UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ

CAMPUS DE FOZ DO IGUAÇU

PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA DE SISTEMAS DINÂMICOS E ENERGÉTICOS

DISSERTAÇÃO DE MESTRADO

ESTUDO DA INFLUÊNCIA DE DIVERSAS MEDIDAS DE SIMILARIDADE NA PREVISÃO DE SÉRIES TEMPORAIS UTILIZANDO O ALGORITMO *KNN-TSP*

JORGE AIKES JUNIOR

FOZ DO IGUAÇU
2012

Jorge Aikes Junior

Estudo da Influência de diversas Medidas de Similaridade na Previsão de Séries Temporais utilizando o Algoritmo *kNN-TSP*

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Sistemas Dinâmicos e Energéticos como parte dos requisitos para obtenção do título de Mestre em Engenharia de Sistemas Dinâmicos e Energéticos. Área de concentração: Sistemas Dinâmicos e Energéticos.

Orientadora: Dr.ª Huei Diana Lee

Foz do Iguaçu
2012

FICHA CATALOGRÁFICA

A291 Aikes Junior, Jorge
Estudo da Influência de diversas Medidas de Similaridade na Previsão de Séries Temporais utilizando o Algoritmo *kNN-TSP*/ Jorge Aikes Junior. - Foz do Iguaçu, 2012.
115p.: il.

Orientadora: Prof [a]. Dr [a]. Huei Diana Lee.
Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia de Sistemas Dinâmicos e Energéticos - Universidade Estadual do Oeste do Paraná.

1. Séries temporais – Análise – Previsão. 2. Mineração de dados.
3. Aprendizagem de máquina. 4. Algoritmos - I. Título.

CDU 519.246
004.6

Miriam Fenner R. Lucas - CRB/9:268 - Unioeste - Campus de Foz do Iguaçu

Estudo da Influência de diversas Medidas de Similaridade na Previsão de Séries Temporais utilizando o Algoritmo *kNN-TSP*

Jorge Aikes Junior

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Engenharia de Sistemas Dinâmicos e Energéticos e aprovada pela Banca Examinadora:
Data da defesa pública: 11/04/2012.

Prof.ª**Dr.ª Huei Diana Lee** - (Orientadora)
Universidade Estadual do Oeste do Paraná - UNIOESTE

Prof. Dr. **Roberto Cayetano Lotero**
Universidade Estadual do Oeste do Paraná - UNIOESTE

Prof. Dr. **Gustavo Enrique Almeida Prado Alves Batista**
Universidade de São Paulo - USP

Resumo

Séries temporais podem ser entendidas como qualquer conjunto de observações que se encontram ordenadas no tempo. Dentre as várias tarefas possíveis com dados temporais, uma que tem atraído crescente interesse, devido a suas várias aplicações, é a previsão de séries temporais. O algoritmo *k-Nearest Neighbor - Time Series Prediction* (*kNN-TSP*) é um método não-paramétrico de previsão de séries temporais que apresenta como uma de suas vantagens a facilidade de aplicação, quando comparado aos métodos paramétricos. Apesar da maior facilidade na determinação de seus parâmetros, algumas questões relacionadas continuam em aberto. Este trabalho está focado no estudo de um desses parâmetros: a medida de similaridade. Esse parâmetro foi avaliado empiricamente utilizando diversas medidas de similaridade em um grande conjunto de séries temporais que incluem séries artificiais, com características sazonais e caóticas, e várias séries reais. Foi realizado também um estudo de caso comparativo entre a precisão da previsão do algoritmo *kNN-TSP* e a dos métodos de Médias Móveis (MA), Autoregressivos de Médias Móveis Integrados Sazonais (SARIMA) univariado e SARIMA multivariado, em uma série de fluxo diário de pacientes na Área de Emergência de um hospital coreano. Neste trabalho é ainda proposta uma abordagem para o desenvolvimento de uma medida de similaridade híbrida, que combine características de várias medidas. Os resultados obtidos neste trabalho demonstram que as medidas da Norma L_p apresentam vantagem sobre as demais medidas avaliadas, devido ao seu menor custo computacional e por apresentar, em geral, maior precisão na previsão de dados temporais utilizando o algoritmo *kNN-TSP*. Apesar de na literatura, em geral, a medida Euclidiana ser adotada como medida de similaridade, a medida Manhattan pode ser considerada candidata interessante para definir a similaridade entre séries temporais, devido a não apresentar diferença estatisticamente significativa com a medida Euclidiana e possuir menor custo computacional. A medida proposta neste trabalho, não apresenta resultados significantes, mas apresenta-se promissora para novas pesquisas. Com relação ao estudo de caso, o algoritmo *kNN-TSP*, com apenas o parâmetro de medida de similaridade otimizado, alcança um erro consideravelmente inferior a melhor configuração com MA, e pouco maior que as melhores configurações dos métodos SARIMA univariado e SARIMA multivariado, sendo essa diferença inferior a um por cento.

Palavras-chave: Séries Temporais, Previsão, *k-Nearest Neighbor - Time Series Prediction*, *k-Nearest Neighbor*.

Abstract

Time series can be understood as any set of observations which are time ordered. Among the many possible tasks appliable to temporal data, one that has attracted increasing interest, due to its various applications, is the time series forecasting. The *k-Nearest Neighbor - Time Series Prediction* (*kNN-TSP*) algorithm is a non-parametric method for forecasting time series. One of its advantages, is its easiness application when compared to parametric methods. Even though its easier to define *kNN-TSP*'s parameters, some issues remain opened. This research is focused on the study of one of these parameters: the similarity measure. This parameter was empirically evaluated using various similarity measures in a large set of time series, including artificial series with seasonal and chaotic characteristics, and several real world time series. It was also carried out a case study comparing the predictive accuracy of the *kNN-TSP* algorithm with the Moving Average (MA), univariate Seasonal Auto-Regressive Integrated Moving Average (SARIMA) and multivariate SARIMA methods in a time series of a Korean's hospital daily patients' flow in the Emergency Department. This work also proposes an approach to the development of a hybrid similarity measure which combines characteristics from several measures. The research's result demonstrated that the L_p Norm's measures have an advantage over other measures evaluated, due to its lower computational cost and for providing, in general, greater accuracy in temporal data forecasting using the *kNN-TSP* algorithm. Although the literature in general adopts the Euclidean similarity measure to calculate de similarity between time series, the Manhattan's distance can be considered an interesting candidate for defining similarity, due to the absence of statistical significant difference and to its lower computational cost when compared to the Euclidian measure. The measure proposed in this work does not show significant results, but it is promising for further research. Regarding the case study, the *kNN-TSP* algorithm with only the similarity measure parameter optimized achieves a considerably lower error than the MA's best configuration, and a slightly greater error than the univariate e multivariate SARIMA's optimal settings presenting less than one percent of difference.

Keywords: Time Series, Forecasting, *k-Nearest Neighbor - Time Series Prediction*, *k-Nearest Neighbor*.

Ao Labi.

Agradecimentos

À professora Huei Diana Lee, por ir muito além do papel de orientadora, me brindando com ensinamentos que transcendem a vida acadêmica. A definição de educador é pequena diante de tanto esforço, dedicação e comprometimento. Obrigado por todos os ensinamentos desses anos de mestrado que serão levados não somente para minha vida, mas para todas àquelas que cruzarem meu caminho.

Aos professores Wu Feng Chung, André Gustavo Maletzke, Carlos Andrés Ferrero, Willian Zalewski e Renato Bobsin Machado por todos os ensinamentos, conselhos e orientações prestados que, novamente, vão além das fronteiras acadêmicas.

Aos mais que colegas, verdadeiros amigos do Labi, Wilson Jung, Luiz Henrique Dutra da Costa, Ricardo Gil Belther Nabo, Antonio Rafael Sabino Parmezan, Leandro Borges dos Santos, Antonio Afonso Dourado Filho, Jefferson Tales Oliva, Simone Aparecida Pinto Romero, Jhonny Marcos Acordi Mertz, Adriele Cristina da Silva, Bianca Espíndola, Dabna Hellen Tomim, Chris Mayara dos Santos, Vanize Meneghetti, Newton Spolaôr, Felipe Conrado Fernandes, Wesley Martins, Pedro Henrique Brusnicki, Alex Guilherme Farina, Leidiane Correa. Obrigado por todos os ensinamentos, curiosidades, risadas e ótimos momentos de descontração passados. Foi um privilégio estar na companhia de vocês.

Aos professores e equipe do PGESDE, por todos os ensinamentos e apoio prestados ao longo desses anos de mestrado. Obrigado também a toda a equipe da UNIOESTE/Foz.

À Fundação Parque Tecnológico Itaipu — FPTI-BR — pelo apoio por meio da linha de financiamento de bolsas prestadas ao PGESDE.

Aos pesquisadores Hye Jin Kam, Jin Ok Sung, Rae Woong Park, do *Department of Biomedical Informatics*, da *School of Medicine, Ajou University* por terem gentilmente cedido os dados da série de fluxo diário de pacientes, utilizados no estudo de caso deste trabalho. Em especial ao Dr. Rae Woong Park, que respondeu aos contatos sobre seus trabalhos de maneira ágil e aberta.

A todos que contribuíram, direta ou indiretamente, para o desenvolvimento deste trabalho e de minha formação, e que por algum motivo não foram mencionados. Perdoem-me por minha omissão.

Ao amigo Wilson Jung, por todos os ensinamentos, conselhos, orientações e principalmente, pela amizade ao longo de todos esses anos. Com certeza tudo isso seria muito mais difícil sem seus conselhos diários. Obrigado por sua presença constante.

Segundo Shakespeare, amigos são a família que Deus nos permitiu escolher. Assim, sinto-me muito afortunado pela minha família, Douglas e Fernanda Comby, Marlon Cezar Gonçalves, Anderson Vieira, Guillherme Wagner e Calvin Berlanda. Obrigado por todo o apoio, incentivo e ajuda prestados durante os anos. Agradeço ainda mais pelos bons momentos passados juntos. A vida não seria nada sem a presença de vocês. Em especial, obrigado Douglas e Marlon, pois os sentimentos de irmãos vão muito além dos que os laços sanguíneos podem determinar. Sinto-me honrado por poder chamá-los de irmãos.

Obrigado a Neide e Néia Balbino por todo o apoio, incentivo, orientação e bons momentos compartilhados. É uma grande alegria para mim nossos caminhos terem se cruzado. Obrigado à Márcia, Vinicius, Ligia e toda família Balbino e Ferreira pelos mesmos motivos.

Em especial, obrigado a Renata Balbino, por todas as alegrias, orientações, conselhos, auxílios, carinhos e atenção prestados. A vida seria em preto e branco sem você. É uma alegria e honra imensa poder estar ao seu lado.

“O que conhecemos é uma gota;
o que desconhecemos é um oceano.”
Isaac Newton

Sumário

Lista de Figuras **xvi**

Lista de Tabelas **xviii**

Lista de Símbolos **xxi**

1 Introdução **1**

1.1 Objetivos 4

1.2 Organização deste Trabalho 5

2 Séries Temporais **7**

2.1 Considerações Iniciais 7

2.2 Notações e Definições 7

2.3 Componentes de Séries Temporais 9

2.3.1 Tendência 9

2.3.2 Sazonalidade 10

2.3.3 Resíduo 11

2.4 Análise de Séries Temporais 12

2.5 Mineração de Dados Temporais 13

2.5.1 Pré-Processamento 14

2.5.2 Recuperação de Conteúdo 15

2.5.3 Agrupamento 16

2.5.4 Classificação de Dados Temporais 17

2.5.5 Detecção de Anomalias 17

2.5.6 Descoberta de *Motifs* 18
2.5.7 Previsão 19
2.6 Aplicações 19
2.7 Considerações Finais 20

3 Previsão de Séries Temporais **21**
3.1 Considerações Iniciais 21
3.2 Métodos para Previsão de Séries Temporais 21
3.2.1 Métodos Paramétricos 22
3.2.2 Métodos Não-Paramétricos 25
3.3 *k-Nearest Neighbor - Time Series Prediction* (*kNN-TSP*) 27
3.3.1 Conceitos de Aprendizagem de Máquina (AM) 27
3.3.2 Descrição do Algoritmo *kNN-TSP* 29
3.4 Aplicações 33
3.5 Considerações Finais 34

4 Similaridade entre Séries Temporais **37**
4.1 Considerações Iniciais 37
4.2 Medidas de Similaridade entre Séries Temporais 37
4.2.1 Norma L_p 39
4.2.2 Canberra 41
4.2.3 Geodésica 42
4.2.4 *Dynamic Time Warping* 43
4.3 Considerações Finais 45

5 Avaliação Experimental **47**
5.1 Considerações Iniciais 47
5.2 Séries Temporais Utilizadas para Avaliação Experimental 47
5.2.1 Séries Artificiais 47

5.2.2 Séries Reais 49
5.3 Configuração Experimental 51
5.4 Resultados e Discussão 55
5.4.1 Séries Artificiais 55
5.4.2 Séries Reais 62
5.4.3 Comparação Geral 68
5.5 Considerações Finais 74

6 Estudo de Caso **75**
6.1 Considerações Iniciais 75
6.2 Previsão do Fluxo Diário de Pacientes 75
6.3 Proposta de uma Medida Composta 78
6.3.1 Seleção de Atributos 80
6.3.2 Composição da Medida 82
6.4 Configuração Experimental 83
6.5 Discussão dos Resultados 84
6.6 Comparativo com o Estudo de Kam 89
6.7 Considerações Finais 90

7 Conclusão **93**
7.1 Principais Contribuições 97
7.2 Limitações 98
7.3 Trabalhos Futuros 99

Referências Bibliográficas **101**

Apêndice A Característica das ST da *NN GC1* **107**

Apêndice B Valores de *MAPE* para as Séries da Avaliação Experimental **109**

Apêndice C Valores de *MAPE* para a Série de Fluxo Diário de Pacientes 115

Lista de Figuras

1.1 Representação esquemática do projeto de Análise Inteligente de Dados de Séries Temporais 4

2.1 Total mensal de passageiros em voos internacionais de 1949 à 1960. 9

2.2 Tendência da série *AirPassengers* 10

2.3 Sazonalidade da série *AirPassengers*. 11

2.4 Resíduo da série *AirPassengers*. 12

3.1 Esquema de funcionamento do *kNN-TSP*. 30

3.2 Exemplo da aplicação do algoritmo *kNN-TSP* 32

3.3 Parâmetros do algoritmo *kNN-TSP* 33

4.1 Efeito da variação de p na Norma L_p para L_1, L_2, L_3 e L_∞. 40

4.2 Efeito da variação de p na Norma L_p para $L_{0.1}$, $L_{0.3}$, $L_{0.5}$ e $L_{0.7}$. 41

4.3 Variação do espaço geométrico de busca da distância Canberra. 42

4.4 Distância Geodésica em um espaço bi-dimensional 43

4.5 Alinhamento entre duas séries 44

4.6 Matriz de custo com rota de alinhamento traçada 45

5.1 Séries temporais artificiais geradas através de modelos sazonais 49

5.2 Séries temporais artificiais geradas por modelos caóticos 49

5.3 Séries reais disponibilizadas pela *NN GC1* 51

5.4 Telas do protótipo para execução de experimentos de previsão de séries temporais 56

5.5 Média e desvio padrão de *MAPE* das séries artificiais para todas as medidas avaliadas. 59

5.6 Média e desvio padrão para todas as medidas avaliadas e séries da *NN GC1*. . . 65

5.7 Média e desvio padrão para todas as medidas avaliadas agrupadas por séries artificiais e da *NN GC1*. 70

6.1 Quantidade de pacientes atendidos pela AE de um hospital coreano entre janeiro/2007 e março/2009. 77

6.2 Média e desvio padrão de *MAPE* da série de fluxo diário de pacientes. 85

Lista de Tabelas

5.1 Características das ST artificiais. 48

5.2 Características das ST disponíveis pela *NN GC1*. 50

5.3 Configuração dos parâmetros *w*, *m* e número de valores previstos para as ST artificiais. 52

5.4 Sazonalidade e quantidade de pontos previstos para as ST da *NN GC1*. 52

5.5 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries artificiais, agrupando os valores de um, cinco e dez vizinhos mais próximos. . . 57

5.6 Comparativo sobre a existência de **d.e.s** entre as medidas de similaridade para as séries artificiais. 61

5.7 Diferença entre t.d.f e t.d.d para as séries artificiais. 62

5.8 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries da *NN GC1*, agrupando os valores de um, cinco e dez vizinhos mais próximos. . . 63

5.9 Comparativo sobre a existência de **d.e.s** entre as medidas de similaridade para as séries da *NN GC1*. 67

5.10 Diferença entre t.d.f e t.d.d para as séries reais. 68

5.11 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para o agrupamento das séries artificiais e das séries da *NN GC1*, incluindo os resultados de um, cinco e dez vizinhos mais próximos. 69

5.12 Comparativo sobre a existência de **d.e.s** entre as medidas de similaridade para as séries artificiais e para as séries da *NN GC1*. 71

