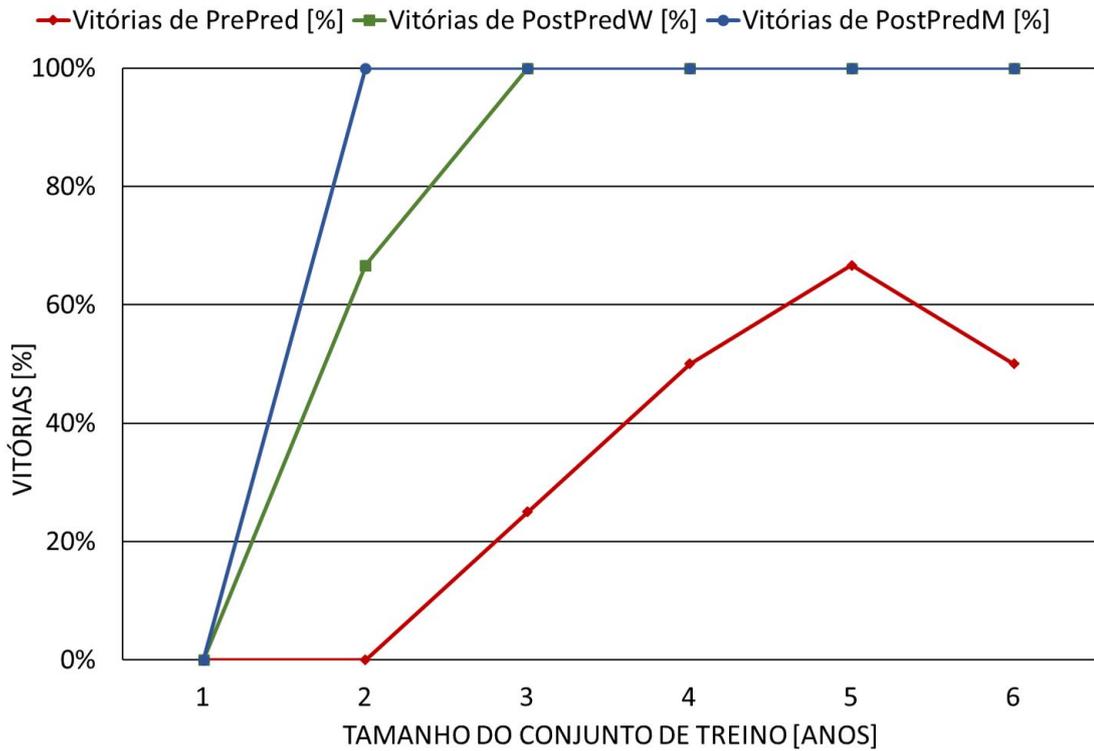
Destacam-se os resultados da comparação estatística entre os erros para cada abordagem aplicada e a base de referência para a geração de futuras agregações semanais na Tabela 5 e agregações mensais na Tabela 6, de acordo com o tamanho dos conjuntos de treino e os horizontes de previsão.

A partir de agora, estuda-se os erros de previsão produzidos pelos experimentos desta pesquisa usando o modelo ARIMA. Começa-se pela análise dos erros obtidos pelos experimentos para geração de previsões agregadas semanais. Para melhor compreensão, define-se *vitória* como o número de vezes em que uma abordagem em particular apresentou uma melhor performance, *i.e.*, menores erros de previsão em comparação a outras abordagens. A Figura 11 mostra um gráfico com as vitórias das abordagens de previsão aplicadas, *PrePred* e *PostPred*, de acordo com o tamanho em anos do conjunto de treino modelado.

Observa-se pela Figura 11 a geral superioridade de performance relativa do *PostPredW* frente ao *PrePred*, particularmente conforme o tamanho do conjunto de treino aumenta. De qualquer forma, *PostPredW* ainda obteve performance melhor na maioria das previsões baseadas em um ano de observações de treino. É importante chamar atenção para este resultado, já



(a) Comparação do ARIMA (*PrePred* e *PostPredW*) contra o *Random Walk* para 12 agregações semanais



(b) Comparação do ARIMA (*PrePred*, *PostPredW* e *PostPredM*) contra o *Random Walk* para 12 agregações mensais

Figura 10: Número relativo de vezes em que cada abordagem de previsão aplicada teve uma melhor performance estatisticamente significativa quando comparada à base de referência *Random Walk*

Tabela 5: Resultados de comparação estatística (p -valores) entre *PrePred* com *Random Walk* e *PostPredW* com ARIMA, com relação às suas performances na geração de previsões agregadas semanais

Horizonte de Previsão [semanas]	Tamanho do Conjunto de Treino [anos]					
	1	2	3	4	5	6
1	1,00	0,96	0,95	0,88	0,93	0,55
2	1,00	0,53	0,19	0,15	0,34	0,05
3	1,00	0,23	0,10	0,01	0,00	0,00
4	1,00	0,16	0,03	0,00	0,00	0,00
5	1,00	0,22	0,03	0,00	0,00	0,00
6	1,00	0,37	0,04	0,00	0,00	0,00
7	1,00	0,41	0,05	0,00	0,00	0,00
8	1,00	0,39	0,06	0,00	0,00	0,00
9	1,00	0,44	0,06	0,00	0,00	0,00
10	1,00	0,48	0,05	0,00	0,00	0,00
11	1,00	0,52	0,05	0,00	0,00	0,00
12	1,00	0,52	0,04	0,00	0,00	0,00