5.13 Diferença entre t.d.f e t.d.d para as séries artificiais e da *NN GC1*. 72

6.1 Descrição das variáveis do estudo de caso 77

6.2 Característica das tabelas atributo-valor utilizadas para a SA. 81

6.3 Exemplo de tabela atributo-valor para a SA. 81

6.4 Resultados da SA sobre os dados experimentais. L_1 representa a medida Manhattan, L_2 a medida Euclidiana e L_3 a Métrica L_3. 83

6.5 Composição da medida proposta para a série de fluxo diário de pacientes. . . . 83

6.6 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para a série de fluxo diário de pacientes, agrupando os valores de um, cinco e dez vizinhos mais próximos. 84

6.7 Comparativo sobre a existência de **d.e.s** entre as medidas de similaridade para a série de fluxo diário de pacientes. 86

6.8 Diferença entre t.d.f e t.d.d para a série de fluxo diário de pacientes. 87

6.9 Três melhores configurações e valores de *MAPE* do estudo inicial de Kam et al. (2010) e deste trabalho para a previsão da série de fluxo de pacientes. 90

A.1 Características das ST disponíveis pela *NN GC1*. 107

B.1 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries artificiais para um vizinho próximo. 109

B.2 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries artificiais para cinco vizinhos próximos. 110

B.3 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries artificiais para dez vizinhos próximos. 110

B.4 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries da *NN GC1* para um vizinho próximo. 111

B.5 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries da *NN GC1* para cinco vizinhos próximos. 111

B.6 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries da *NN GC1* para dez vizinhos próximos. 112

B.7 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para o agrupamento das séries artificiais e das séries da *NN GC1* para um vizinho próximo. . 112

B.8 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para o agrupamento das séries artificiais e das séries da *NN GC1* para cinco vizinhos próximos. 112

B.9 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para o agrupamento das séries artificiais e das séries da *NN GC1* para dez vizinhos próximos. 113

C.1 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para a série de quantidade de fluxo diário de pacientes para um vizinhos próximo. 115

C.2 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para a série de quantidade de fluxo diário de pacientes para cinco vizinhos próximo. 115

C.3 Valores de média, desvio padrão, máximo e mínimo de *MAPE* para a série de quantidade de fluxo diário de pacientes para dez vizinhos próximo. 115

Lista de Símbolos

δ	Medida de distância
λ	Distância calculada por uma medida de similaridade
ω	Peso de ponderação da medida híbrida
Φ	Valor de ponderação do componente auto-regressivo de um modelo SARIMA
ϕ	Valor de ponderação de um modelo AR
Θ	Valor de ponderação do componente de média móvel de um modelo SARIMA
θ	Valor de ponderação de um modelo MA
a	Valor previsto
B	Operador de defasagem
D	Ordem de diferenciação sazonal do modelo SARIMA
d	Grau do operador de diferença (modelo ARIMA)
e	Fator inovação (ruído) de modelos de séries temporais.
f	Função de previsão
h	Ordem de um modelo AR
i	Índice
j	Índice
k	Quantidade de vizinhos próximos
L	Tamanho da rota de alinhamento de duas séries temporais
M	Defasagem de uma série
m	Tamanho de uma série temporal aleatória
N	Ruído de uma série temporal

n	Tamanho de uma série temporal aleatória
$O(.)$	Custo computacional de determinada operação
P	Tendência de uma série temporal
p	Valor que identifica a medida de similaridade a ser usada na Norma L_p
Q	Ordem do coeficiente de médias móvel sazonal do modelo SARIMA
q	Ordem de um modelo MA
R	Rota de alinhamento de duas séries temporais
S	Sazonalidade de uma série temporal
s	Ordem da sazonalidade do modelo SARIMA
T	Série temporal aleatória
t	Instante de tempo
u	Função dinâmica desconhecida
v	Função dinâmica desconhecida
W	Quantidade de dimensões em um vetor
w	Tamanho de janela de previsão
x	Vetor aleatório
y	Vetor aleatório
Z	Série temporal aleatória
AE	Área de Emergência
AG	Algoritmo Genético
AM	Aprendizagem de Máquina
AR	Auto-regressivos
ARIMA	Auto-regressivos de Médias Móveis Integrados
ARMA	Auto-regressivos de Médias Móveis
C_k	Critério para seleção dos vizinhos próximos

d.e.s	Diferença estatisticamente significativa
DF	Dimensão Fractal
DTW	*Dynamic Time Warping*
IC	Intra Correlação
kNN	*k-Nearest Neighbor*
kNN-TSP	*k-Nearest Neighbor - Time Series Prediction*
L_1	Distância Manhattan
L_2	Distância Euclidiana
L_3	Métrica L_3
L_∞	Distância de Chebyshev
L_p	Relativo à Norma L_p– conjunto de medidas de similaridade
LS	*Laplacian Score*
MA	Médias Móveis
MAPE	*Mean Absolute Percentage Error*
MD	Mineração de Dados
MDT	Mineração de Dados Temporais
M_s	Medida de Similaridade
NN	*Nearest Neighbors*
NN GC1	*Time Series Forecasting Grand Competition for Computational Intelligence*
RE	*Representation Entropy*
RNA	Rede Neuronal Artificial
SA	Seleção de Atributos
SARIMA	Auto-regressivos de Médias Móveis Integrados Sazonais
STC	Série Temporal de Modelo Caótico
STS	Série Temporal de Modelo Sazonal

t.d.d	Total de **d.e.s** desfavorável
t.d.f	Total de **d.e.s** favorável
UTI	Unidade de Terapia Intensiva

Capítulo 1

Introdução

O avanço da tecnologia em diversas áreas do conhecimento, impulsionado tanto pela crescente capacidade computacional quanto pela diminuição de custos, tem motivado a utilização de sistemas computacionais para a aquisição e o gerenciamento de dados em diversas áreas. Esses sistemas possibilitam, dependendo do objetivo, o armazenamento de grandes volumes de dados em diferentes formatos. Por exemplo, em processos de monitoramento para analisar fenômenos biológicos, pode ser de interesse o armazenamento de dados numéricos, áudio e vídeo, bem como da informação temporal que permita organizá-los cronologicamente (Ferrero et al., 2009; Maletzke, 2009; Verplancke et al., 2010; Lo, 2011).

A possibilidade de armazenar dados que mantenham sua característica temporal permite uma série de análises sobre os fenômenos de origem desses dados, atraindo assim grande atenção. Dados, em geral, podem ser classificados em quatro grandes categorias (Roddick and Spiliopoulou, 2002):

Estáticas: Os dados não apresentam nenhum contexto temporal e não é possível realizar inferências sobre esse contexto;

Sequências: Os dados são ordenados como uma lista de eventos, isto é, apresentam uma ordenação sem marcações temporais. Assim, dados dessa categoria permitem algumas poucas relações como “antes de” e “depois de”;

Com Marcação Temporal: Os dados estão ordenados e representam uma sequência programada de eventos, geralmente adquiridos em intervalos com certa regularidade, e com marcação do tempo;

Completamente Temporal: Os dados estão ordenados e apresentam uma ou mais relações temporais vinculadas. Assim, dados dessa categoria podem possuir relações temporais multivariadas, como a marcação do tempo e o tempo de transação.

A análise manual desses dados, com o intuito de extrair conhecimento e padrões relevantes, pode ser uma tarefa demorada e sujeita à subjetividade, tornando-se, em alguns casos,

inviável devido à alta complexidade da relação entre os dados. Nesse sentido, métodos e ferramentas computacionais têm sido desenvolvidos para auxiliar na análise mais completa desses dados. A área de Mineração de Dados, com o apoio da área de Inteligência Artificial, tem como principal objetivo desenvolver métodos e ferramentas para a extração de conhecimento em bases de dados (Han and Kamber, 2006). No entanto, os métodos tradicionais de Mineração de Dados para a construção de modelos não levam em consideração a característica temporal na análise dos dados. Desse modo, pesquisas têm sido desenvolvidas propondo a adaptação desses métodos para a análise de dados, nos quais o tempo constitui um fator importante (Chiu et al., 2003; Ferrero, 2009; Ding et al., 2010).

De acordo com diferentes objetivos de análise e compreensão de dados temporais, foram propostas diversas tarefas de interesse, dentre as quais podem ser citadas: a classificação, a recuperação por conteúdo, o agrupamento, a detecção de anomalias, a identificação de padrões (*Motifs*), a extração de regras de associação e a previsão de valores futuros (Last et al., 2004).

A previsão de dados temporais, que consiste em calcular dados desconhecidos tomando como base um conjunto de dados conhecidos, é uma tarefa que tem atraído a atenção de pesquisadores de diversas áreas do conhecimento. A utilização dos termos "previsão" e "predição" é fruto de algumas discussões. Alguns autores adotam o termo predição para descrever métodos subjetivos de estimação e previsão para descrever métodos objetivos, enquanto outros utilizam o termo previsão para descrever qualquer meio de estimar o futuro e predição para métodos sistemáticos de realização dessa estimação. Há ainda um terceiro grupo que emprega os termos indistintamente, geralmente utilizando com mais frequência o termo previsão (Chatfield, 2004; Morettin and Toloi, 2006). Neste trabalho será adotado o termo previsão para determinar a estimação de valores futuros de uma série temporal.

Ao longo do tempo foram desenvolvidas diversas abordagens para a previsão de dados. Como exemplos dessas abordagens têm-se as paramétricas, que assumem que os dados respeitam alguma distribuição conhecida e modelam parâmetros de funções que se ajustem a essa distribuição dos dados; e as não-paramétricas, que buscam modelar o comportamento sem o conhecimento prévio da distribuição dos dados (Chatfield, 2004; Morettin and Toloi, 2006; Cryer and Chan, 2008).

As abordagens não-paramétricas podem ainda ser dividas em globais e locais. As primeiras empregam funções de aproximação global, utilizando toda a série temporal; as segundas dividem a série em conjuntos menores de séries, apresentando funções de aproximação para cada um dos subconjuntos, sendo que a função de aproximação de um subconjunto é válida apenas para esse subconjunto (Karunasinghe and Liong, 2006).

Dentre as abordagens para a previsão de séries temporais não-paramétricas de aproximação local existentes, pode-se citar a adaptação do algoritmo de Aprendizagem de Máquina (AM) *k-Nearest Neighbor* (*kNN*). Neste trabalho foi empregada a adaptação desse algoritmo de-

nominada *k-Nearest Neighbor - Time Series Prediction* (*kNN-TSP*) proposta por Ferrero (2009). O *kNN-TSP* consiste em encontrar as k sequências mais similares dentro da série, a partir de uma sequência de referência, e utilizar os valores dessas sequências similares e uma função de previsão, para realizar o cálculo do valor futuro da série.

Nessa abordagem, um dos parâmetros que influencia a exatidão do algoritmo consiste na medida de similaridade selecionada para identificar as sequências similares na série. Assim, determinar a influência desse parâmetro constitui tema de importância. Desse modo, neste trabalho é apresentado um estudo da influência de diversas medidas de similaridade na previsão de dados em diversas séries temporais, tanto artificiais quanto reais, utilizando como método de previsão o algoritmo não-paramétrico *kNN-TSP*. Além disso, é introduzida a proposta inicial de uma medida de similaridade composta por outras medidas, buscando associar características de diferentes medidas.

Este trabalho faz parte do projeto de Análise Inteligente de Dados em uma parceria entre o Laboratório de Bioinformática (LABI) da Universidade Estadual do Oeste do Paraná (UNIOESTE)/Foz do Iguaçu, o Serviço de Coloproctologia da Faculdade de Ciências Médicas da Universidade Estadual de Campinas (UNICAMP)/Campinas, o Laboratório de Inteligência Computacional (LABIC) do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP)/São Carlos e o Grupo Interdisciplinar em Mineração de Dados e Aplicações (GIMDA) da Universidade Federal do ABC (UFABC)/Santo André. Nesse projeto colaboram pesquisadores de distintas áreas, como ciência da computação, biologia e saúde no desenvolvimento de métodos e ferramentas que auxiliem na identificação, na avaliação e no monitoramento de eventos relacionados principalmente à saúde.

Na Figura 1.1 é ilustrado o projeto de Análise Inteligente de Dados de Séries Temporais. Esse projeto está dividido em três etapas:

Primeira Etapa: Essa etapa consiste na realização do pré-processamento de séries temporais, possibilitando assim uma melhor descrição e entendimento dos dados;

Segunda Etapa: Essa etapa objetiva a extração de conhecimento relevante sobre o conjunto de dados, através da aplicação de métodos computacionais. A realização da extração de conhecimento relevante, como a geração de padrões, facilita o entendimento de diversos fatores comportamentais dos dados, bem como a possibilidade de extrair conhecimento não detectável através de outros métodos;

Terceira Etapa: Essa etapa consiste na avaliação e na validação do conhecimento extraído junto a especialistas do domínio, objetivando a análise de dados biomédicos e o suporte à tomada de decisão.

Neste trabalho, situado na segunda etapa do projeto, é estudada a influência de um dos

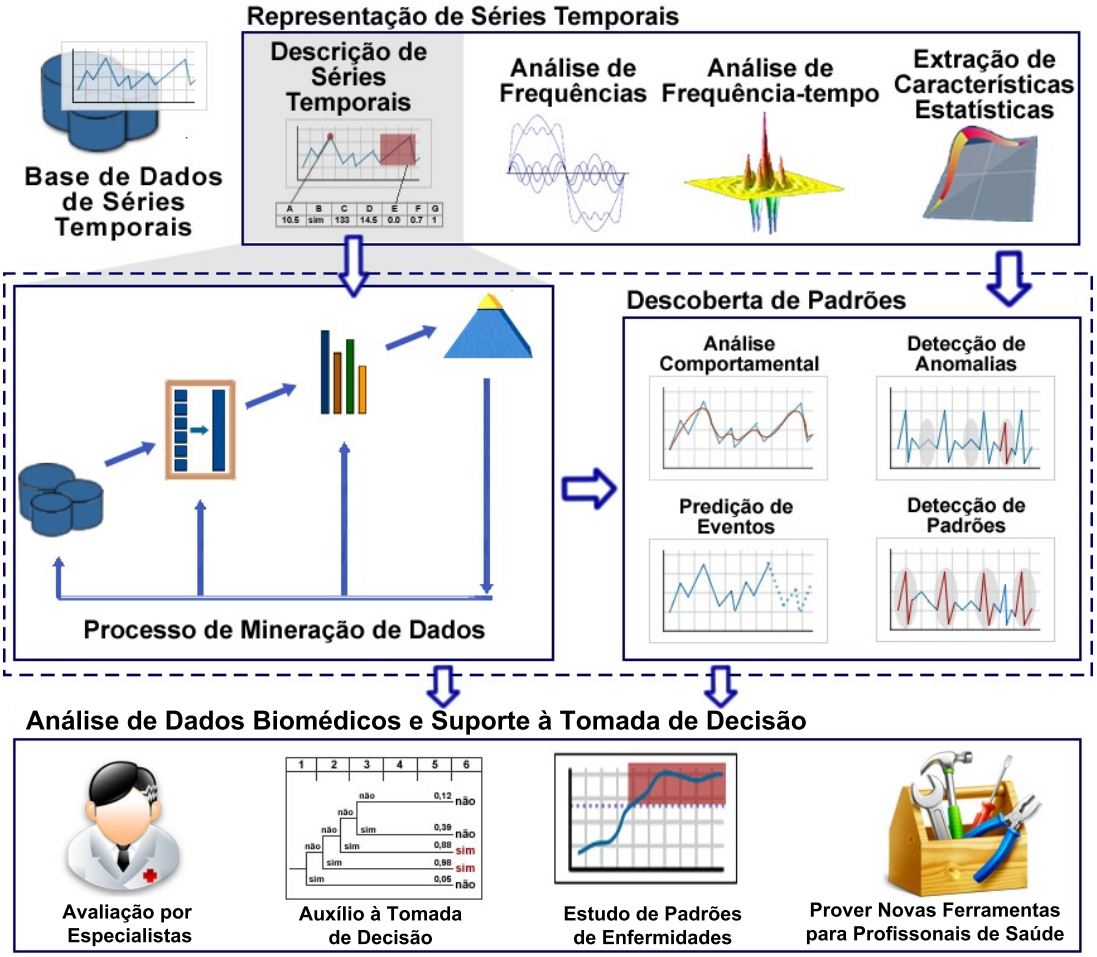


Figura 1.1: Representação esquemática do projeto de Análise Inteligente de Dados de Séries Temporais (Modificado de Ferrero (2009) e Maletzke (2009)).

parâmetros do algoritmo *kNN-TSP* na previsão de séries temporais.

1.1 Objetivos

Os principais objetivos deste trabalho são:

Objetivo Geral:

O objetivo geral deste trabalho consiste em estudar a influência do parâmetro de medida de similaridade na exatidão da previsão de série temporais utilizando o algoritmo *kNN-TSP*. O conhecimento da influência desse, bem como dos demais parâmetros, possibilita a realização de ajustes otimizados ao algoritmo, podendo levar a previsões mais exatas.