■ *PostPredW* é estatisticamente superior ao *PrePred*

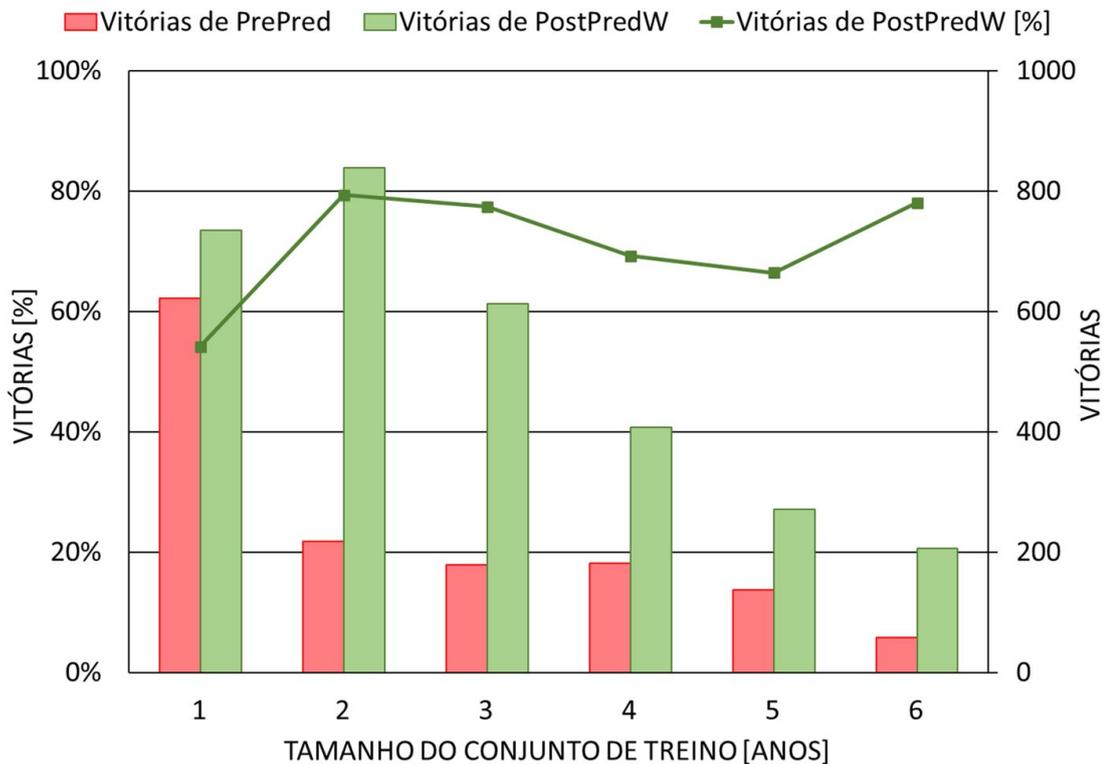


Figura 11: Gráfico das vitórias de cada abordagem de previsão com relação às suas performances na geração de até doze previsões agregadas semanais

equivalente ao tamanho do conjunto de treino.

Para uma análise mais aprofundada, foi analisada a significância estatística da diferença

Tabela 6: Resultados de comparação estatística (p -valores) entre *PrePred* com *Random Walk* e *PrePred*, *PostPredW* e *PostPredM* com ARIMA, com relação às suas performances na geração de previsões agregadas mensais

		HP ¹	Tamanho do Conjunto de Treino [anos]					
			1	2	3	4	5	6
(1) <i>PrePred</i> (<i>Random Walk</i>) vs. (2) <i>PrePred</i> (ARIMA)	1	1,00	1,00	1,00	0,98	0,65	0,83	
	2	1,00	1,00	1,00	0,96	0,51	0,55	
	3	1,00	1,00	1,00	0,83	0,34	0,57	
	4	1,00	1,00	1,00	0,72	0,15	0,32	
	5	1,00	1,00	0,99	0,41	0,04	0,17	
	6	1,00	1,00	0,93	0,11	0,02	0,11	
	7	1,00	1,00	0,63	0,04	0,01	0,05	
	8	1,00	1,00	0,32	0,02	0,00	0,03	
	9	1,00	0,98	0,11	0,00	0,00	0,01	
	10	1,00	0,91	0,02	0,00	0,00	0,01	
	11	1,00	0,85	0,00	0,00	0,00	0,00	
	12	1,00	0,86	0,00	0,00	0,00	0,00	
(1) <i>PrePred</i> (<i>Random Walk</i>) vs. (2) <i>PostPredW</i>	1	1,00	0,15	0,01	0,00	0,00	0,00	
	2	1,00	0,43	0,04	0,00	0,00	0,00	
	3	1,00	0,30	0,01	0,00	0,00	0,00	
	4	1,00	0,10	0,00	0,00	0,00	0,00	
	5	1,00	0,02	0,00	0,00	0,00	0,00	
	6	1,00	0,00	0,00	0,00	0,00	0,00	
	7	1,00	0,00	0,00	0,00	0,00	0,00	
	8	0,99	0,00	0,00	0,00	0,00	0,00	
	9	0,91	0,00	0,00	0,00	0,00	0,00	
	10	0,67	0,00	0,00	0,00	0,00	0,00	
	11	0,48	0,00	0,00	0,00	0,00	0,00	
	12	0,52	0,00	0,00	0,00	0,00	0,00	
(1) <i>PrePred</i> (<i>Random Walk</i>) vs. (2) <i>PostPredM</i>	1	1,00	0,04	0,00	0,01	0,00	0,01	
	2	1,00	0,00	0,00	0,00	0,00	0,00	
	3	1,00	0,00	0,00	0,00	0,00	0,00	
	4	1,00	0,00	0,00	0,00	0,00	0,00	
	5	1,00	0,00	0,00	0,00	0,00	0,00	
	6	1,00	0,00	0,00	0,00	0,00	0,00	
	7	1,00	0,00	0,00	0,00	0,00	0,00	
	8	1,00	0,00	0,00	0,00	0,00	0,00	
	9	1,00	0,00	0,00	0,00	0,00	0,00	
	10	1,00	0,00	0,00	0,00	0,00	0,00	
	11	1,00	0,00	0,00	0,00	0,00	0,00	
	12	1,00	0,00	0,00	0,00	0,00	0,00	

Abordagem (2) é estatisticamente superior à abordagem (1)

¹Horizonte de Previsão [meses]

Tabela 7: Resultados de comparação estatística (p -valores) entre *PrePred* e *PostPredW*, com relação às suas performances na geração de previsões agregadas semanais

Horizonte de Previsão [semanas]	Tamanho do Conjunto de Treino [anos]					
	1	2	3	4	5	6
1	0,85	0,00	0,01	0,08	0,42	0,20
2	0,16	0,00	0,00	0,00	0,05	0,00
3	0,02	0,00	0,00	0,00	0,01	0,00
4	0,00	0,00	0,00	0,00	0,00	0,00
5	0,00	0,00	0,00	0,00	0,00	0,00
6	0,00	0,00	0,00	0,00	0,00	0,00
7	0,00	0,00	0,00	0,00	0,00	0,00
8	0,00	0,00	0,00	0,00	0,00	0,00
9	0,00	0,00	0,00	0,00	0,00	0,00
10	0,00	0,00	0,00	0,00	0,00	0,00
11	0,00	0,00	0,00	0,00	0,00	0,00
12	0,00	0,00	0,00	0,00	0,00	0,00

■ *PostPredW* é estatisticamente superior ao *PrePred*

entre os erros de previsão obtidos pelo *PrePred* e *PostPredW*. Os p -valores resultantes são apresentados na Tabela 7. Foram agrupados todos os erros selecionados gerados a partir de previsões com conjuntos de treino de mesmo tamanho e com os mesmos horizontes de previsão. Embora existam algumas exceções para erros em previsões de curto prazo, de duas semanas ou menos, novamente, a superioridade estatística da performance relativa de *PostPredW* sobre *PrePred* é notável para quase todos os experimentos de previsão similares, assumindo-se 95% de confiança.

Em seguida, são analisados os erros de previsão obtidos pela geração de previsões agregadas mensais. Um gráfico análogo, na Figura 12, apresenta as vitórias de *PrePred*, *PostPredW*, e *PostPredM* de acordo com o tamanho em anos dos conjuntos de treino modelados.

A Figura 12 apresenta uma notável diferença na quantidade relativa de vezes em que *PostPredM* obteve melhores resultados que *PostPredW* e *PrePred*, especialmente conforme o tamanho dos conjuntos de treino aumentava. Ainda assim, para conjuntos de treino com apenas um ano de observações, *PostPredW* apresentou a maioria dos melhores resultados, ultrapassando ambos *PostPredM* e *PrePred*. Assim como para as agregações semanais, os erros obtidos a partir da previsão de agregações mensais futuras foram submetidos a uma análise estatística análoga da significância da diferença entre os resultados obtidos por *PrePred*, *PostPredW* e *PostPredM*. Os p -valores resultantes são apresentados na Tabela 8.

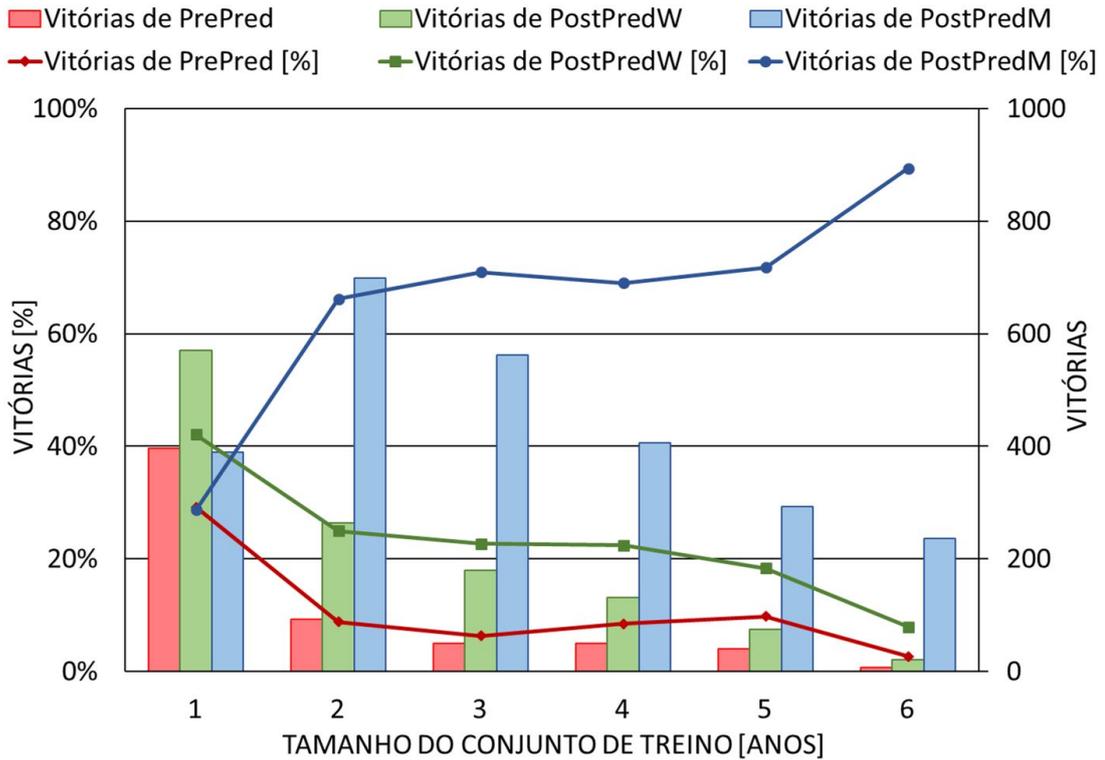


Figura 12: Gráfico das vitórias de cada abordagem de previsão com relação às suas performances na geração de até doze previsões agregadas mensais

A primeira sub-tabela da Tabela 8 apresenta erros para *PrePred* e *PostPredW*. Apesar de poucas exceções, *PostPredW* tem uma performance estatisticamente melhor que *PrePred* independentemente do tamanho dos conjuntos de treino usados e dos horizontes de previsão. Similarmente, a segunda sub-tabela compara *PostPredM* com *PrePred*. *PostPredM* também parece superar *PrePred* em todos os experimentos observados, exceto para conjuntos de treino pequenos (com um ano de observações). Um comportamento similar pode ser notado na terceira sub-tabela, onde *PostPredW* e *PostPredM* são comparados. Novamente, *PostPredM* tem uma performance melhor para previsões com conjuntos de treino com dois ou mais anos de observações e horizontes de previsão maiores ou iguais a dois meses.