Objetivos Específicos:

- Estudo de diversas medidas de similaridade, de maneira a conhecer as suas características;

- Estudo do comportamento do algoritmo *kNN-TSP* utilizando várias medidas frente a séries temporais de diferentes características;
- Estudo comparativo do algoritmo *kNN-TSP* com abordagens tradicionais da literatura;
- Proposta inicial de uma medida de similaridade.

1.2 Organização deste Trabalho

O presente trabalho está organizado do seguinte modo:

Capítulo 2 — Séries Temporais: Nesse capítulo são introduzidos alguns conceitos sobre séries temporais, suas componentes, definições e notações utilizadas. São apresentadas também as etapas de análise de séries temporais, bem como várias tarefas de mineração de dados temporais, além de alguns exemplos de aplicação dessas tarefas;

Capítulo 3 — Previsão de Séries Temporais: Nesse capítulo é tratada a tarefa de previsão de valores de séries temporais. São descritos alguns dos principais métodos paramétricos e não-paramétricos de previsão de séries temporais existentes. O algoritmo *kNN-TSP*, método de previsão empregado neste trabalho, bem como conceitos de AM necessários para sua compreensão também são brevemente descritos neste capítulo, além de alguns exemplos de aplicações que realizam previsão de dados temporais;

Capítulo 4 — Similaridade entre Séries Temporais: Nesse capítulo são descritas as medidas de similaridade avaliadas neste trabalho. A maneira que elas definem a semelhança entre indivíduos, sua interpretação geométrica e o custo computacional também são discutidos;

Capítulo 5 — Avaliação Experimental: Nesse capítulo é apresentada a avaliação experimental realizada. São descritas as características das séries temporais artificiais e reais utilizadas nos experimentos, bem como a configuração experimental e os meios de análise. A discussão e a análise dos resultados também são descritas nesse capítulo;

Capítulo 6 — Estudo de Caso: Nesse capítulo é apresentada a problemática de superlotação em Áreas de Emergência de hospitais, bem como um estudo realizado por Kam et al. (2010) que busca amenizar esse problema empregando métodos paramétricos de previsão de séries temporais, para estimar o fluxo diário de pacientes nessa área em um hospital coreano. Uma avaliação das medidas de similaridade utilizando o algoritmo *kNN-TSP* para a previsão desse mesmo fluxo diário de pacientes, bem como um estudo comparativo entre as duas abordagens é também realizado. Além disso, uma proposta inicial de desenvolvimento de uma medida de similaridade composta é descrita nesse capítulo;

Capítulo 7 — Conclusão: Nesse capítulo são apresentadas as conclusões do trabalho, bem como as principais contribuições, limitações e trabalhos futuros.

Capítulo 2

Séries Temporais

2.1 Considerações Iniciais

Dados temporais apresentam como uma de suas características mais importantes a relação temporal entre as observações. Essa característica faz com que a representação e a análise de dados com fatores temporais apresentem peculiaridades, em relação à análise de dados tradicionais. Essa relação temporal permite também análises sob novas perspectivas, as quais podem auxiliar a descoberta de novos conhecimentos sob o conjunto de dados.

Neste capítulo são apresentados conceitos e definições referentes às Séries Temporais (ST). São abordadas as características das séries[1], suas componentes e os objetivos de diversos métodos de análise de ST, bem como detalhes de tarefas de Mineração de Dados (MD) voltadas a dados temporais. Ao final do capítulo são mencionados alguns exemplos de aplicações.

2.2 Notações e Definições

As ST podem ser entendidas como qualquer conjunto de observações que se encontram ordenadas no tempo e podem ser originadas a partir da saída de um sistema dinâmico (Morettin and Toloi, 2006). Considere um sistema dinâmico representado pela Equação 2.1:

$$\begin{aligned} v_{t+1} &= f(v_t, u_t) \\ z_t &= g(v_t) \end{aligned} \tag{2.1}$$

onde u e v representam os estados das entradas do sistema e $u \in \Re$ e $v \in \Re$; f e g são funções não-lineares desconhecidas; e z_t é uma saída escalar conhecida. Assim, pode-se definir uma ST Z de tamanho m como um conjunto ordenado de valores, ou seja, $Z = (z_1, z_2, \ldots, z_m)$ onde $z_t \in \Re$ e representa uma observação z em um instante t (Chiu et al., 2003).

[1] Neste trabalho, os termos série e série temporal são usados indistintamente.

As ST apresentam-se em diversas áreas do conhecimento, adquiridas através da mensuração de várias situações. Dentre os numerosos exemplos existentes, alguns podem ser citados (Chatfield, 2004):

- Séries Temporais Econômicas, como preços de produtos, quantidade de entradas (ou saídas) de mercadorias em sucessivos meses e lucro da companhia durante os anos;
- Séries Temporais Físicas, que ocorrem em ciências físicas, como meteorologia e geofísica. Exemplos desse tipo de série incluem quantidade de chuva em sucessivos dias e temperatura do ar durante os meses do ano;
- Séries Temporais Demográficas, geralmente usadas em estudos de população, ou previsão de comportamento populacional. Exemplos incluem a observação do crescimento populacional e a taxa de óbitos anuais;
- Controle de Processos, nos quais as ST podem ser utilizadas para detectar alterações no desempenho de um processo ou manufatura. Como exemplos, podem-se citar a verificação gráfica do processo de manufatura e a análise de passos ou processos, que se afastam de um valor alvo ao longo do tempo;
- Séries Temporais de Processos Binários, ocorrem em dados que podem assumir apenas dois valores, os quais geralmente são representados por zeros (0) e uns (1). Por exemplo, processos de comunicação, nos quais existem os estados "ligado" e "desligado", podem ser representados por esse tipo de série.

Na Figura 2.1 pode-se observar um exemplo de série temporal (*AirPassangers*) relativa ao total mensal de passageiros em voos internacionais, entre os anos de 1949 e 1960[2]. O eixo das abscissas apresenta o tempo (em anos) e o eixo das ordenadas a quantidade de passageiros (em milhares) em voos internacionais.

As séries temporais podem ser subdivididas em dois grupos, de acordo com o intervalo em que os dados são adquiridos (Brockwell and Davis, 2002):

Série Temporal Discreta: São as ST cujas observações dos dados são realizadas de maneira discreta, ou seja, em um determinado período de tempo, geralmente igualmente espaçado;

Série Temporal Contínua: Nestas séries, as observações são realizadas em pequenos intervalos de tempo buscando chegar próximo a um intervalo contínuo de tempo.

Ainda como característica de uma ST, podem-se verificar três componentes básicos: tendência, sazonalidade e resíduo (Morettin and Toloi, 2006), os quais são apresentados a seguir.

[2]Dados disponíveis no pacote "*The R Datasets Package*" do ambiente R (http://www.r-project.org/).

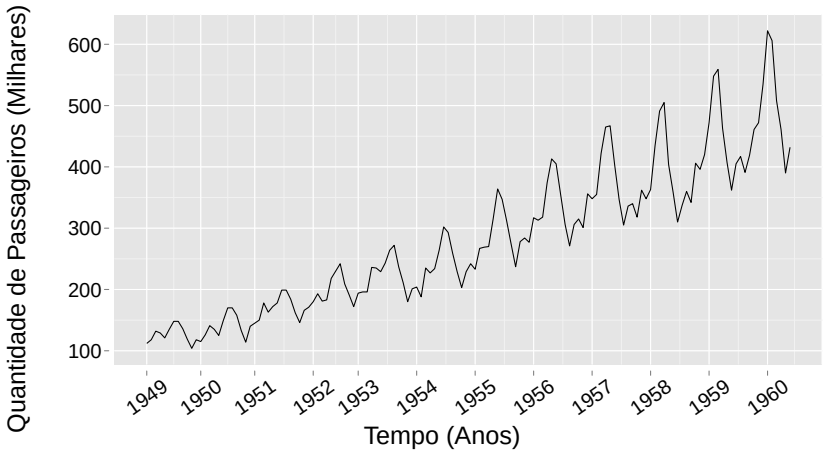


Figura 2.1: Total mensal de passageiros em voos internacionais de 1949 à 1960 (Modificado de *The R Datasets Package*).

2.3 Componentes de Séries Temporais

Como mencionado, uma série temporal pode ser descrita por um conjunto de componentes: tendência, sazonalidade e resíduo. Muitos métodos de análise de ST separam esses componentes, buscando apresentar características importantes da série representadas por eles (Pyle, 1999; Morettin and Toloi, 2006). Desse modo, a composição de uma série temporal pode ser descrita pela Equação 2.2:

$$Z_t = I_t + S_t + N_t \tag{2.2}$$

onde Z representa a série e I, S e N representam a tendência, a sazonalidade e o resíduo (ruído) em um instante t, respectivamente. A seguir, cada um desses componentes é apresentado em detalhes.

2.3.1 Tendência

A tendência de uma ST pode ser entendida como o movimento regular e lentamente desenvolvido durante a série. Apresenta-se como um movimento de longa duração em uma série (Brocklebank and Dickey, 2003).

Essa tendência pode ser tanto crescente quanto decrescente e pode assumir uma grande variação de padrões, dentre esses pode-se citar (Ehlers, 2009):

Crescimento Linear: Caracterizado por seguir uma proporção linear, por exemplo, sempre apresentar um crescimento constante para os dados;

Crescimento Exponencial: Ocorre quando a taxa de crescimento é proporcional a uma função

exponencial, por exemplo, crescimento cujo valor seguinte é o quadrado do anterior;

Crescimento Amortecido: Ocorre quando a taxa de crescimento de dados futuros é menor que os dados atuais, como em situações em que para um determinado ano o crescimento esperado é de setenta por cento (70%) do ano anterior.

Na Figura 2.2, pode-se observar a tendência linear da ST *AirPassengers*, representada pela linha tracejada.

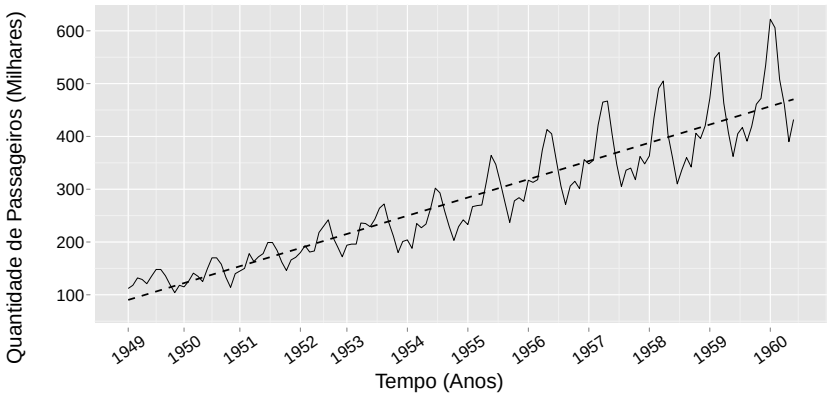


Figura 2.2: Tendência, representada pela linha tracejada, da série *AirPassengers* (Modificado de *The R Datasets Package*).

2.3.2 Sazonalidade

A sazonalidade corresponde a movimentos similares dentro de uma ST. Ela pode ser consideravelmentes regular ou também sofrer alterações suaves ao longo de um grande período de tempo (Brocklebank and Dickey, 2003).

A sazonalidade pode ocorrer em vários casos, estando presente em uma grande quantidade de séries. Um exemplo de sazonalidade ocorre quando determinado dado é coletado mensalmente e verifica-se que o valor do dado de um determinado mês está bastante relacionado com o valor desse dado no mesmo mês de anos anteriores. ST de vendas ou leituras de temperaturas podem exibir variações anuais regulares fáceis de serem detectadas indicando, assim, sua sazonalidade (Chatfield, 2004).

Dependendo da análise realizada, o componente sazonal pode ser de interesse e por isso ele pode ser explicitamente mensurado. Em contraste, em alguns casos, como na interpretação de fatores estatísticos em séries temporais econômicas, por exemplo, desemprego, estes fatores sazonais devem ser reconhecidos e removidos, fornecendo assim dados sem sazonalidade. Esse processo de remoção de fatores sazonais é conhecido como ajuste de sazonalidade (Brockwell and Davis, 2002; Chatfield, 2004).

A sazonalidade pode ser classificada em dois tipos (Ehlers, 2009):

Sazonalidade Aditiva: Séries que se incluem nessa classificação possuem flutuações sazonais independentes do nível global da série, ou seja, suas flutuações são próximas a constantes. Por exemplo, espera-se que um determinado produto tenha um crescimento de vendas de mil unidades no mês de julho em todos os anos. Esse valor é independente da quantidade vendida no restante do ano;

Sazonalidade Multiplicativa: Séries nessa classificação possuem flutuações sazonais dependentes do nível global da série. Por exemplo, espera-se que a venda de um determinado produto em julho aumente trinta e cinco por cento (35%). Verifica-se nesse caso, que em anos com maior quantidade de vendas o aumento no mês de julho é maior do que em anos com menores vendas.

Na Figura 2.3 pode-se observar a sazonalidade da ST *AirPassengers*.

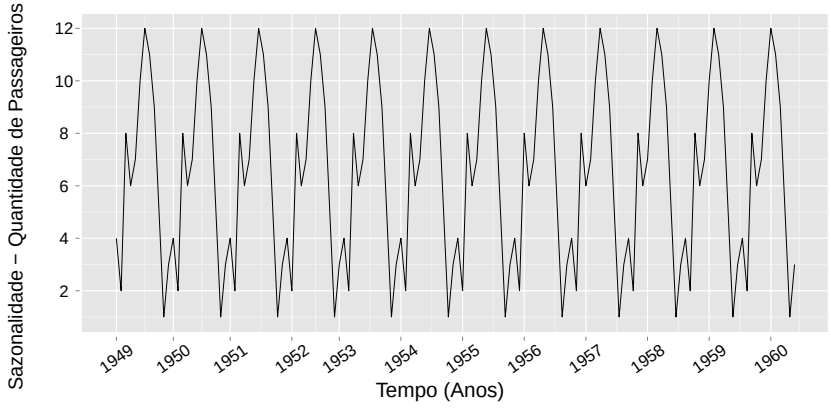


Figura 2.3: Sazonalidade da série *AirPassengers*.

2.3.3 Resíduo

O resíduo, também chamado de ruído, é um componente composto por todos os movimentos que não pertencem à sazonalidade ou à tendência. A existência do resíduo em algumas análises pode trazer problemas, como no caso de estimação estatística de séries econômicas, que pode ser influenciada pela autocorrelação dos resíduos (Kirchgässner and Wolters, 2007).

O ruído N de um instante t de uma ST pode ser dado pela seguinte equação (Ferrero, 2009):

$$N_t = Z_t - (I_t + S_t) \tag{2.3}$$

onde Z representa a série e T e S representam a tendência e a sazonalidade, respectivamente.

Os resíduos podem ser aleatórios ou seguir alguma distribuição. Para a análise de resíduos existem vários métodos, tais como Médias Móveis ou modelos Auto-regressivos. A análise de resíduos pode ser importante para a verificação de variações cíclicas que podem ocorrer dentro do conjunto de dados classificado como resíduo (Chatfield, 2004).

Na Figura 2.4 pode ser verificado o resíduo da ST *AirPassengers*.

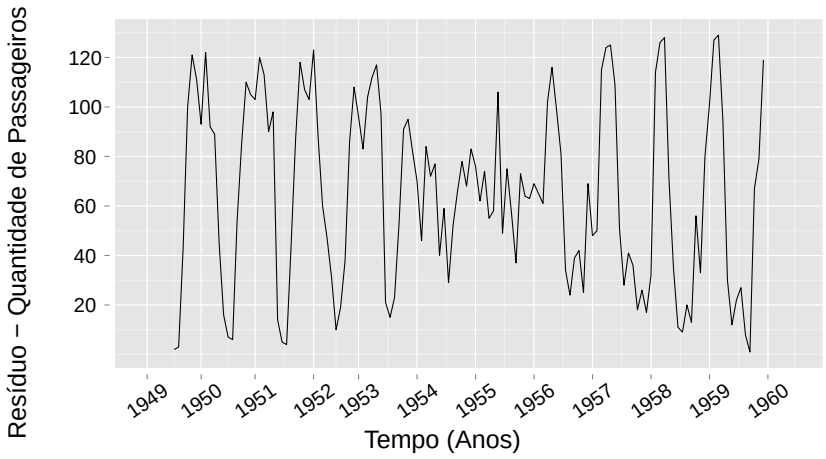


Figura 2.4: Resíduo da série *AirPassengers*.

2.4 Análise de Séries Temporais

A análise de uma ST pode ser realizada seguindo diversos objetivos, os quais podem exigir análises sob diferentes perspectivas. De maneira geral, os objetivos da análise de dados temporais podem ser classificados em quatro grupos (Chatfield, 2004):

Descrição: Essa análise visa descrever os comportamentos da ST, tais como a existência ou não de tendência, sazonalidade, dados discrepantes (*outliers*), alterações da estrutura da série, como a existência de pontos de curva (mudança de padrão de uma tendência ou sazonalidade crescente para decrescente), entre outros. A análise descritiva da série é geralmente um dos primeiros passos em qualquer análise de ST;

Explicação: Essa análise utiliza-se de duas ou mais variáveis. Busca encontrar as relações entre duas séries, ou seja, explicar a variação de uma série com base em outra;

Previsão: Nessa análise, dados os valores passados de uma ST, objetiva-se prever os possíveis valores futuros da série;

Controle: Essa análise é utilizada quando os valores de uma ST representam dados de controle sobre um determinado processo, visando mensurar a qualidade desse processo.