Esses resultados confirmam a adequação da abordagem *PostPred* para a geração de previsões agregadas semanais e mensais de dados diários de TSM com horizontes de longo prazo, como foi observado por muitas pesquisas na literatura. Entretanto, os tamanhos dos conjuntos de dados de treino apresentaram uma influência muito relevante com relação às vantagens da aplicação da agregação temporal. Em particular, *PostPred* apresentou uma melhora significativa em performance a medida em que foram aumentados o tamanho dos conjuntos de dados de treino e o nível de agregação. Em geral, *PrePred* apresentou melhores performances para con-

Tabela 8: Resultados de comparação estatística (p -valores) entre *PrePred*, *PostPredW* e *PostPredM*, com relação às suas performances na geração de previsões agregadas mensais

		HP ¹	Tamanho do Conjunto de Treino [anos]					
			1	2	3	4	5	6
(1) <i>PrePred</i> vs. (2) <i>PostPredW</i>	1		0,00	0,00	0,00	0,00	0,01	0,00
	2		0,00	0,00	0,00	0,00	0,00	0,00
	3		0,01	0,00	0,00	0,00	0,00	0,00
	4		0,03	0,00	0,00	0,00	0,00	0,00
	5		0,04	0,00	0,00	0,00	0,00	0,00
	6		0,04	0,00	0,00	0,00	0,00	0,00
	7		0,05	0,00	0,00	0,00	0,00	0,00
	8		0,08	0,00	0,00	0,00	0,00	0,01
	9		0,03	0,00	0,00	0,00	0,00	0,03
	10		0,01	0,00	0,01	0,00	0,00	0,05
	11		0,00	0,00	0,01	0,00	0,00	0,09
	12		0,00	0,00	0,01	0,00	0,00	0,08
(1) <i>PrePred</i> vs. (2) <i>PostPredM</i>	1		1,00	0,00	0,00	0,00	0,00	0,00
	2		1,00	0,00	0,00	0,00	0,00	0,00
	3		1,00	0,00	0,00	0,00	0,00	0,00
	4		1,00	0,00	0,00	0,00	0,00	0,00
	5		1,00	0,00	0,00	0,00	0,00	0,00
	6		1,00	0,00	0,00	0,00	0,00	0,00
	7		1,00	0,00	0,00	0,00	0,00	0,00
	8		1,00	0,00	0,00	0,00	0,00	0,00
	9		1,00	0,00	0,00	0,00	0,00	0,00
	10		1,00	0,00	0,00	0,00	0,00	0,00
	11		1,00	0,00	0,00	0,00	0,00	0,00
	12		1,00	0,00	0,00	0,00	0,00	0,00
(1) <i>PostPredW</i> vs. (2) <i>PostPredM</i>	1		1,00	0,41	0,07	0,49	0,13	0,16
	2		1,00	0,00	0,00	0,01	0,01	0,00
	3		1,00	0,00	0,00	0,00	0,00	0,00
	4		1,00	0,00	0,00	0,00	0,00	0,00
	5		1,00	0,00	0,00	0,00	0,00	0,00
	6		1,00	0,00	0,00	0,00	0,00	0,00
	7		1,00	0,00	0,00	0,00	0,00	0,00
	8		1,00	0,00	0,00	0,00	0,00	0,00
	9		1,00	0,00	0,00	0,00	0,00	0,00
	10		1,00	0,00	0,00	0,00	0,00	0,00
	11		1,00	0,00	0,00	0,00	0,00	0,00
	12		1,00	0,00	0,00	0,00	0,00	0,00

Abordagem (2) é estatisticamente superior à abordagem (1)

¹ Horizonte de Previsão [meses]

juntos de treino consideravelmente pequenos e para previsões de curto prazo. Especificamente para a geração de previsões mensais, *PostPredW* também superou *PrePred* na maioria dos experimentos. Além disso, *PostPredW* surpreendentemente superou *PostPredM* nas previsões com conjuntos de treino pequenos e previsões de curto prazo. Este resultado é particularmente interessante pois sugere as vantagens do uso de dados agregados para previsão, mesmo que estes não estejam na mesma unidade de tempo que as previsões desejadas.

Capítulo 7

Conclusões

A previsão de longo prazo da TSM no Oceano Atlântico Tropical é extremamente relevante como uma forma de auxiliar as agências governamentais e sociedade a prepararem-se para a ocorrência futura de eventos extremos, como as secas. Assim, melhorar a previsão da TSM em diferentes horizontes é uma questão primordial. Esta pesquisa avalia o uso da agregação temporal e suas consequências em diferentes horizontes de previsão de TSM. Além dos dados diários da TSM obtidos do projeto PIRATA, foram modeladas e avaliadas séries temporais semanais e mensais derivadas para analisar o impacto da agregação temporal na previsão de futuras observações agregadas de TSM. Esta pesquisa também explora diferentes tamanhos de conjuntos de dados de treino. Isto se torna particularmente importante quando dados são coletados por sensores (sensores de IoT) onde dados faltantes podem interferir nos tamanhos dos conjuntos de dados de treino.

Apesar dos efeitos da agregação temporal na previsão de séries temporais já terem sido extensivamente abordados na literatura [Tiao, 1972; Wei, 1978; Silvestrini and Veredas, 2008], as pesquisas mostraram que as vantagens e desvantagens da agregação temporal estão fortemente ligadas a propriedades das séries temporais básicas (de menor unidade de tempo) disponíveis. Propriedades como estacionariedade e sazonalidade se mostraram muito importantes para a análise da técnica de agregação temporal, entretanto análises estatísticas demonstraram grande diversidade em tais propriedades observadas na séries temporais de TSM. Com isso, os conhecimentos prévios quanto aos efeitos da agregação temporal não podem ser diretamente aplicados, e um maior estudo se faz necessário.

O estudo consistiu em medir a diferença entre prever-se com dados de alta frequência (séries temporais diárias) e então computar-se os dados agregados (previsões semanais e mensais derivadas de previsões diárias), versus prever-se diretamente com dados agregados (séries temporais semanais e mensais derivadas de séries temporais diárias). Quando previsões foram feitas antes da agregação temporal, tal abordagem foi nomeada como *PrePred*, enquanto que quando previsões foram feitas após a agregação temporal, tal abordagem foi nomeada como *PostPred*. Além disso, por uma questão de justa comparação, o modelo ARIMA foi utilizado como mé-

todo de previsão para ambos *PrePred* e *PostPred* a fim de focar no tamanho da influência da agregação temporal. Adicionalmente, adotou-se o *Random Walk* como um modelo de previsão usado como base de referência (*baseline*), como uma forma de medir a qualidade de ambos *PrePred* e *PostPred*.

Os modelos ARIMA e *Random Walk* são modelos estatísticos bem-estabelecidos que permitem a produção de uma família de outros modelos lineares de previsão como o modelo autorregressivo, e de médias móveis, além de lidar com não-estacionariedade e sazonalidade em séries temporais. Suas previsões inspiram maior confiança em cientistas de dados por serem modelos interpretáveis, e não "caixas-pretas" como são considerados alguns métodos de aprendizado de máquina. Estas características motivaram o uso destes modelos para se executar esta pesquisa em que a análise do comportamento das séries temporais é um recurso interessante.

Já que o processo de obtenção de modelos ARIMA adequados não é usualmente uma tarefa trivial, foi desenvolvido um pacote de funções R [R Development Core Team, 2008] chamado *TSPred* [Salles and Ogasawara, 2015] para automatizar o treinamento e previsão de séries temporais com modelos lineares, entre eles, os modelos ARIMA. Estas funções selecionam um modelo ARIMA otimizado, enquanto também consideram sazonalidade, usando a função *auto.arima* do pacote-R *forecast* [Hyndman and Khandakar, 2008]. A função *auto.arima* seleciona um modelo ARIMA otimizado através da definição dos valores de seus parâmetros, p , d e q , computados através de uma variação do algoritmo de Hyndman e Khandakar [Hyndman and Khandakar, 2008; Hyndman and Athanasopoulos, 2013]. Dado um conjunto de dados, *TSPred* é capaz de computar o modelo linear mais adequado dentre os modelos ARIMA, regressão polinomial, entre outros, a ser utilizado como base de referência (*baselines*) para previsão de séries temporais com métodos de aprendizado de máquina. Ao se utilizar o pacote de funções desenvolvido, o risco de ajustamento inadequado de modelos lineares é minimizado, e análises tendenciosas de performance de métodos de previsão podem ser evitadas. O pacote de funções R desenvolvido nesta pesquisa está publicamente disponível [Salles and Ogasawara, 2015] a fim de se proporcionar novas aplicações.

Embora na revisão da literatura tenha sido observado que *PrePred* é a abordagem mais adotada durante previsões de TSM, as análises experimentais apontaram que esta foi geralmente superada pelo *Random Walk* em previsões semanais e mensais. Tal comportamento não ocorreu quando *PostPred* foi usado. Quando o tamanho do conjunto de treinamento foi maior que dois anos, *PostPred* superou o *Random Walk*. Ademais, foram feitos testes estatísticos sob

diferentes horizontes de previsão entre *PrePred* e *PostPred* e na maioria dos casos, *PostPred* foi significativamente melhor que *PrePred*. *PrePred* foi competitivo apenas quando ambos o horizonte de previsão e o tamanho do conjunto de treinamento eram pequenos.

Os resultados apontaram para a importância da aplicação da agregação temporal para fornecer previsões de longo prazo para a TSM. Além disso, esta pesquisa destaca a influência dos horizontes de previsão e do tamanho dos conjuntos de dados de treinamento sobre os possíveis benefícios da aplicação de agregação temporal para a previsão da TSM, especialmente quando existe significativa diversidade nas características estatísticas das séries temporais disponíveis, cenário que impossibilita o uso direto de grande parte do conhecimento prévio quanto a influência da agregação temporal para a previsão de séries temporais. Esta pesquisa é complementar às muitas outras pesquisas em previsão de TSM estudadas, uma vez que destaca o valor de se estudar a necessidade de aplicação de agregação temporal ao se realizar previsões de TSM. Não apenas os resultados desta pesquisa podem ser usados como critério para o planejamento de experimentos de previsão de TSM, mas também esta metodologia pode servir como base para estudos futuros sobre o valor prático de diferentes estratégias de agregação temporal para a previsão de diferentes conjuntos de dados de TSM de diferentes tamanhos.

Essa pesquisa resultou na produção de dois artigos científicos. O primeiro deles, feito em parceria e baseado principalmente no trabalho de Teixeira [2014], foi submetido e publicado nos anais do congresso IX *Brazilian e-Science Workshop* [Teixeira et al., 2015], que aconteceu na cidade de Recife, Pernambuco, onde o trabalho foi apresentado. O segundo artigo gerado com base nos resultados desta pesquisa, foi submetido à revista *Ecological Informatics*.

Além disso, durante esta pesquisa foram realizados estudos complementares sobre a performance comparativa de diversos modelos lineares, entre eles, os modelos ARIMA, para a previsão e imputação de séries temporais, não apenas de dados de TSM, mas provenientes de quaisquer fontes, frente a métodos de aprendizado de máquina. Estes estudos auxiliaram na escolha dos métodos de previsão lineares aplicados nesta pesquisa, além disso, ajudaram a guiar as atividades de pré-processamento dos dados de TSM.