Atualmente, existe uma grande quantidade de métodos e processos que podem auxiliar nessas análises. Alguns desses processos serão apresentados nas seções a seguir, relativos à Mineração de Dados Temporais (MDT).

2.5 Mineração de Dados Temporais

Uma das dificuldades da área de MD é lidar com dados que possuem informações temporais. A dificuldade recai, nestes casos, na necessidade da visualização dos dados como uma sequência de eventos, para assim permitir compreender o fenômeno completamente. Os atributos com dependências temporais necessitam ser tratados de maneira diferente dos outros tipos de atributos, mas, apesar disso, a maior parte das técnicas de MD tratam os dados temporais como uma coleção de eventos desordenados, ignorando então detalhes de informações temporais (Antunes and Oliveira, 2001). Dessa forma, faz-se necessária a adaptação de métodos tradicionais de MD, adequando-os para eventos temporais.

De maneira semelhante à MD, o objetivo da MDT é descobrir relações ocultas entre as sequências e subsequências de eventos. Para isso, de acordo com o objetivo da análise a ser realizada, uma ou mais tarefas de MDT podem ser empregadas. Dentre essas tarefas, podem ser destacadas (Keogh and Kasetty, 2003; Lin et al., 2004):

Pré-Processamento: Objetiva preparar os dados de maneira a facilitar as demais tarefas;

Indexação (Recuperação de Conteúdo): Busca por ST semelhantes em uma base de dados de ST;

Agrupamento: Busca identificar agrupamentos de ST, baseado na natureza dos dados, utilizando-se de alguma medida de similaridade;

Classificação: Objetiva, tomando como base as características de uma ST não-rotulada, classificar essa ST em uma das classes pré-definidas;

Detecção de Anomalias: Procura detectar padrões raros ou pouco frequentes no conjunto de dados;

Descoberta de *Motifs*: Visa a descoberta de padrões similares repetidos durante a série;

Previsão: Objetiva estimar dados futuros da série.

2.5.1 Pré-Processamento

O pré-processamento de ST, do mesmo modo que o pré-processamento de dados que não possuem relação com o tempo, visa preparar os dados para auxiliar na qualidade da análise e melhorar o desempenho computacional. Entretanto, ao se manipular dados temporais, deve-se utilizar métodos que forneçam uma atenção especial para a dimensão temporal, visando preservá-la (Morchen, 2006).

Dentre os fatores que devem receber atenção na etapa de pré-processamento de uma ST, podem ser destacados (Pyle, 1999):

Valores Faltantes: A existência de valores faltantes é uma característica muito indesejável para uma ST. Os valores faltantes podem ser de dois tipos: valores faltantes do índice temporal, o que geralmente é tratado utilizando o próximo valor de índice, e valores faltantes do dado sendo observado. Este último é o mais complexo de ser tratado, pois, em alguns casos, podem estar faltando séries de valores. Uma possível forma de tratar os valores faltantes é através de técnicas de interpolação ou auto-regressão, de maneira a preencher esses valores faltantes tomando como base algum padrão existente na série. Deve-se atentar que a inclusão desses valores pode ressaltar certas tendências nos dados, o que pode ressaltar padrões existentes, sejam eles verdadeiros ou não, de maneira equivocada. Esse realce de característica dos dados pode ser contornado adicionando-se ruído aos valores faltantes sendo preenchidos;

***Outliers*:** São valores que fogem do limiar padrão dos dados. Podem aparecer isoladamente ou em grupos dentro da série. Nem sempre representam dados errôneos e por isso devem ser analisados com cautela. Quando representarem dados incorretos, devem ser tratados de maneira a não prejudicar a evolução da série. Uma maneira de tratá-los é substituir seu valor por outro, utilizando abordagem semelhante à usada em casos de dados faltantes;

Amostragem Irregular: Apesar da maioria das vezes as ST possuírem suas mensurações realizadas em intervalos idênticos de amostragem, ou seja, possuem o eixo temporal equiespaçado, isso pode não ocorrer em todos os casos. Certos métodos de análise de ST lidam apenas com séries equiespaçadas, logo, se faz necessária a transformação da série ajustando os valores para que os mesmos reflitam os valores que deveriam ser observados, caso a série apresentasse um intervalo de amostragem regular;

Tendência: A maior parte dos modelos possuem dificuldades em considerar o componente de tendência, portanto, quando utilizar um desses modelos, a tendência deve ser removida da série.

O pré-processamento de ST realiza operações de maneira a tornar mais adequada a representação da série de acordo com as características dos métodos empregados nas etapas seguin-

tes. A remoção de ruídos e de *outliers*, o preenchimento de valores faltantes e a remoção de tendências estão entre as possíveis operações a serem realizadas nessa tarefa (Morchen, 2006).

2.5.2 Recuperação de Conteúdo

O tema de recuperação de conteúdo em ST tem recebido uma crescente atenção. A recuperação de conteúdo, também chamada consulta por exemplo (*query by example*), consiste em encontrar as séries (ou subsequências) que possuem uma grande similaridade com um determinado exemplo que está sendo consultado.

A recuperação por conteúdo em ST pode ser divida em duas categorias (Chen and Ozsu, 2003):

***Pattern Existence Queries*:** Nessa categoria buscam-se ST que possuem certo padrão nos dados;

***Exact Match Queries*:** Nessa categoria buscam-se ST através da especificação exata de valores a serem encontrados, além de fornecer informações temporais detalhadas do que se está procurando.

Armazenar e realizar consultas de similaridades nas séries, em seu formato original, é custoso computacionalmente, principalmente tratando-se de sistemas interativos, além de não destacar características chaves dos dados. Logo, buscando a diminuição no tempo de comparação da consulta, os dados originais podem ser representados em mais alto nível, utilizando, por exemplo, métodos como *Piecewise Aggregate Approximation* (Keogh and Pazzani, 1999).

Além do método de representação da ST, outro fator importante para a recuperação de conteúdo em ST é o método de busca. Ao lidar com ST, o método de busca deve levar em consideração alguns pontos específicos do domínio de ST, os quais também podem ser tratados na etapa de pré-processamento, como (Hetland, 2004):

- O intervalo de valores geralmente não é finito, ou mesmo discreto;
- A taxa de amostragem pode não ser constante;
- Deve permitir medidas de similaridade bastante flexíveis, visando melhores resultados em séries com ruídos.

Entre as possíveis aplicações da recuperação por conteúdo, podem ser citadas: identificação de companhias com padrões de crescimento similares, identificação de padrões musicais similares em bases de direitos autorais, encontrar porções de ondas sísmicas não-similares, a fim de verificar irregularidades geológicas, entre outras (Hetland, 2004).

2.5.3 Agrupamento

As tarefas de agrupamento (*clustering*) consistem na separação de um determinado conjunto de objetos em grupos, baseado em sua similaridade com os demais objetos do grupo. Objetos que apresentam maior similaridade encontram-se em um mesmo grupo e objetos que apresentam maior dissimilaridade encontram-se em grupos separados. Verifica-se então que o agrupamento visa identificar grupos de afinidades baseado em suas características (Rodrigues et al., 2007).

As abordagens de agrupamento de séries temporais podem ser divididas em três categorias principais (Morchen, 2006):

***Whole Series Clustering*:** Nessa abordagem, realiza-se o agrupamento de ST numéricas baseado em determinada medida de similaridade, utilizando algoritmos de agrupamentos conhecidos, sem que estes sejam específicos para tratar dados temporais;

***Sub-series Clustering*:** Nessa abordagem, realiza-se a construção de um conjunto de ST tomando como base a extração de segmentos de uma ST mais longa, ou seja, a divisão de uma ST extensa em séries de menor comprimento;

***Time Point Clustering*:** Nessa categoria enquadra-se o processo de agrupar os pontos de uma ST baseado na combinação de proximidade temporal e a similaridade dos valores correspondentes a estes pontos.

Na tarefa de agrupamento de dados não há classes pré-definidas, consistindo então, em um processo não-supervisionado. Validar a qualidade do particionamento dos dados é uma tarefa muito complexa, dificultando o processo de avaliação do particionamento. Independentemente da medida de avaliação adotada, essa pode apenas ser usada como um indicador da qualidade, sendo que a correta avaliação poderá apenas ser feita por especialistas do domínio (Rodrigues, 2008).

O procedimento de avaliar o resultado de um agrupamento é conhecido como validação de agrupamento (*cluster validity*), o qual pode ser classificado em três abordagens (Theodoridis and Koutroumbas, 2009):

Validação Interna: Geralmente essa abordagem de avaliação consiste em verificar os dados, baseado em medidas envolvendo os dados propriamente ditos, ou seja, as ligações internas entre os dados;

Validação Externa: Essa abordagem de avaliação é baseada em estruturas pré-definidas que refletem o conhecimento já existente (ou até mesmo intuição) do conjunto de dados;

Validação Relativa: Nessa abordagem ocorrem comparações dos agrupamentos encontrados com agrupamentos resultantes de outras técnicas de agrupamento. Essas técnicas podem incluir a utilização do mesmo algoritmo de agrupamento com diferentes configurações, ou até mesmo outros algoritmos.

A tarefa de agrupamento de ST pode ser aplicada, por exemplo, para agrupar padrões de tráfego em comunicação de dados, para o reconhecimento de assinaturas e para a análise de padrões de consumo de energia elétrica (Morchen, 2006; Rodrigues et al., 2007).

2.5.4 Classificação de Dados Temporais

A tarefa de classificação é definida com um caso especial de aprendizagem de máquina. Consiste em, tomando por base um conjunto de exemplos composto por vários atributos, sendo que esses exemplos possuem classes conhecidas, classificar exemplos desconhecidos baseado nos valores de seus atributos (Larose, 2005). A maioria dos métodos de classificação assumem dados históricos, ou seja, assumem que os exemplos envolvidos na construção do modelo, são os melhores estimadores das classificações futuras. Assim, características temporais destes exemplos podem ser um dos fatores que auxiliam na eficiência do processo de classificação (Last et al., 2004).

A tarefa de classificação em ST pode ser dividida em dois grupos principais (Morchen, 2006):

Classificação de Séries Temporais: Classificam-se as séries como um todo, atribuindo-se uma etiqueta para cada série de treino. Muitas abordagens, nesta categoria, baseiam-se na combinação de extração de características temporais e métodos de classificação convencionais;

Classificação de Pontos Temporais: Classificam-se pontos das séries. Para isso são atribuídas etiquetas para cada ponto e o treinamento utiliza-se destas etiquetas e dos valores dos pontos das séries. Não há a necessidade de se etiquetar todos os pontos, apenas aqueles que apresentam interesse para a detecção de eventos a serem analisados.

A classificação de ST pode ser aplicada em diversas situações, tais como a classificação de batimentos cardíacos e a indexação de documentos manuscritos (Wei and Keogh, 2006).

2.5.5 Detecção de Anomalias

Detecção de anomalias, também chamada de detecção de novidades ou detecção de raridades, pode ser entendida como a detecção automática de fenômenos anormais ou imprevistos

se comparados com um grande conjunto de dados ditos como normais (Ma and Perkins, 2003).

Um aspecto que deve ser considerado quando se discute raridade, é a quantidade de dados sendo observados. Por exemplo, em um conjunto de dados pequeno, no qual um determinado caso cobre um por cento (1%) dos dados, este pode ser considerado raro. Em uma base na qual existam dez milhões de exemplos, o caso que cobre um por cento dos dados talvez possa não ser mais considerado raro (Weiss, 2004).

Os métodos para detecção de anomalias podem ser classificados em dois grupos principais (Rebbapragada et al., 2008):

Série Temporal Única: Neste método analisa-se a ST de maneira isolada, como se fosse única. O objetivo desta abordagem é encontrar sub-regiões anômalas. Em alguns casos, é possível converter uma única ST em uma base de ST através da utilização de janelas deslizantes;

Base de Séries Temporais: Nesse método analisa-se múltiplas séries, buscando-se exemplos anômalos.

A detecção de anomalias pode ser utilizada para diversos fins, tais como a identificação de transações fraudulentas, detecção e predição de falhas de equipamentos, detecção de vazamento de óleo por imagens de satélite, entre outros (Weiss, 2004).

2.5.6 Descoberta de *Motifs*

Motifs em ST podem ser definidos como padrões similares repetidos durante a série. O objetivo da descoberta de *motifs* é encontrar padrões de interesse previamente desconhecidos em uma ST (Yankov et al., 2007).

A descoberta de *motifs* é realizada por meio de buscas em uma ST objetivando encontrar casamento de padrões não-triviais. Tais casamentos não-triviais consistem em descobrir padrões que possuam uma distância mínima entre eles, de maneira a evitar que padrões com distâncias muito próximas sejam encontrados, dessa maneira garantindo que os padrões sejam mutuamente exclusivos (Lin et al., 2002).

Uma preocupação adicional da descoberta de *motifs* é tratar situações em que a ST possua ruído, pois eles acabam por dificultar o processo de determinação dos padrões a serem encontrados. Além dessa, há ainda a preocupação em se detectar falsos *motifs*, que podem ocorrer, por exemplo, em situações em que a série possui picos altos e baixos, favorecendo então encontrar segmentos de retas nestes picos (Chiu et al., 2003).

A descoberta desses padrões permite então uma descrição estrutural da ST. A análise

de *motifs* pode ser aplicada a diversas áreas da MD, como descoberta de regras, detecção de novidades e agrupamentos (Yankov et al., 2007).

2.5.7 Previsão

A previsão em ST busca, através da exploração de dados passados, projetar os dados futuros. Tanto os valores do passado da série quanto o valor atual da série são usados buscando-se estimar os valores futuros (Sorjamaa et al., 2007). A previsão de um momento Z_{m+1} de uma série $Z = (Z_1, Z_2, \ldots, Z_m)$ pode ser descrita pela seguinte equação (Lütkepohl, 2005):

$$Z_{m+1} = f(Z_m, Z_{m-1}, Z_{m-2}, \ldots) \tag{2.4}$$

onde $f(.)$ representa uma função de previsão que se utiliza dos valores passados Z_m, Z_{m-1}, $Z_{m-2}, \ldots$ para realizar a estimação.

De modo geral, as abordagens de previsão podem ser classificadas em paramétricas e não-paramétricas. Os métodos paramétricos assumem que os dados respeitam alguma distribuição e que podem ser modelados a partir de um conjunto de parâmetros. Entre os principais modelos paramétricos podem ser citados: Auto-regressivos (AR), Médias Móveis (MA), Auto-regressivos de Médias Móveis (ARMA), Auto-regressivos de Médias Móveis Integrados (ARIMA) e Auto-regressivos de Médias Móveis Integrados Sazonais (SARIMA) (Morettin and Toloi, 2006). Os métodos não-paramétricos não definem parâmetros de uma distribuição específica e têm a capacidade de se adaptar a diferentes comportamentos ao longo do tempo. Entre os métodos propostos encontram-se os baseados nos conceitos de Redes Neuronais Artificiais (RNA) e variações do algoritmo dos vizinhos mais próximos ou *k-Nearest Neighbor* (*kNN*) (McNames, 1999; Ferrero et al., 2009).

A aproximação de métodos de aprendizagem de máquina e da mineração de séries temporais tem proporcionado o desenvolvimento de novas abordagens para a previsão de ST. A tarefa de previsão de dados temporais, bem como alguns métodos para a realização dessa tarefa, são descritos em maiores detalhes no Capítulo 3.

2.6 Aplicações

Diversos eventos de diferentes áreas podem ser representados através de ST, bem como diversas técnicas focadas em áreas de conhecimento específicas foram desenvolvidas. No campo industrial, há um crescimento no interesse pela área de ST, visando aperfeiçoar a manutenção de equipamentos que sofrem desgastes, objetivando obter o máximo da produção e do equipa-

mento, mantendo estável ou até diminuindo os custos de produção. Podem ser adicionados sensores a equipamentos visando monitorar seu estado de funcionamento durante o tempo, transformando assim essa informação em uma ST. Uma vez obtida a ST, e, conhecendo-se o modelo de falha (ou desenvolvendo-se o mesmo através de observações), podem ser utilizados métodos de previsão de ST para predizer a necessidade de manutenção da máquina. Seguindo modelo semelhante pode-se verificar se a máquina apresenta falhas (Xie and Ho, 1999).

Existem também aplicações dentro da área de astronomia. Em Hanslmeier et al. (2004) pode ser observada a utilização de análise de ST para o estudo da dinâmica de granulação solar através da obtenção de espectros 2-D. Segundo os autores, pela primeira vez foi explicitamente mostrada como a evolução em um campo selecionado da fotosfera influenciava a evolução da estrutura granular/intergranular.