Esses estudos realizados se justificam pela observação de que avaliações sistemáticas de métodos lineares estabelecidos para comparação de performance de métodos de aprendizado de máquina são muitas vezes necessárias, mas frequentemente negligenciadas durante processos de previsão ou imputação de séries temporais. Sendo assim, nesses estudos explora-se os modelos lineares, entre eles, os ARIMA, como base de referência para a previsão e/ou imputação de

séries temporais.

Os estudos complementares realizados culminaram na produção de dois artigos científicos, sendo um deles submetido à revista *Journal of Applied Statistics*, e o outro submetido e publicado nos anais do XXX Simpósio Brasileiro de Banco de Dados [Salles et al., 2015], que aconteceu na cidade de Petrópolis no Rio de Janeiro, onde o trabalho foi apresentado.

Referências Bibliográficas

- Abraham, B. (1982). Temporal Aggregation and Time Series. International Statistical Review / Revue Internationale de Statistique, 50(3):285–291.
- Aguilar-Martinez, S. and Hsieh, W. W. (2009). Forecasts of Tropical Pacific Sea Surface Temperatures by Neural Networks and Support Vector Regression. International Journal of Oceanography, 2009:e167239.
- Alves, J. M. B., Souza, R., and Campos, J. N. B. (2003). Previsão da anomalia de temperatura da superfície do mar (tsm) no atlântico tropical, com a equação da difusão de temperatura. Revista Climanalise, ano, 3:163–172.
- Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., and Petropoulos, F. (2015). Forecasting with temporal hierarchies. Technical report, University Library of Munich, Germany.
- Atzori, L., Iera, A., and Morabito, G. (2010). The internet of things: A survey. Computer networks, 54(15):2787–2805.
- Barth, A., Alvera-Azcárate, A., Beckers, J. M., and Rixen, M. (2006). Coupling a two-way nested primitive equation model and a statistical SST predictor of the Ligurian Sea via data assimilation. Ocean Modelling, 13(3–4):255–270.
- Beniston, M., Stephenson, D. B., Christensen, O. B., Ferro, C. A. T., Frei, C., Goyette, S., Halsnaes, K., Holt, T., Jylhä, K., Koffi, B., Palutikof, J., Schöll, R., Semmler, T., and Woth, K. (2007). Future extreme events in European climate: an exploration of regional climate model projections. Climatic Change, 81(1):71–95. 00715.
- Bourlès, B., Lumpkin, R., McPhaden, M. J., Hernandez, F., Nobre, P., Campos, E., Yu, L., Planton, S., Busalacchi, A., Moura, A. D., Servain, J., and Trotte, J. (2008). The Pirata Program: History, Accomplishments, and Future Directions. Bulletin of the American Meteorological Society, 89(8):1111–1125.
- Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (2008). Time Series Analysis: Forecasting and Control. Wiley, Hoboken, N.J, 4 edition edition.

- Chattopadhyay, S. (2007). Feed forward Artificial Neural Network model to predict the average summer-monsoon rainfall in India. Acta Geophysica, 55(3):369–382.
- Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., and Rong, X. (2015). Data mining for the internet of things: literature review and challenges. International Journal of Distributed Sensor Networks, 2015:12.
- Chen, M., Mao, S., and Liu, Y. (2014). Big data: a survey. Mobile Networks and Applications, 19(2):171–209.
- Chen, Z. and Yang, Y. (2004). Assessing forecast accuracy measures. Preprint Series, (2004-2010):2004–10.
- Cheng, H., Tan, P.-N., Gao, J., and Scripps, J. (2006). Multistep-Ahead Time Series Prediction. In Ng, W.-K., Kitsuregawa, M., Li, J., and Chang, K., editors, Advances in Knowledge Discovery and Data Mining, number 3918 in Lecture Notes in Computer Science, pages 765–774. Springer Berlin Heidelberg.
- Cho, J., Yeh, P. J. F., Lee, Y.-W., Kim, H., Oki, T., Kanae, S., Kim, W., and Otsuki, K. (2010). A study on the relationship between Atlantic sea surface temperature and Amazonian greenness. Ecological Informatics, 5(5):367–378.
- Collins, D. C., Reason, C. J. C., and Tangang, F. (2004). Predictability of Indian Ocean sea surface temperature using canonical correlation analysis. Climate Dynamics, 22(5):481–497.
- Durand, B., Servain, J., Laurent, H., and Machado, L. A. T. (2005). Tropical Atlantic Moisture Flux, Convection over Northeastern Brazil, and Pertinence of the PIRATA Network*. Journal of Climate, 18(12):2093–2101.
- EM-DAT (2016). The international disasters database.
- Englund, P., Quigley, J. M., and Redfearn, C. L. (1999). The choice of methodology for computing housing price indexes: comparisons of temporal aggregation and sample definition. The journal of real estate finance and economics, 19(2):91–112.
- Feng, S., Hu, Q., and Oglesby, R. J. (2010). Influence of Atlantic sea surface temperatures on persistent drought in North America. Climate Dynamics, 37(3-4):569–586.

- Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W., Zeileis, A., and R-Core (2015). *car: Companion to Applied Regression*.
- Fu, R., Dickinson, R. E., Chen, M., and Wang, H. (2001). How do tropical sea surface temperatures influence the seasonal distribution of precipitation in the equatorial Amazon? *Journal of Climate*, 14(20):4003–4026.
- GLOBAL RISK DATA PLATAFORM (2016). *Global Assessment Reports on Disaster Risk Reduction*.
- GOOS-Brasil (2015). *PIRATA Dataset*. Technical report, <http://www.goosbrasil.org>.
- Grodsky, S. A., Carton, J. A., Provost, C., Servain, J., Lorenzzetti, J. A., and McPhaden, M. J. (2005). 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*, 110(C8).
- Gruber, N., Keeling, C. D., and Bates, N. R. (2002). Interannual Variability in the North Atlantic Ocean Carbon Sink. *Science*, 298(5602):2374–2378.
- Gujarati, D. (2002). *Basic Econometrics*. McGraw-Hill/Irwin, Boston; Montreal, 4 edition edition.
- Hastenrath, S. (1984). Interannual Variability and Annual Cycle: Mechanisms of Circulation and Climate in the Tropical Atlantic Sector. *Monthly Weather Review*, 112(6):1097–1107.
- Hastenrath, S. (2011). Exploring the climate problems of Brazil’s Nordeste: a review. *Climatic Change*, 112(2):243–251.
- Hertig, E. and Jacobeit, J. (2010). Predictability of Mediterranean climate variables from oceanic variability. Part II: Statistical models for monthly precipitation and temperature in the Mediterranean area. *Climate Dynamics*, 36(5-6):825–843.
- Hippert, H. S., Pedreira, C. E., and Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *Power Systems, IEEE Transactions on*, 16(1):44–55.

- Hothorn, T., Zeileis, A., (pan.f), R. W. F., (pan.f), C. C., Millo, G., and Mitchell, D. (2015). lmtest: Testing Linear Regression Models.
- Hu, Z.-Z., Kumar, A., Huang, B., Wang, W., Zhu, J., and Wen, C. (2012). Prediction skill of monthly SST in the North Atlantic Ocean in NCEP Climate Forecast System version 2. Climate Dynamics, 40(11-12):2745–2759.
- Hyndman, R. J. and Athanasopoulos, G. (2013). Forecasting: principles and practice. OTexts.
- Hyndman, R. J. and Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software, 27(3):1–22.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer, 1st ed. 2013. corr. 4th printing 2014 edition edition.
- Kara, A. B., Wallcraft, A. J., Hurlburt, H. E., and Loh, W.-Y. (2009). Quantifying SST errors from an OGCM in relation to atmospheric forcing variables. Ocean Modelling, 29(1):43–57.
- Kourentzes, N. and Petropoulos, F. (2015). Forecasting with multivariate temporal aggregation: The case of promotional modelling. International Journal of Production Economics.
- Kourentzes, N., Petropoulos, F., and Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting, 30(2):291–302.
- Lendasse, A., Oja, E., Simula, O., and Verleysen, M. (2004). Time Series Prediction Competition: The CATS Benchmark. In IJCNN 2004, International Joint Conference on Neural Networks, volume 2, pages 1615–1620, Budapest, Hungary.
- Lins, I. D., Araujo, M., Moura, M. d. C., Silva, M. A., and Droguett, E. L. (2013). Prediction of sea surface temperature in the tropical Atlantic by support vector machines. Computational Statistics & Data Analysis, 61:187–198.
- Liu, W. T. and Juárez, R. I. N. (2001). ENSO drought onset prediction in northeast Brazil using NDVI. International Journal of Remote Sensing, 22(17):3483–3501.
- Liu, Y., Stanturf, J., and Goodrick, S. (2010). Wildfire potential evaluation during a drought event with a regional climate model and NDVI. Ecological Informatics, 5(5):418–428.

- McKee, T. B., Doesken, N. J., Kleist, J., et al. (1993). The relationship of drought frequency and duration to time scales. In Proceedings of the 8th Conference on Applied Climatology, volume 17, pages 179–183. American Meteorological Society Boston, MA, USA.
- Mishra, A. K. and Singh, V. P. (2010). A review of drought concepts. Journal of Hydrology, 391(1–2):202–216.
- Mishra, A. K. and Singh, V. P. (2011). Drought modeling – A review. Journal of Hydrology, 403(1–2):157–175. 00128.
- Moura, A. D. and Shukla, J. (1981). On the dynamics of droughts in northeast Brazil: Observations, theory and numerical experiments with a general circulation model. Journal of the Atmospheric Sciences, 38(12):2653–2675.
- Muller, H., Dumas, F., Blanke, B., and Mariette, V. (2007). High-resolution atmospheric forcing for regional oceanic model: the Iroise Sea. Ocean Dynamics, 57(4-5):375–400.
- Nelson, C. R. and Plosser, C. R. (1982). Trends and random walks in macroeconomic time series: Some evidence and implications. Journal of Monetary Economics, 10(2):139–162.
- NOAA (2016). NOAA’s El Niño Page. National Oceanic and Atmospheric Administration.
- Nobre, P. and Shukla, J. (1996). Variations of Sea Surface Temperature, Wind Stress, and Rainfall over the Tropical Atlantic and South America. Journal of Climate, 9(10):2464–2479.
- Perera, C., Liu, C. H., Jayawardena, S., and Chen, M. (2014). A survey on internet of things from industrial market perspective. IEEE Access, 2:1660–1679.
- Petropoulos, F. and Kourentzes, N. (2014). Improving forecasting via multiple temporal aggregation. Foresight: The International Journal of Applied Forecasting, 2014(34):12–17.
- Pfaff, B. and Stigler, M. (2013). *urca: Unit root and cointegration tests for time series data*.
- R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