Com relação à área de estudos ambientais, as ST podem ser de utilidade para verificação de estados, de composição e de funcionamento de ecossistemas. É possível com ST verificar dados de oceanografia, como processos climáticos que causem variabilidades, dinâmicas da cadeia alimentar, depósitos de carbono, entre outros (Ducklow et al., 2009). Podem ser utilizadas também para verificar a dinâmica de comportamento das espécies, visando detectar quais espécies seguem tendências de comportamento similares e quais divergem do comportamento mais comum (Nye et al., 2009).

Na área de saúde, verifica-se uma grande inserção da análise de ST. No trabalho de Layte et al. (2010), ST foram analisadas em busca das proporções de diminuição da mortalidade por causas circulatórias, bem como os motivos dessa diminuição, durante os anos de 1995 e 2005. A análise de séries provenientes de eletrocardiogramas, séries que refletem os mecanismos de controle fisiológico da frequência cardíaca, podem trazer diversas informações sobre o estado do coração, como verificar a possível existência de isquemia miocárdica (Cammarota and Curione, 2008).

2.7 Considerações Finais

As ST estão presentes em diversas aplicações, estando várias vezes relacionadas ao processo de tomada de decisões. Isso desperta cada vez mais o interesse sobre esse tema, demonstrando sua importância. Neste capítulo foram apresentados definições e conceitos de séries temporais e seus componentes. Foram apresentadas também tarefas de análise e mineração de ST, com exemplos de algumas de suas aplicações. No próximo capítulo, a tarefa de previsão de dados temporais, foco deste trabalho, é apresentada em maiores detalhes. Essa é uma tarefa de interesse em diversas áreas, pois busca estimar valores de interesse futuros, através da análise de dados do passado.

Capítulo 3

Previsão de Séries Temporais

3.1 Considerações Iniciais

Um dos principais objetivos da construção de modelos utilizando ST é a possibilidade de estimar valores da série para instantes de tempos futuros (Cryer and Chan, 2008). A previsão de ST é importante para o auxílio à tomada de decisões, entretanto, ela não constitui um fim, apenas fornece um conjunto de informações para a tomada de decisão (Morettin and Toloi, 2006).

As aplicações que utilizam previsões de séries temporais podem ser as mais variadas. Previsões a curto prazo podem ser usadas para monitoração de processos, antecipando falhas de equipamentos ou monitoração da qualidade de produção de determinado produto. Previsões a longo prazo podem auxiliar a tomada de decisões de planejamento urbano, para crescimento populacional ou análise de modificações climáticas.

Existem diversos métodos para a realização dessas previsões. Neste capítulo são apresentados alguns desses métodos, bem como alguns conceitos de Aprendizagem de Máquina (AM) necessários para a compreensão do método de previsão utilizado neste trabalho.

3.2 Métodos para Previsão de Séries Temporais

Existem diversos métodos para previsão de ST, que utilizam desde complexos modelos estatísticos a modelos intuitivos e simples, cada um apresentando suas próprias capacidades e limitações. Uma mesma série pode ser analisada e prevista por vários desses métodos. Assim, para uma melhor seleção do método de previsão a ser empregado, é necessário não somente se ter conhecimento do comportamento do fenômeno observado, mas também da natureza e do objetivo da análise e do método utilizado (Mueller, 1996).

Uma das possíveis formas de classificação das abordagens de previsão é a divisão entre métodos paramétricos e não-paramétricos (Morettin and Toloi, 2006; Aguirre, 2007). A seguir

são descritos alguns dos principais modelos de ambas as categorias.

3.2.1 Métodos Paramétricos

Os métodos que se utilizam de modelos paramétricos dependem explicitamente de um conjunto de parâmetros finitos. Esses parâmetros devem ser encontrados de maneira a otimizar os resultados da previsão. Entre os principais modelos paramétricos encontrados na literatura podem ser destacados: Auto-regressivos (AR), Médias Móveis (MA), Auto-regressivos de Médias Móveis (ARMA), Auto-regressivos de Médias Móveis Integrados (ARIMA) e Auto-regressivos de Médias Móveis Integrados Sazonais (SARIMA) (Mueller, 1996; Morettin and Toloi, 2006; Aguirre, 2007; Cryer and Chan, 2008).

Auto-regressivos (AR)

Os modelos auto-regressivos buscam estimar valores futuros levando em consideração que um determinado valor da série é resultante de uma combinação linear de seus valores passados, acrescidos de um fator "inovação" (também chamado por alguns autores de ruído branco). Esse fator "inovação" representa qualquer fator desconhecido que não pode ser explicado pelos valores do passado da série. Assim, o modelo AR de ordem h, AR(h), deve satisfazer a Equação 3.1 (Cryer and Chan, 2008):

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_h z_{t-h} + e_t \tag{3.1}$$

onde ϕ representa o valor de ponderação do valor observado; h representa o número de observações consideradas; e e_t representa o fator "inovação". Assume-se também, que para cada instante t do tempo, e_t é independente dos valores passados da série $(z_{t-1}, z_{t-2}, \ldots)$.

Pode-se reduzir a notação da Equação 3.1 utilizando-se o operador de defasagem (também chamado de translação ou *backshift*) B: $Bz_t = z_t - 1, B^m z_t = z_{t-m}$. Assim, reescrevendo a Equação 3.1 utilizando o operador de defasagem, obtém-se a Equação3.2 (Morettin and Toloi, 2006):

$$\phi(B) z_t = e_t \tag{3.2}$$

onde $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_h B^h$.

Médias Móveis (MA)

Os modelos de médias móveis procuram estimar valores levando em consideração que cada valor futuro é uma combinação linear dos valores $e_t, e_{t-1}, e_{t-2}, \ldots, e_{t-q}$. Assim, o modelo MA de ordem q, MA(q), é definido pela Equação 3.3 (Mueller, 1996):

$$z_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} \tag{3.3}$$

onde θ representa os pesos de ponderação aplicados; e q a quantidade de valores considerados. A Equação 3.3 assemelha-se a Equação 3.1, observando-se a diferença que o valor estimado depende dos valores e_t observados em cada período passado, ao invés das observações da série propriamente ditas.

O funcionamento do método consiste em aplicar os pesos 1, $-\theta_1, -\theta_2, \ldots, -\theta_q$ para as variáveis de inovação $e_t, e_{t-1}, e_{t-2}, \ldots, e_{t-q}$. Quando se deseja conhecer o valor futuro z_{t+1}, "move-se" os pesos considerados, aplicando-os nas variáveis $e_{t+1}, e_t, e_{t-1}, \ldots, e_{t-q+1}$.

Pode-se novamente reduzir a notação, utilizando-se o operador de defasagem, de maneira semelhante a Equação 3.2. Desse modo, obtém-se a Equação 3.4 (Aguirre, 2007):

$$z_t = \theta(B) e_t \tag{3.4}$$

onde $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$.

Auto-regressivos de Médias Móveis (ARMA)

Os modelos auto-regressivos de médias móveis são uma combinação dos modelos AR e MA. Dessa maneira, esses modelos buscam estimar um valor considerando que a série é descrita em parte por um processo auto-regressivo, e em parte por um processo envolvendo média móvel. O modelo ARMA de ordem h, q, ARMA(h, q), pode ser descrito pela Equação 3.5 (Cryer and Chan, 2008):

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_h z_{t-h} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} \tag{3.5}$$

Esse modelo considera tanto os valores do passado da série quanto os valores dos elementos "inovativos" da série para estimar o valor futuro. Uma das vantagens da utilização dos modelos ARMA, é que muitas séries reais podem ser descritas e ter seus valores estimados por esse modelo considerando poucos parâmetros.

A notação reduzida do modelo ARMA(h, q), utilizando o operador de defasagem, é apre-

sentada na Equação3.6 (Aguirre, 2007):

$$\phi(B)z_t = \theta(B)e_t \tag{3.6}$$

Auto-regressivos de Médias Móveis Integrados (ARIMA)

Os modelos auto-regressivos de médias móveis integrados são os modelos com aplicações mais abrangentes dentre os mencionados até o momento. Suas propriedades permitem, em teoria, manipular séries temporais de qualquer natureza. O diferencial desse modelo para os anteriormente citados é o fato dele estar preparado para lidar com ST não-estacionárias. O modelo ARIMA de ordem (h,d,q), ARIMA(h,d,q), é descrito pela Equação 3.7 (Mueller, 1996):

$$z_t = \phi_1 M_{t-1} + \phi_2 + M_{t-2} + \ldots + \phi_h M_{t-h} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} \tag{3.7}$$

onde $M_t = z_t - z_{t-d}$; d representa o grau do operador de diferença; ϕ_h e θ_q são os parâmetros do processos auto-regressivos e de médias móveis, possuindo ordem h e q respectivamente (ARMA(h,q)); e e_t corresponde ao elemento "inovativo" que não pode ser explicado pelo modelo.

Para a utilização desse modelo, supõe-se que a d-ésima diferença entre as observações da ST pode ser representada por um processo estacionário capaz de ser estimado por um modelo ARMA. Desse modo, séries que apresentam tendência não-explosiva, ou seja, não-estacionaridade homogênea, bem como séries estacionárias podem ser descritas por esse modelo.

A notação compacta do modelo ARIMA(h,d,q), é apresentada na Equação3.8 (Morettin and Toloi, 2006):

$$\phi(B)M_t = \theta(B)e_t \tag{3.8}$$

Auto-regressivos de Médias Móveis Integrados Sazonais (SARIMA)

Os modelos auto-regressivos de médias móveis integrados sazonais são uma variação do modelo ARIMA que ampliam sua abrangência, permitindo a utilização em séries que apresentam sazonalidade. Esse modelo é importante, devido ao fato de um grande número de ST reais possuir algum componente sazonal, repetido a cada s observações. O exemplo mais comum dessa característica são os dados mensais, onde, em geral, é possível encontrar uma dependência entre z_t e z_{t-12}. O modelo SARIMA de ordem $(h,d,q)\times(H,D,Q)_s$, SARIMA$(h,d,q)\times(H,D,Q)_s$,

é descrito em sua forma compacta[1] pela Equação 3.9 (Morettin and Toloi, 2006; Aguirre, 2007):

$$\phi(B)\Phi(B^s)(1-B^s)^D(1-B)^d z_t = \theta(B)\Theta(B^s)e_t \tag{3.9}$$

onde, $\Phi(B^s)$ representam os coeficientes sazonais do processo auto-regressivo ($\Phi(B^s) = 1 - \Phi_1 B^s - \ldots - \Phi_H B^{sH}$); $\Theta(B^S)$ representam os coeficientes sazonais das médias móveis ($\Theta(B^s) = 1 - \Theta_1 B^s - \ldots - \Theta_Q B^{sQ}$); e $(1-B^s)^D$ é o operador da diferença de ordem D para a diferenciação sazonal da série.

Assim, o modelo SARIMA torna-se mais adequado em situações onde os dados possuem variações sazonais que não são adequadamente tratados pela primeira diferença ($z_t - z_{t-1}$). Nesses casos, esse modelo é vantajoso por diferenciar os dados no período sazonal.

3.2.2 Métodos Não-Paramétricos

Os métodos que se utilizam de modelos não-paramétricos, em contrapartida, não dependem explicitamente de parâmetros para modelar o comportamento do sistema. Dessa maneira podem ser mais facilmente adaptados a diferentes conjuntos de dados ou ainda à variação de um mesmo conjunto. Dentre os métodos não-paramétricos existentes na literatura, podem ser destacados os baseados nos conceitos de Redes Neuronais Artificiais e variações do algoritmo *k-Nearest Neighbor* (Mitchell, 1997; Pyle, 1999; Haykin, 1999; Russel and Norvig, 2004; Han and Kamber, 2006; Ferrero, 2009).

Rede Neuronal Artificial (RNA)

Sistemas biológicos possuem a capacidade de realizar tarefas extremamente complexas de maneira eficiente. Esse fato acabou por inspirar abordagens que visam simular sistemas biológicos. A Rede Neuronal Artificial é uma dessas abordagens. Ela consiste em simular o funcionamento do cérebro humano em computadores (Haykin, 1999; Berthold and Hand, 2003).

As RNA possuem uma grande quantidade de unidades de processamento simples, chamadas neurônios. Cada uma dessas unidades possui uma grande quantidade de conexões com outros neurônios, sendo que essas conexões recebem pesos de ponderação e podem estar dispostas em mais de uma camada. A configuração das interações entre os neurônios e seus pesos define o modelo a ser representado (Haykin, 1999).

A disposição dessas ligações, do número de neurônios e das camadas é chamada topologia. A escolha da melhor topologia é dependente do problema em análise, não havendo então

[1] Apenas a forma compacta foi apresentada de maneira a simplificar a notação.

uma topologia fixa que possa ser considerada a melhor. De maneira geral, a topologia pode ser composta por três componentes principais (Witten and Frank, 2005; Han and Kamber, 2006):

Camada de Entrada: Essa camada é responsável por receber os dados de entrada e transmiti-los para as camadas mais internas (camadas escondidas);

Camada Escondida: Nessa camada está disposta a maior parte dos neurônios responsáveis por modelar a série. Aqui ocorrem os cálculos e as ponderações necessários para a previsão, e em seguida esses resultados são transmitidos à próxima camada, que pode ser outra camada escondida ou a camada de saída. Pode haver uma ou várias camadas escondidas. A existência de camadas escondidas é uma das características das redes *multilayer perceptron*, de grande utilização atualmente. Modelos de RNA mais simples podem não conter essa camada;

Camada de Saída: Essa camada é responsável por retornar ao usuário o resultado dos processamentos.

As RNA necessitam de períodos de treinamento antes de estarem adequadas para utilização. Esse treinamento visa ajustar sua configuração de maneira a obter os resultados apropriados (Han and Kamber, 2006).

k-*Nearest Neighbor* (*kNN*)

O algoritmo *k-Nearest Neighbor* (*kNN*) é um algoritmo de aprendizagem supervisionada que consiste em encontrar, segundo alguma medida de similaridade, os k exemplos mais próximos de um exemplo ainda não-rotulado e, baseado nos rótulos desses k exemplos próximos, rotular o novo exemplo (Han and Kamber, 2006).

Desse modo, quando se utiliza, por exemplo, 1-*NN*, o novo exemplo recebe o mesmo rótulo da classe do vizinho mais próximo encontrado; caso sejam considerados $k > 1$ vizinhos, por exemplo, 5-*NN*, deve-se então definir como será determinado o rótulo do novo exemplo. Uma das abordagens mais simples consiste em utilizar a classe majoritária, ou seja, a classe predominante entre esses cinco exemplos. Outra possível abordagem é a utilização de pesos para cada um dos vizinhos próximos de acordo com algum critério, como a proximidade desse vizinho, ou seja, quanto mais próximo o vizinho maior a influência de sua classe na decisão. Assim, é possível perceber que a decisão de quantos vizinhos próximos devem ser considerados para a classificação, pode exercer influência no resultado do funcionamento do algoritmo. Esse valor é particular a cada problema, levando à necessidade de avaliação dos possíveis valores de k a serem considerados.

Alguns pontos principais devem ser levados em consideração quando se utiliza o algoritmo *kNN* (Larose, 2005):

- Definição da quantidade de vizinhos a ser considerada, ou seja, o valor de k;
- Definição da medida de similaridade de acordo com a característica dos exemplos;
- Definição de como serão combinadas as informações de mais de um exemplo.

Outro fator de importância consiste na quantidade de exemplos usados para treinamento. Esses exemplos influenciam diretamente no tempo dispendido para a execução do algoritmo, pois, para se encontrar os k vizinhos próximos, se faz necessária uma busca por todos os exemplos do conjunto de treinamento. Assim, torna-se ideal armazenar a menor quantidade de exemplos possíveis, armazenando apenas os mais representativos (Alpaydin, 2004).

O algoritmo *kNN* foi adaptado por Ferrero (2009) para utilização com dados temporais. A adaptação permite a utilização de ST como forma de entrada de dados, focando como tarefa de interesse a previsão de dados temporais. Esse algoritmo adaptado é descrito a seguir.

3.3 *k-Nearest Neighbor - Time Series Prediction* (*kNN-TSP*)

O algoritmo *k-Nearest Neighbor - Time Series Prediction* é derivado de uma série de conceitos de AM. De maneira a buscar um melhor entendimento sobre seu funcionamento e de sua classificação dentro da área de AM, alguns conceitos fundamentais dessa área são apresentados.

3.3.1 Conceitos de Aprendizagem de Máquina (AM)

A AM é uma área da inteligência artificial que tem como objetivo simular o processo de aprendizagem por meio do desenvolvimento de técnicas computacionais, bem como a construção de sistemas que possuem a capacidade de adquirir conhecimentos de maneira automática (Mitchell, 1997). Assim, a AM consiste em programar computadores para otimizar (melhorar com experiência) a sua resposta dada uma determinada situação, usando dados de exemplos ou situações passadas (Alpaydin, 2004).