- Ronchail, J., Cochonneau, G., Molinier, M., Guyot, J.-L., De Miranda Chaves, A. G., Guimarães, V., and de Oliveira, E. (2002). Interannual rainfall variability in the Amazon basin and sea-surface temperatures in the equatorial Pacific and the tropical Atlantic Oceans. International Journal of Climatology, 22(13):1663–1686.
- Rostami-Tabar, B., Babai, M. Z., Syntetos, A., and Ducq, Y. (2013). Demand forecasting by temporal aggregation. Naval Research Logistics (NRL), 60(6):479–498.
- Rostami-Tabar, B., Babai, M. Z., Syntetos, A., and Ducq, Y. (2014). A note on the forecast performance of temporal aggregation. Naval Research Logistics (NRL), 61(7):489–500.
- Salles, R., Bezerra, E., Soares, J., and Ogasawara, E. (2015). Evaluating Linear Models as a Baseline for Time Series Imputation. In XXX Simpósio Brasileiro de Banco de Dados, Petrópolis, RJ. 00000.
- Salles, R. P. (2015). EIC-Pesquisa: Evaluating Temporal Aggregation for Predicting the Sea Surface Temperature of the Atlantic Ocean. Technical report, <https://sourceforge.net/p/gpca/wiki/SSTPred/>.
- Salles, R. P. and Ogasawara, E. (2015). TSPred: Functions for Baseline-Based Time Series Prediction. Technical report, <http://cran.r-project.org/web/packages/TSPred/index.html>.
- Sapankevych, N. I. and Sankar, R. (2009). Time series prediction using support vector machines: a survey. Computational Intelligence Magazine, IEEE, 4(2):24–38.
- Servain, J., Busalacchi, A. J., McPhaden, M. J., Moura, A. D., Reverdin, G., Vianna, M., and Zebiak, S. E. (1998). A pilot research moored array in the tropical Atlantic (PIRATA). Bulletin of the American Meteorological Society, 79(10):2019–2031.
- Servain, J., Clauzet, G., and Wainer, I. C. (2003). Modes of tropical Atlantic climate variability observed by PIRATA. Geophysical Research Letters, 30(5). 00008.
- Shumway, R. H. and Stoffer, D. S. (2010). Time Series Analysis and Its Applications: With R Examples. Springer, New York, 3rd ed. 2011 edition edition.
- Silvestrini, A. and Veredas, D. (2008). Temporal aggregation of univariate and multivariate time series models: a survey. Journal of Economic Surveys, 22(3):458–497.

- Skliris, N., Sofianos, S., Gkanasos, A., Mantziafou, A., Vervatis, V., Axaopoulos, P., and Lascaratos, A. (2011). Decadal scale variability of sea surface temperature in the Mediterranean Sea in relation to atmospheric variability. Ocean Dynamics, 62(1):13–30.
- Stockdale, T. N., Anderson, D. L. T., Balmaseda, M. A., Doblas-Reyes, F., Ferranti, L., Mogenssen, K., Palmer, T. N., Molteni, F., and Vitart, F. (2011). ECMWF seasonal forecast system 3 and its prediction of sea surface temperature. Climate Dynamics, 37(3-4):455–471.
- Stram, D. O. and Wei, W. W. (1986). Temporal aggregation in the arima process. Journal of Time Series Analysis, 7(4):279–292.
- Sukov, A. I., Soldatov, V. Y., Krapivin, V. F., Cracknell, A. P., and Varotsos, C. A. (2008). A sequential analysis method for the prediction of tropical hurricanes. International Journal of Remote Sensing, 29(9):2787–2798.
- Svoboda, M., LeComte, D., Hayes, M., Heim, R., Gleason, K., Angel, J., Rippey, B., Tinker, R., Palecki, M., Stooksbury, D., Miskus, D., and Stephens, S. (2002). The Drought Monitor. Bulletin of the American Meteorological Society, 83(8):1181–1190. 00464.
- Tadesse, T., Wilhite, D. A., Harms, S. K., Hayes, M. J., and Goddard, S. (2004). Drought monitoring using data mining techniques: A case study for nebraska, usa. Natural Hazards, 33(1):137–159.
- Teixeira, P. M. (2014). Uso de Técnicas de Mudança de Periodicidade de Séries Temporais para Previsão da Temperatura de Superfície do Mar do Oceano Atlântico. Dissertação (Mestrado), Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, Rio de Janeiro.
- Teixeira, P. M., Dubois-Iorgulescu, A.-M., Salles, R., Bezerra, E., Ogasawara, E., and Lima, L. S. d. (2015). Uso de Redes Neurais para Previsão da Temperatura da Superfície do Mar do Oceano Atlântico Tropical. In IX Brazilian e-Science Workshop, Recife, PE. 00000.
- Tiao, G. C. (1972). Asymptotic behaviour of temporal aggregates of time series. Biometrika, 59(3):525–531.
- Trapletti, A., Hornik, K., and code), B. L. B. t. (2015). tseries: Time Series Analysis and Computational Finance.

- Tsai, C.-W., Lai, C.-F., Chiang, M.-C., and Yang, L. T. (2014). Data mining for internet of things: a survey. IEEE Communications Surveys & Tutorials, 16(1):77–97.
- Vianna, M. L., Menezes, V. V., Pezza, A. B., and Simmonds, I. (2010). Interactions between Hurricane Catarina (2004) and warm core rings in the South Atlantic Ocean. Journal of Geophysical Research: Oceans, 115(C7):n/a–n/a.
- Wang, G., Dommenges, D., and Frauen, C. (2014). An evaluation of the CMIP3 and CMIP5 simulations in their skill of simulating the spatial structure of SST variability. Climate Dynamics, 44(1-2):95–114.
- Ward, M. N. and Folland, C. K. (1991). Prediction of seasonal rainfall in the north nordeste of Brazil using eigenvectors of sea-surface temperature. International Journal of Climatology, 11(7):711–743.
- Wei, W. W. (1978). Some consequences of temporal aggregation in seasonal time series models. In Seasonal analysis of economic time series, pages 433–448. NBER.
- Wei, W. W. S. (2005). Time Series Analysis : Univariate and Multivariate Methods. Pearson, Boston, 2 edition edition.
- Weigend, A. S. (1993). Time Series Prediction: Forecasting The Future And Understanding The Past. Westview Press, Reading, MA.
- Wu, A., Hsieh, W. W., and Tang, B. (2006). Neural network forecasts of the tropical Pacific sea surface temperatures. Neural Networks: The Official Journal of the International Neural Network Society, 19(2):145–154.
- Yoon, J.-H. and Zeng, N. (2009). An Atlantic influence on Amazon rainfall. Climate Dynamics, 34(2-3):249–264.
- Yozgatligil, C., Aslan, S., Iyigun, C., and Batmaz, I. (2012). Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theoretical and Applied Climatology, 112(1-2):143–167.
- Zheng, F. and Zhu, J. (2010). Coupled assimilation for an intermediated coupled ENSO prediction model. Ocean Dynamics, 60(5):1061–1073.