A AM pode ser classificada de uma maneira ampla, de acordo com a presença de rótulos nos exemplos, em duas categorias principais (Russel and Norvig, 2004):

Aprendizagem Supervisionada: Envolve as situações em que se têm exemplos de treinamento com informações de entrada e de saída, ou seja, o rótulo de classificação do exemplo

é conhecido. Esse rótulo pode ser fornecido tanto por um instrutor supervisionando a aprendizagem, quanto estar contido no próprio exemplo, como um de seus atributos;

Aprendizagem Não-supervisionada: Envolve as situações em que se têm exemplos de treinamento com as informações de entrada, porém, não apresentam as informações de saída. Dessa forma, os exemplos não apresentam os rótulos de sua classificação.

Além da classificação de acordo com a presença dos rótulos nos exemplos, pode-se classificar a AM de acordo com seu paradigma de aprendizagem (Rezende, 2005):

Simbólico: Nesse paradigma, a aprendizagem consiste em utilizar-se da análise de exemplos e contra-exemplos de um determinado conceito, representando o conceito embutido de maneira simbólica. Exemplos desse paradigma incluem árvores de decisão e regras de decisão;

Estatístico: Nesse paradigma, a aprendizagem ocorre por meio do desenvolvimento de modelos estatísticos que buscam encontrar a aproximação de um conceito. Um dos exemplos mais expressivos desse paradigma é a aprendizagem Bayesiana;

Baseado em Exemplos: Nesse paradigma, a aprendizagem ocorre por meio da utilização de exemplos previamente conhecidos. Quando é fornecido ao sistema um novo exemplo, o sistema busca, entre seus exemplos conhecidos, aqueles que mais se assemelham com o fornecido, e baseado nesses exemplos conhecidos classifica o novo exemplo. O algoritmo de vizinhos mais próximos (*Nearest Neighbors*) pertence a esse paradigma;

Conexionista: Nesse paradigma, a aprendizagem consiste na utilização de unidades de processamento altamente conectadas, onde cada unidade realiza individualmente pequenas porções de processamento. Exemplos desse paradigma consistem nas Redes Neuronais Artificiais;

O algoritmo *kNN-TSP* pertence à categoria de algoritmos Baseados em Exemplos, a qual é descrita a seguir.

Algoritmos Baseados em Exemplos

Algoritmos Baseados em Exemplos (também chamados de *Instance Based*) são caracterizados por não construírem um modelo geral que descreve explicitamente o conjunto de dados de treinamento. Esse modelo é construído pelo simples armazenamento desse conjunto de dados. A generalização sobre o conjunto de dados é realizada a cada momento em que é solicitado ao algoritmo uma nova classificação ou previsão (Mitchell, 1997).

No momento em que o algoritmo recebe uma solicitação de classificação ou de previsão para um novo exemplo, um conjunto de exemplos similares é recuperado e utilizado para a realização da tarefa. Essa abordagem resulta na criação de diferentes funções de aproximação local, uma para cada consulta realizada, tornando-a vantajosa em casos que a função que descreve os dados é muito complexa e pode ser descrita por uma coleção de funções menos complexas (Mitchell, 1997; Han and Kamber, 2006).

Algoritmos baseados em casos comumente não apresentam conhecimento explícito. Porém, realizando generalização de exemplares, as regras produzidas por esses algoritmos podem ser comparadas com outros métodos de aprendizagem. Devido a sua característica incremental, é possível a realização de consultas ao algoritmo durante a construção de sua base de exemplos, porém, caso isso seja feito, deve ser realizado com cautela, pois nessa fase a exatidão do algoritmo em descrever o conjunto de dados pode estar reduzida (Witten and Frank, 2005).

Uma desvantagem dos algoritmos baseados em casos que pode ser destacada é a possibilidade do custo de classificação ou previsão ser elevado. Esse possível custo elevado deve-se ao fato de que praticamente toda a computação necessária pelo algoritmo ocorre apenas no momento em que é solicitada a classificação ou previsão. Devido a essa característica, técnicas de armazenamento e de implementação de paralelismo eficientes podem ser de grande auxílio para um bom desempenho computacional (Mitchell, 1997; Han and Kamber, 2006).

Um dos algoritmos Baseados em Exemplos mais conhecidos é o *k-Nearest Neighbor*, que foi brevemente descrito na Seção 3.2.2.

3.3.2 Descrição do Algoritmo *kNN-TSP*

Como mencionado, o algoritmo *kNN-TSP* é uma adaptação do algoritmo *kNN* (Ferrero, 2009). Ele busca estimar um valor z_{t+1} de uma ST Z, utilizando os valores anteriores dessa série, ou seja, $z_t, z_{t-1}, z_{t-2}, \ldots, z_{t-m+1}$ onde m corresponde à quantidade de valores do passado da série a serem considerados.

A ideia central do funcionamento do algoritmo consiste em calcular um valor futuro $\hat{z}_{t+1}$, sendo esse valor uma aproximação do desconhecido valor verdadeiro. Para esse cálculo se utilizam as k sequências de tamanho w mais próximas ao final da série, sendo essa sequência final denominada de sequência de referência. Essas sequências próximas são selecionadas através de alguma medida de similaridade.

Na Figura 3.1 é ilustrado de maneira esquemática o funcionamento do algoritmo *kNN-TSP* em forma de blocos, onde cada bloco representa:

1. Os dados de entrada para o algoritmo consistem na ST que terá seus valores estimados, o

tamanho da janela (w), a função de previsão utilizada para cálculo dos valores futuros e a quantidade de vizinhos próximos a serem considerados (k);

2. Nesse passo do algoritmo é quantificada a similaridade, de acordo com alguma medida de similaridade, entre os w pontos finais da ST e todas as sequências candidatas a vizinhos próximos. Uma janela deslizante percorre todas as possíveis sequências da série;

3. Nesse momento o algoritmo já possui a similaridade de todas as sequências armazenadas. Essas sequências são então ordenadas de acordo com a sua similaridade, sendo que as k sequências mais similares são selecionadas;

4. Nesse passo é realizado o cálculo do valor futuro estimado. A função de previsão recebe como dados de entrada as k sequências selecionadas no passo anterior e então realiza o cálculo do valor futuro;

5. Os dados de saída retornados pelo algoritmo consistem nos valores futuros estimados para a ST.

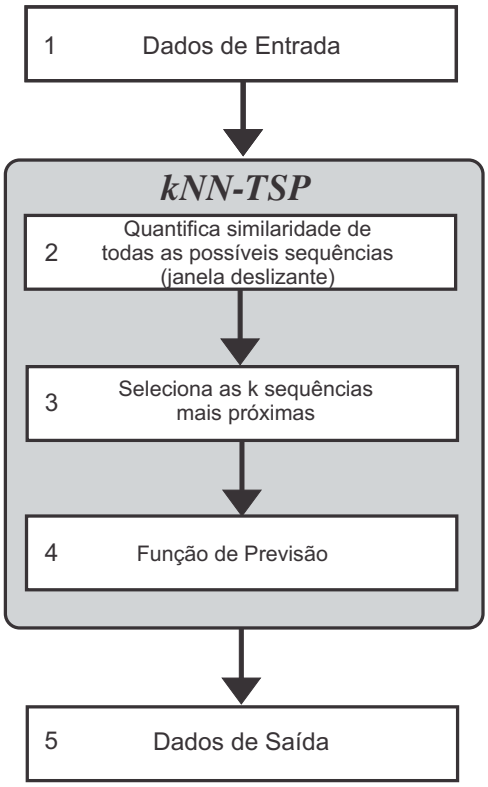


Figura 3.1: Esquema de funcionamento do *kNN-TSP*.

No Algoritmo 1 (Ferrero, 2009) é apresentado o pseudocódigo, em alto nível, do algoritmo *kNN-TSP*, onde:

- Z representa a ST utilizada;

- w representa o tamanho da janela para busca das sequências;
- M_s representa a medida de similaridade utilizada;
- C_k representa o critério utilizado para a seleção dos vizinhos próximos;
- k representa a quantidade de vizinhos mais próximos; e
- f representa a função de previsão utilizada para o cálculo do valor futuro.

Algoritmo 1: *kNN-TSP*.

```
   Entrada: Z, w, M_s, C_k, k, f
   Saída: ẑ_{n+1};
1  início
2      // Construção do conjunto de séries de treinamento S a partir da série temporal Z
3      // e tamanho de janela w
4      S ← series_de_treinamento(Z, w);
5      // Definição da sequência de referência U
6      U ← (z_n);
7      // Obtenção das k sequências mais próximas a U contidas em S, considerando a
8      // medida de similaridade M_s e o critério de seleção de vizinhos próximos C_k
9      S' ← vizinhos_proximos(S, U, M_s, C_k, k);
10     // Cálculo do valor futuro da sequências de referências, utilizando f(S')
11     ẑ_{n+1} ← f(S');
12     retorna ẑ_{n+1}
13 fim
```

Na linha 4 é atribuída à variável S o conjunto de treinamento no qual será realizada a busca pelos vizinhos mais próximos, isto é, a ST que terá seu valor previsto é particionada em sequências de tamanho w. Em seguida, na linha 6, é armazenada a sequência de referência, ou seja, a sequência que será utilizada como parâmetro para a busca por sequências similares no passado da série. Na linha 9, têm-se a busca pelos k vizinhos próximos no conjunto de treinamento S. Usa-se como referência de busca a sequência U, sendo verificada a similaridade através de uma medida M_s e um critério de seleção de vizinhos C_k. A função de previsão $f(.)$ realiza, na linha 11, o cálculo do valor futuro utilizando como entrada os vizinhos próximos armazenados em S'. Ao final, na linha 12, é retornado o valor futuro estimado pelo algoritmo.

Um exemplo da aplicação do algoritmo *kNN-TSP* para a previsão da ST de Mackey-Glass é apresentado na Figura 3.2. Nessa figura, a linha mais clara representa os valores reais observados na ST; a linha de cor preta, a sequência dos cinco últimos valores ocorridos; a linha de cor cinza escuro com asteriscos, as sequências similares encontradas pelo algoritmo; e o quadrado, o valor estimado calculado pela função de previsão. Nesse exemplo, deseja-se prever o último valor da série, dessa maneira o tamanho da janela selecionado foi cinco e, baseado na última sequência de cinco pontos e em uma medida de similaridade, foram selecionadas duas sequências similares. Baseado no valor do ponto futuro dessas duas sequências similares foi calculado o valor previsto para o final da série.

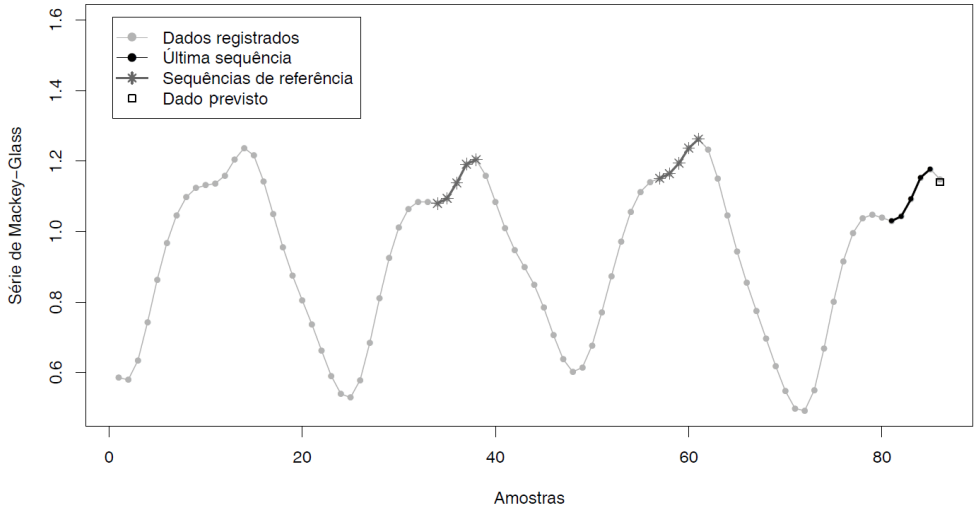


Figura 3.2: Exemplo da aplicação do algoritmo *kNN-TSP*.

Como pôde ser observado, a previsão com o algoritmo *kNN-TSP* depende de alguns parâmetros descritos a seguir e ilustrados pela Figura 3.3 (Ferrero et al., 2009):

(a) Tamanho *w* da janela para extrair as sequências: Refere-se ao tamanho das sequências a serem consideradas para o cálculo do valor futuro na ST. Deve-se dividir a série de entrada em um conjunto de sequências que irão constituir os exemplos de treinamento, sendo que esses exemplos possuem tamanho w. Isso possibilita analisar localmente o fenômeno que está sendo avaliado, permitindo assim previsões locais ou de curto prazo. O valor de w é dependente do problema em análise, pois fatores como o domínio dos dados e frequências de aquisição dos mesmos tem grande influência no formato dos dados;

(b) Conjunto de exemplos de treinamento: Consiste no conjunto de sequências pertencentes à ST a serem consideradas para constituir o conjunto de treinamento. O tempo e o espaço de armazenagem despendidos, nesse algoritmo, dependem do tamanho do conjunto de treinamento. Desse modo, é importante realizar uma seleção das sequências de treinamento de maneira a se armazenar apenas as sequências relevantes;

(c) Medida de similaridade: Refere-se à medida utilizada para quantificar a similaridade entre os exemplos. Quanto maior a similaridade, menor a distância entre as ST. Formas de quantificar a similaridade entre ST serão discutidas em maiores detalhes no Capítulo 4;

(d) Cardinalidade do conjunto de sequências similares: Refere-se à quantidade (k) de sequências mais próximas a serem consideradas para a previsão do valor futuro. Esse valor pode ser definido inicialmente como constante, por exemplo, através da análise de um especialista, ou de maneira experimental, avaliando o desempenho de diversos valores de k. Outra possível abordagem é a utilização de um limiar de similaridade, onde todas as sequências que obtiverem um valor maior que um determinado limiar são consideradas como sendo vizinhos mais próximos, não havendo então um valor definido para k;

(e) Função de previsão: Refere-se a função utilizada para determinar a maneira como serão considerados os valores das sequências mais próximas para estimar o valor futuro. Essa função pode ser então adaptada de acordo com o domínio em estudo, de maneira a considerar fatores específicos de cada situação.

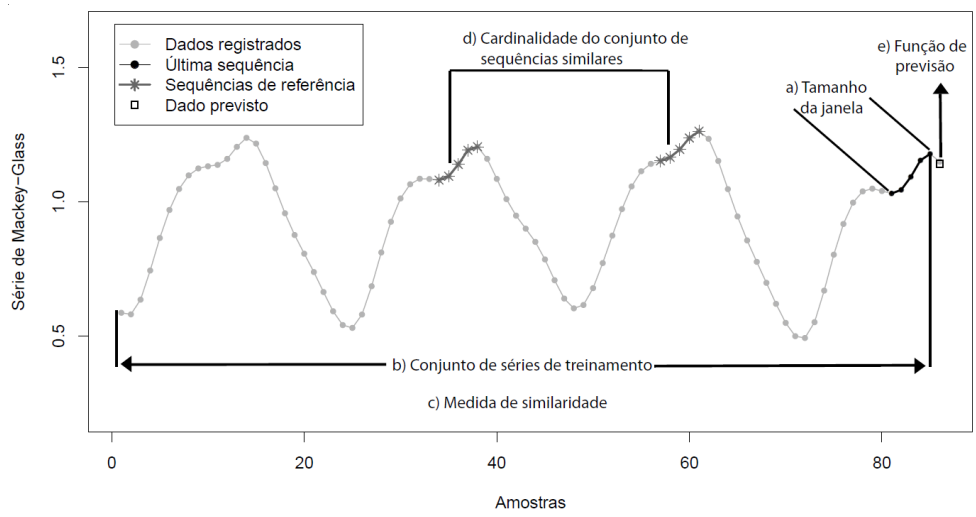


Figura 3.3: Parâmetros do algoritmo *kNN-TSP* (Modificado de Ferrero (2009)).

A escolha desses parâmetros influencia diretamente nos resultados da estimativa do valor futuro. A melhor configuração desse conjunto de parâmetros é particular a cada domínio e problema, não havendo então uma configuração ótima geral. Dessa forma, se faz necessária a investigação da influência de cada um desses parâmetros.

3.4 Aplicações

A previsão de ST, focando o uso de métodos de AM, apresenta diversas aplicações em múltiplas áreas do conhecimento. Dentre os numerosos exemplos, podem ser citados: em economia, Hanias and Curtis (2008) utilizaram uma variação do algoritmo *kNN* para estimar a taxa cambial entre o dólar e o euro, utilizando uma ST de frequência diária cobrindo dados entre 01/01/2001 e 31/08/2001. Foram realizadas previsões variando de um a sessenta dias à frente. Os autores demonstraram que a série utilizada possui características caóticas, e identificaram um melhor balanceamento entre os erros de previsão para a janela de trinta dias à frente.

Em hidrologia, Aitkenhead and Cooper (2008) empregaram um método baseado na utilização de uma RNA para prever a possibilidade de inundação e de alta vazão de água em nascentes no nordeste da Escócia. Foram utilizadas ST relativas a vinte e uma variáveis ambientais, coletadas como médias horárias no período entre 20/07/2004 e 01/11/2004. Buscaram realizar previsões com horizontes de até vinte e quatro horas. Os resultados indicaram que a

quantidade de variáveis bem previstas diminuía à medida que se aumentava a quantidade de horas à frente previstas. Porém, o modelo foi considerado promissor mesmo em um horizonte de previsão de vinte e quatro horas, indicando sua possibilidade de integração com os sistemas de monitoramento tradicionais. Odan et al. (2009) realizaram um comparativo entre a exatidão de previsão de um método baseado na Transformada de Fourier e o algoritmo *kNN-TSP*, para a previsão de consumo em sistemas de abastecimento de água. Os dados utilizados datam de 19/01/2005 à 26/03/2005 e 01/05/2005 à 26/08/2005. O horizonte de previsão utilizado foi de vinte e quatro horas. Nesse trabalho verificaram-se melhores resultados de previsão utilizando o algoritmo *kNN-TSP* em comparação com a Transformada de Fourier.

Em monitoração ambiental, Ferrero et al. (2008) aplicaram o algoritmo *k-Nearest Neighbor* para a previsão de dados ambientes referentes à temperatura da água do lago de Itaipu. Foram utilizadas 44 observações de temperatura da água coletadas trimestralmente entre os anos de 1994 e 2004, visando prever um valor futuro. Especialistas da área consideraram os resultados promissores para a previsão de dados ambientais.

Em medicina, Verplancke et al. (2010), utilizaram uma RNA para a previsão da necessidade de diálise em uma Unidade de Terapia Intensiva (UTI). Foram coletados dados de duas variáveis biológicas de pacientes entre 31/05/2003 e 17/11/2007. Nesse trabalho, os autores buscaram estimar a necessidade de diálise entre o quinto e o décimo dia de internação de pacientes, utilizando os dados coletados dos três primeiros dias. Os resultados demonstraram-se promissores a serem incorporados às técnicas de estimativas das UTIs, por apresentarem resultados semelhantes a outros dois métodos tradicionais empregados (*Support Vector Machine* (*SVN*) e *Naive Bayes*), porém com menor custo computacional.

Na computação, Lo (2011) utilizou uma abordagem híbrida baseada em um modelo ARIMA e um modelo *SVN* para a previsão da quantidade de falhas em *software* durante o desenvolvimento de sistemas. O modelo híbrido foi desenvolvido considerando o fato das falhas de *software* apresentarem tanto características paramétricas, as quais são mais bem detectadas pelo modelo ARIMA, quanto não-paramétricas, as quais são mais bem detectadas pelo modelo *SVN*. Os resultados demonstraram que a abordagem híbrida se aproxima mais dos valores reais do que se fossem empregadas isoladamente.

3.5 Considerações Finais

Diversos métodos de previsão de dados temporais podem ser empregados. Neste capítulo foram apresentados alguns dos métodos paramétricos e não-paramétricos mais mencionados na literatura. Foram também introduzidos alguns conceitos de AM necessários para um melhor entendimento do algoritmo *kNN-TSP*, sendo esse o método de previsão adotado neste trabalho.

Ao final do capítulo foram apresentadas aplicações da tarefa de previsão de ST em casos reais. O próximo capítulo apresenta considerações sobre diversas medidas de similaridade e suas características, tema importante para se compreender sua influência no resultado das previsões utilizando o algoritmo *kNN-TSP*, foco deste trabalho.

Capítulo 4

Similaridade entre Séries Temporais

4.1 Considerações Iniciais

O conceito de similaridade é importante para diversas tarefas de MD tradicionais e aplicadas à ST. Por exemplo, as tarefas de agrupamento dependem do conceito de similaridade entre os membros de um conjunto para decidir em que grupo um determinado membro se enquadra melhor.

Essa semelhança entre membros é definida, geralmente, através da utilização de uma medida de similaridade e o conhecimento das características dessas medidas permite uma melhor escolha entre as medidas disponíveis. Assim, neste capítulo são apresentadas características e particularidades de algumas das mais conhecidas medidas de similaridade utilizadas com dados temporais.

4.2 Medidas de Similaridade entre Séries Temporais

A medida de similaridade define o critério para quantificar quão similares são duas sequências e decidir se serão denominadas como pertencentes ou não a determinado padrão. A aplicação dessa definição em ST é bastante subjetiva, pois é dependente de diversos fatores. Entre os fatores que exercem grande influência, estão o domínio da aplicação e as características do método escolhido para o cálculo da similaridade (Hetland, 2004).

As medidas de similaridade podem ser agrupadas em três categorias, dependendo da maneira de considerar os dados (Lhermitte et al., 2011):

Diretamente com os dados originais da ST: Nessa categoria enquadram-se as medidas que permitem calcular a distância entre as séries diretamente através de seus dados originais, ou seja, não é aplicada nenhuma transformação nos dados. Pode-se subdividir essa categoria em duas abordagens: (1) distâncias, sendo os exemplos mais expressivos as métricas

derivadas da Norma L_p, tais como a distância Euclidiana, Manhattan e Mahalanobis; e (2) medidas de correlação, que quantificam o grau de relacionamento linear entre as séries, como a correlação de Pearson;

Transformando os dados originais: Nessa categoria enquadram-se as medidas que permitem calcular a distância entre as séries após aplicar alguma transformação em seus dados. De maneira geral, essas transformações possuem dois objetivos principais: (1) reduzir a dimensão da série, buscando sempre perder a menor quantidade de informações possível; e (2) isolar características específicas de seus componentes. Dentre os principais exemplos de transformação podem ser citados o *Principal Components Analisys* (*PCA*) e a Transformada de Fourier. Assim, a similaridade entre as séries é calculada após a realização dessas transformações;

Aplicando métricas aos dados: Nessa categoria enquadram-se as medidas que permitem calcular a distância entre as séries através de transformações de seus dados, utilizando um conjunto de parâmetros que descrevem a série. De posse desse conjunto de parâmetros descritores, realizam-se os cálculos de similaridade entre as séries. Desse modo, caso haja interesse apenas em características específicas, como a duração de tendências, pode-se criar medidas para calcular o tempo de duração das tendências e se realizar o cálculo da similaridade utilizando-se essa característica. Caso a transformação do dado permita ressaltar corretamente a característica desejada, medidas nessa categoria apresentam a vantagem de serem boas descritoras de séries específicas, porém, necessitam de conhecimento externo para sua interpretação. Entretanto, por realizar transformações levando em consideração características específicas da série, essas medidas podem não ser bem aplicadas para outras séries.

Diversas técnicas para quantificar a similaridade entre ST têm sido propostas na literatura e avaliadas em função de alguma tarefa de interesse. Nesse sentido, podem ser citados: o estudo de Fabris et al. (2008), no qual é proposta a combinação de várias medidas de similaridade, através da atribuição de pesos, como uma alternativa à distância Euclidiana. Experimentos foram conduzidos comparando a medida proposta com a *Dynamic Time Warping* (*DTW*), demonstrando a viabilidade da medida proposta para a classificação de ST. Em Ding et al. (2010), as distâncias Euclidiana, *Pattern Distance* (*PD*) e *Included Angle Distance* (*IAD*), sendo a última proposta no trabalho, foram comparadas para verificar suas eficiências no reconhecimento ST de mercado financeiro. Os resultados demonstraram que a distância proposta é tão eficiente em encontrar as similaridades quanto à distância Euclidiana, porém, com menor custo computacional. No trabalho de Yan et al. (2010), a distância proposta, *Key Points Based Similarity Measure* (*KPDIST*) é comparada com as distâncias Euclidianas e *Dynamic Time Warping* utilizando o algoritmo *k-Nearest Neighbor for Continuous Time Series* (*kNNC*) e com as distâncias *Simple Matching Coefficient* (*SMC*) e *normalized Longest Common Subsequence* (*nLCS*) usando o

algoritmo *k-Nearest Neighbor for Discrete Time Series* (*k-NND*), com o intuito de detectar anomalias no comportamento de ST. Os resultados desses experimentos mostram que, para dados contínuos, a medida proposta é competitiva com as medidas de distâncias comparadas e, para dados discretos, o método proposto mostra-se mais eficiente que as medidas estudadas para a detecção de anomalias.

Observa-se na literatura a predominância da utilização de certas medidas, por representarem a noção de similaridade de maneira intuitiva e, em geral, alcançarem bons resultados com os métodos propostos. A seguir são descritas algumas das principais medidas de similaridade aplicadas em ST.

4.2.1 Norma L_p

As medidas da Norma L_p, em especial a distância Euclidiana, estão entre as medidas de distância mais conhecidas e exploradas na literatura, geralmente tendo sua aplicação expandida para dados bidimensionais, tridimensionais, ou de maior número de dimensões. A forma intuitiva de considerar a distância, bem como a simplicidade de implementação e interpretação, favorece a seleção de medidas dessa norma para calcular a similaridade em um grande número de aplicações.

Para o cálculo da distância baseado na Norma L_p, cada sequência é considerada um ponto no espaço W-dimensional. Desse modo, a similaridade entre essas sequências é dada pela diferença entre esses pontos. Essa norma é definida pela Equação 4.1 (Aggarwal et al., 2001):

$$L_p(x,y) = \left(\sum_{i=1}^{W} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{4.1}$$

onde x e y são vetores W-dimensionais, os quais representam sequências contidas em uma ST; e p define a medida de distância a ser utilizada.

Apesar de serem amplamente utilizadas para quantificar a similaridade de ST, as medidas da Norma L_p apresentam limitações. Entre essas limitações, verificam-se a baixa capacidade para o tratamento de *outliers* (valores fora do intervalo de determinada distribuição) e o fato de serem sensíveis a pequenas distorções e deslocamentos no eixo temporal da série. Outra desvantagem das medidas dessa família é o fato de estarem suscetíveis à chamada "maldição da dimensionalidade", que se refere ao fato de que a complexidade do cálculo da similaridade é proporcional à quantidade de dimensões dos dados, ou seja, quanto maior a dimensionalidade dos dados, maior a quantidade de operações necessárias (Vlachos et al., 2004).

De acordo com o valor de p, têm-se medidas com nomes e comportamento específicos. Neste trabalho, as medidas da Norma L_p foram subdividas em dois grupos: L_p Inteiras, que

representam valor de p igual ou superiores a um (1) sem casas decimais, e L_p Fracionárias, representando valores de p superiores a zero (0) e inferiores a um (1), indicando assim valores fracionários. Ambos os grupos são detalhados a seguir.

Norma L_p Inteiras

Como mencionado, essas medidas são bem conhecidas pela comunidade cientifica, sendo empregadas em diversos trabalhos nos mais variados domínios. Dentre as mais conhecidas estão a distância Euclidiana e a distância Manhattan. As medidas L_p Inteiras são nomeadas de acordo com o valor de p (Felipe et al., 2006):

- $p = 1$: Manhattan, também conhecida como *City Block* (L_1);
- $p = 2$: Euclidiana (L_2);
- $p = 3$: Métrica L_3 (L_3);
- $p = \infty$: Chebychev, também denominada Infinita (L_∞).

A distância L_1 define um espaço geométrico no qual todos os pontos possuem o mesmo valor da soma das diferenças absolutas de cada ponto; a distância L_2 define um espaço geométrico em forma de circunferência onde todos os pontos estão equidistantes em relação ao centro; a distância L_3 define um espaço geométrico quadrado de cantos arredondados; e a L_∞ define um espaço quadricular, onde a distância entre dois pontos é a maior de suas diferenças. A variação dos espaços geométricos dessas medidas pode ser visibilizada na Figura 4.1.

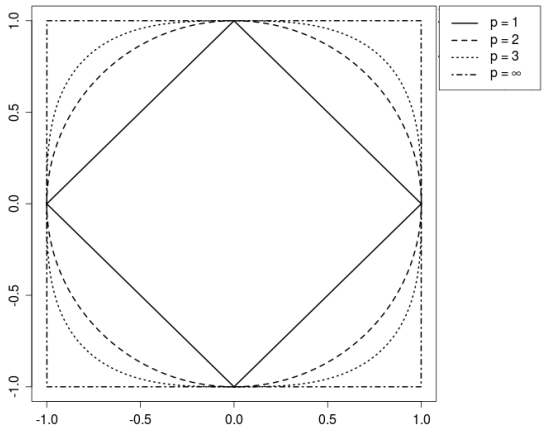


Figura 4.1: Efeito da variação de p na Norma L_p para L_1, L_2, L_3 e L_∞.

A partir da Equação 4.1 é possível observar que o custo computacional para calcular a distância está relacionado ao valor de p. Desse modo, para valores de p inteiros, quanto maior esse valor, maior a quantidade de operações a serem realizadas. Assim, entre as medidas Manhattan, Euclidiana e Métrica L_3, a distância Manhattan é a de menor custo computacional, a

Métrica L_3 a de maior custo e a Euclidiana apresenta custo intermediário entre as duas distâncias anteriores.

Norma L_p Fracionárias

Aggarwal et al. (2001) demonstraram que valores de p da Norma L_p menores que um, em conjuntos de grande dimensionalidade, apresentam vantagens se comparados às medidas com valores inteiros. Essa vantagem se dá pelo fato das medidas fracionárias atribuírem mais peso a pequenas variações entre os dados.

Enquanto as medidas L_p Inteiras tendem a formar figuras quadriculares de cantos arredondados, a medida que se aumenta o valor de p de um a L_∞ (infinito), até se alcançar um figura quadrática no espaço de busca, as medidas fracionárias tendem a se retrair. Assim, o losango formado por $p = 1$ tende a ter suas laterais atraídas ao centro do espaço de busca. A variação dos espaços geométricos dessas medidas, para valores de p iguais à $p = 0.1, 0.3, 0.5$ e 0.7 pode ser visibilizada na Figura 4.2.

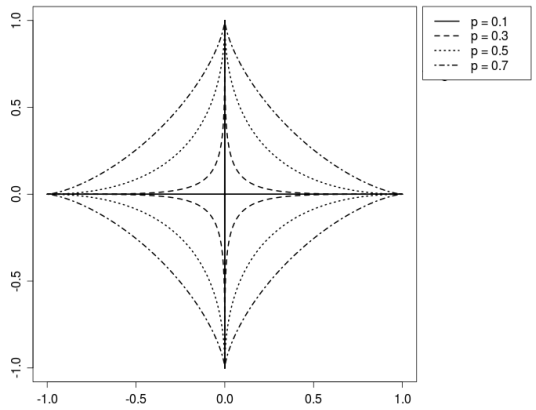


Figura 4.2: Efeito da variação de p na Norma L_p para $L_{0.1}$, $L_{0.3}$, $L_{0.5}$ e $L_{0.7}$.

4.2.2 Canberra

A distância Canberra (c_a) assemelha-se à distância Manhattan, calculando a diferença absoluta entre dois vetores. A diferença decorre do fato da distância Canberra dividir a diferença absoluta pela soma dos valores absolutos, antes de realizar a soma. Assim, dados dois vetores, x e y, W-dimensionais, a distância Canberra é dada pela Equação 4.2 (Deza and Deza, 2006):

$$c_a(x,y) = \sum_{i=1}^{W} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{4.2}$$

A distância Canberra é uma versão ponderada da distância Manhattan, onde as diferenças

entre valores próximos à origem apresentam-se maiores quando comparados aos valores mais distantes da origem (Jurman et al., 2009). Dessa maneira, essa medida é menos influenciada por variáveis que apresentam valores elevados, quando comparada com a distância Manhattan.

Em relação ao custo computacional, assim como a distância Manhattan, a distância Canberra apresenta complexidade linear. A diferença com relação à distância Manhattan é a adição que ocorre no denominador, fazendo com que a distância Canberra apresente maior custo computacional (quantidade de operações a serem realizadas) que a Manhattan, porém menor que as demais medidas anteriormente apresentadas.

A Figura 4.3[1] demonstra a variação do espaço geométrico de busca da distância Canberra. Pode ser observada, nessa figura, a maior sensibilidade da distância Canberra para valores próximos à origem.

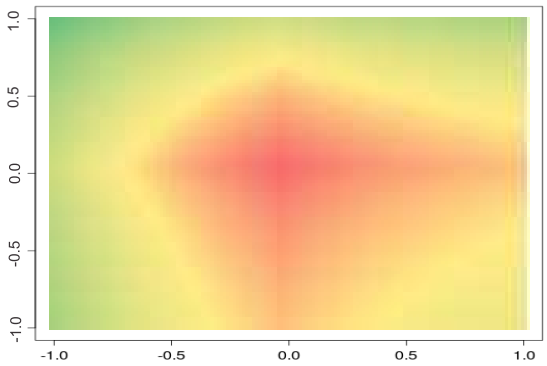


Figura 4.3: Variação do espaço geométrico de busca da distância Canberra (Modificado de http://people.revoledu.com/kardi/tutorial/Similarity/).

4.2.3 Geodésica

A distância Geodésica (g_e), nome derivado de Geodésia, ciência de mensuração do tamanho e forma da Terra, é uma generalização da noção de linha reta para espaços curvos. Em espaços planos a Geodésica é um segmento de reta, enquanto que em outras superfícies isso pode não ser verdade. Em um plano cilíndrico, por exemplo, a Geodésica depende dos pontos do plano escolhidos, podendo ser uma reta, uma circunferência ou até mesmo uma hélice. Já em um plano esférico, a Geodésica consiste em um arco.

Assim, um segmento geodésico é a menor curva entre dois pontos. Um espaço de métrica pode ser considerado geodésico se quaisquer dois pontos do espaço estão unidos por um segmento geodésico. Desse modo, a distância Geodésica é o comprimento de um segmento geodésico entre dois pontos em um espaço geodésico (Deza and Deza, 2006).

[1]Teknomo, Kardi.*Similarity Measurement*. Disponível em: http://people.revoledu.com/kardi/tutorial/Similarity/

Neste trabalho será utilizada a distância Geodésica segundo a implementação de Meyer and Bucht (2011). Assim, dados dois vetores, x e y, W-dimensionais, a distância Geodésica é dada pela Equação 4.3:

$$g_e(x,y) = \arccos\left(\frac{xy}{\sqrt{xx \times yy}}\right) \tag{4.3}$$

onde xy indica a multiplicação escalar dos vetores x e y; e xx e yy indicam a multiplicação escalar do vetor x com ele próprio e do vetor y com ele próprio, respectivamente. A distância Geodésica entre os vetores é dada pelo ângulo desses vetores no espaço W-dimensional.

A complexidade para o cálculo da distância Geodésica é linear, onde a maior quantidade de operações ocorre nas multiplicações escalares efetuadas. Assim, essa grande quantidade de multiplicações necessárias acaba por tornar essa distância mais custosa computacionalmente do que as distâncias anteriores.

A Figura 4.4 ilustra a distância Geodésica em um espaço bidimensional. A linha tracejada representa o ângulo entre os vetores x e y.

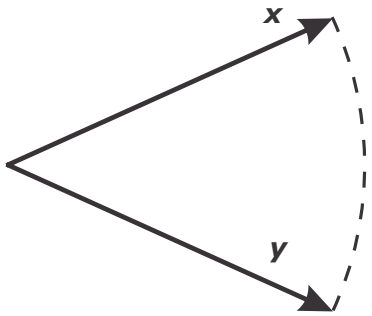


Figura 4.4: Distância Geodésica em um espaço bi-dimensional. A distância é o comprimento da curva (ângulo) representada pela linha tracejada.

4.2.4 *Dynamic Time Warping*

A técnica *Dynamic Time Warping* (*DTW*), desenvolvida baseada na distância de Levenshtein (Levenshtein, 1966), foi apresentada para a comunidade de MD no trabalho de Berndt and Clifford (1994). Ela já era utilizada com sucesso na área de reconhecimento de fala, e foi introduzida para a análise de séries temporais como uma solução para um dos principais problemas das distâncias da Norma L_p, adicionando robustez à comparação de ST defasadas no eixo do tempo.

Essa distância busca alinhar, da maneira mais adequada possível, os valores das séries a serem comparadas (Petitjean et al., 2010). Isso permite que duas ST globalmente similares, mas que estejam fora de alinhamento no eixo temporal, possam ser alinhadas para pos-

terior comparação ponto-a-ponto. Desse modo, a *DTW* permite encontrar semelhança entre séries as quais não seriam possíveis utilizando apenas a abordagem ponto-a-ponto tradicional (Ratanamahatana and Keogh, 2005). O processo de alinhamento consiste em mapear pontos de uma série em outra, e pode ser observado pela Figura 4.5.

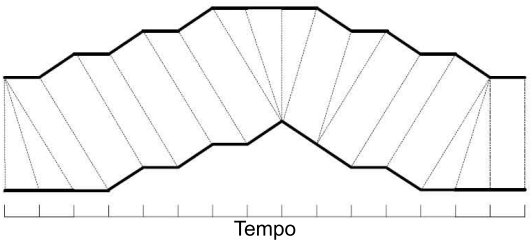


Figura 4.5: Alinhamento entre duas séries (Modificado de Salvador and Chan (2007)).

A técnica *DTW*, em alguns casos, pode apresentar resultados melhores na comparação de séries do que a distância Euclidiana, e a última pode ser considerada um caso especial da primeira, onde as séries possuem a mesma dimensão. As principais desvantagens do *DTW* estão relacionadas à sua complexidade $O(m.n)$ (usando programação dinâmica), sendo m e n os tamanhos da primeira e segunda série, respectivamente, e o fato da complexidade de memória possuir uma relação quadrática ao tamanho das séries (Chu et al., 2002).

Para alinhar duas séries Z e T, de tamanhos m e n, respectivamente, o algoritmo constrói uma matriz $m \times n$. Nessa matriz, cada elemento (i, j) corresponde ao valor da distância entre os pontos (Z_i, T_j). De posse dessa matriz busca-se uma rota R, que alinhe as séries Z e T, conforme a Equação 4.4.

$$R = (r_1, r_2, \ldots, r_L) \tag{4.4}$$

onde cada r_l corresponde a um mapeamento $(i, j)_L$ para $l = 1, \ldots, L$ e L, representando o tamanho da rota, está restrito à condição $max(m,n) \leq L < m+n$ (Salvador and Chan, 2007).

Diversas rotas podem ser encontradas, sendo que a que se deseja encontrar é aquela que minimize a distância (custo de alinhamento), conforme a Equação 4.5 (Berndt and Clifford, 1994). Esse processo pode ser visibilizado na Figura 4.6.

$$DTW(Z,T) = min_R \left[\sum_{l=1}^{L} \delta(r_L) \right] \tag{4.5}$$

onde δ representa a medida de distância utilizada.

De maneira a reduzir o custo computacional, a busca dessa rota é realizada empregando-se programação dinâmica. Sua utilização apresenta uma formulação simples na qual, soma-se à distância acumulada da célula atual a menor de suas três adjacentes, sendo essas a célula à

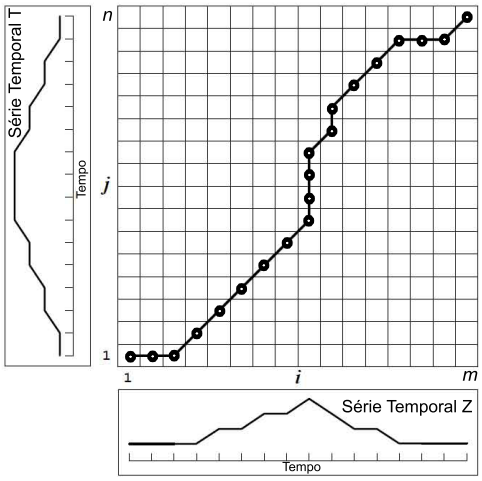


Figura 4.6: Matriz de custo com rota de alinhamento traçada (Modificado de Salvador and Chan (2007)).

esquerda, superior ou diagonal superior direita. Uma vez preenchida essa matriz, pode-se iniciar a busca da rota de alinhamento que apresente o menor custo acumulado (Ratanamahatana and Keogh, 2005).

A ideia da utilização da *DTW* como medida de distância entre ST consiste em considerar o custo de alinhamento como distância entre as séries. Assim, quanto maior o custo de alinhamento das séries, isto é, maior quantidade de mapeamentos entre os pontos da série, maior a distância entre elas. Com relação ao seu custo, essa medida é a mais custosa dentre as apresentadas, devido à quantidade de operações necessárias para a construção das matrizes de distância e de alinhamento entre as séries.

4.3 Considerações Finais

Diversas medidas de similaridade podem ser encontradas na literatura, cada uma apresentando características e limitações específicas. Essas medidas podem levar em consideração os dados originais das ST ou realizar transformações sobre os mesmos de maneira a ressaltar características específicas das séries, podendo favorecer determinadas análises. Neste capítulo foram apresentadas algumas das medidas de similaridades mais presentes na literatura. Foram apresentadas também as características dessas medidas, de maneira a se compreender seu funcionamento.

No próximo capítulo serão apresentadas as discussões e resultados experimentais realizados com o algoritmo *kNN-TSP* utilizando diversas medidas de similaridade, tanto em ST geradas através de equações matemáticas como com séries reais provenientes de dados de transporte.

Capítulo 5

Avaliação Experimental

5.1 Considerações Iniciais

Neste capítulo são apresentados os dados experimentais, compostos por ST artificiais, isto é, séries geradas por meio de funções matemáticas, que permitem um melhor controle da avaliação, bem como diversas séries reais provenientes de dados relacionados ao transporte. A configuração dos experimentos, bem como a análise e a discussão dos resultados dos experimentos, realizados com o objetivo de estudar a influência das medidas de similaridade utilizando o algoritmo *kNN-TSP*, também são apresentados neste capítulo.

5.2 Séries Temporais Utilizadas para Avaliação Experimental

De maneira a avaliar o comportamento do algoritmo *kNN-TSP* frente a ST de diferentes características, foram selecionadas séries artificiais e reais para a realização dos experimentos. As diferentes características dos dois conjuntos de séries permitem uma melhor avaliação da influência das medidas de similaridade no algoritmo frente a uma grande quantidade de situações.

5.2.1 Séries Artificiais

Foram utilizadas cinco séries temporais artificiais para avaliar o comportamento do algoritmo utilizando as diferentes medidas de similaridade. Essas séries estão agrupadas em duas famílias de acordo com as suas características: séries temporais de modelos sazonais (STS) e séries temporais de modelos caóticos (STC). As séries de ambas as famílias são apresentadas a seguir e sumarizadas na Tabela 5.1.

Tabela 5.1: Características das ST artificiais.

Séries Temporais de Modelos Sazonais		
Id	**Série Temporal**	***m***
STS1	Dependência sazonal	2200
STS2	Sazonalidade multiplicativa	590
STS3	Alta frequência	550
Séries Temporais de Modelos Caóticos		
STC1	Lorenz	551
STC2	Mackey-Glass	551

Séries Temporais de Modelos Sazonais (STS)

Essas séries permitem avaliar o algoritmo de previsão em comportamentos previsíveis. Em geral, apresentam tendência definida e mudança de amplitude de modo sazonal. As séries utilizadas neste trabalho são exemplificadas na Figura 5.1. Detalhes das equações utilizadas podem ser encontrados em (Kulesh et al., 2008; Ferrero, 2009).

Série temporal de dependência sazonal (STS1) — a série gerada por esse modelo possui sazonalidade constante e tendência linear;

Série temporal de sazonalidade multiplicativa (STS2) — a série resultante considera variação de tendência não-linear e sazonalidade multiplicativa. Assim, as oscilações crescem ao longo do tempo;

Série temporal de alta frequência (STS3) — a série desenvolvida possui dados que consideram sazonalidade multiplicativa e constante aumento de amplitude.

Séries Temporais de Modelos Caóticos (STC)

Essas séries temporais permitem avaliar algoritmos de previsão, frente a comportamentos pouco previsíveis e que apresentam ciclos não-repetitivos. Desse modo, podem apresentar comportamentos globais similares ao passado, porém não-idênticos. Esse fato aumenta a dificuldade de previsão dessas séries em relação às de modelos sazonais. As séries temporais de modelos caóticos utilizadas neste trabalho são exemplificadas na Figura 5.2 (McNames, 1999; Kulesh et al., 2008). Detalhes das equações, bem como os parâmetros utilizados podem ser encontrados em McNames (1999).

Sistema de Lorenz (STC1) — permite a geração de ST com comportamento não-periódico e imprevisível por meio de um sistema de equações diferenciais;

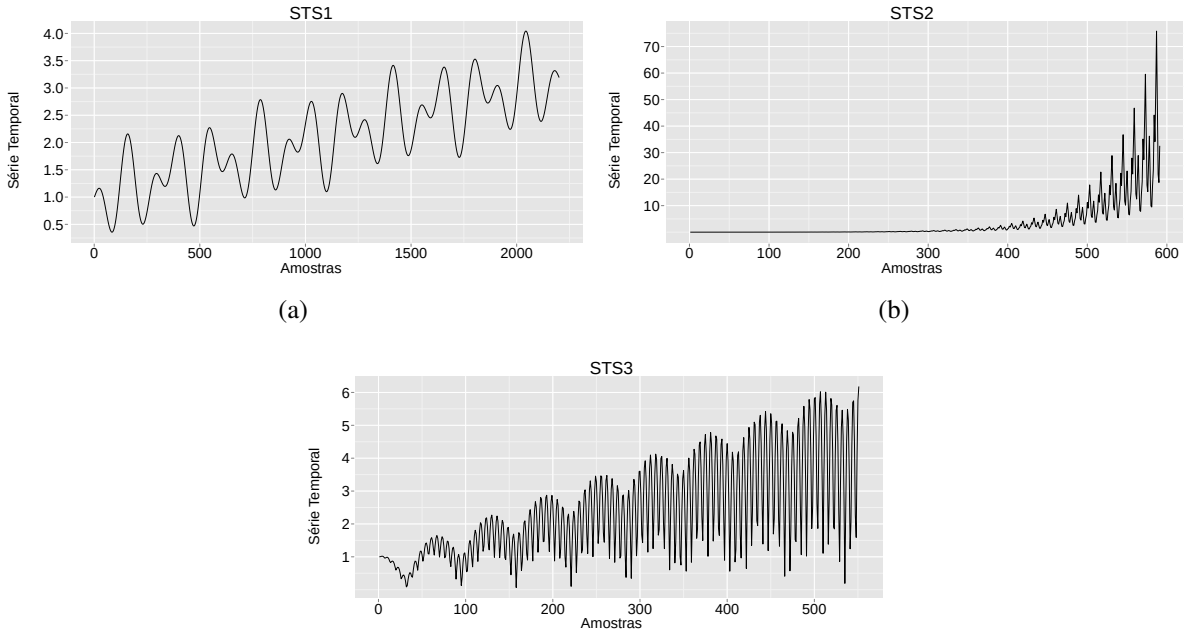


Figura 5.1: Séries temporais artificiais geradas através de modelos sazonais: (a) STS1, (b) STS2 e (c) STS3 (Modificado de Kulesh et al. (2008)).

Sistema de Mackey-Glass (STC2) — permite a criação de sistemas caóticos por meio de um sistema de equações originalmente desenvolvido para modelar a formação de linfócitos. Essa série é atualmente muito utilizada na literatura por trabalhos que avaliam o comportamento de seus métodos em séries caóticas.

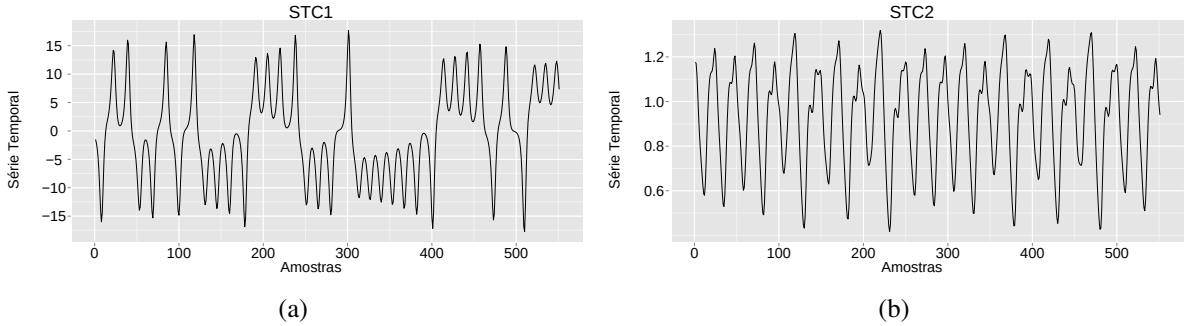


Figura 5.2: Séries temporais artificiais geradas por modelos caóticos: (a) STC1 e (b) STC2 (Modificado de McNames (1999) e Kulesh et al. (2008)).

5.2.2 Séries Reais

De maneira a verificar o comportamento do algoritmo e das medidas de similaridade frente a ST reais, foram utilizadas diversas séries disponibilizadas pela edição do ano de 2010

da *Time Series Forecasting Grand Competition for Computational Intelligence (NN GC1)*[1]. Ela é uma extensão das edições *NN3* (dos anos 2006-2007) e *NN5* (do ano de 2008) da *Time Series Forecasting Competition for Computational Intelligence*.

A *NN GC1*, é uma competição de previsão de ST com o objetivo de avaliar a acurácia de métodos de Inteligência Computacional aplicados à previsão de dados temporais. Desse modo, contribui para disseminar conhecimento de melhores práticas em séries temporais de diferentes frequências, bem como a evolução desses métodos. A competição aceita qualquer abordagem que se utilize de Inteligência Computacional, tais como redes neuronais, preditores fuzzy, algoritmos genéticos e evolucionários, árvores de decisão, entre outros.

Os competidores podem treinar e avaliar seus algoritmos antes da submissão ao evento. Para isso, estão disponíveis seis bases de dados, cada uma contendo onze ST, com diferentes intervalos de aquisição, tamanhos e datas de início e fim de aquisições (essas características de todas as séries da *NN GC1* utilizadas neste trabalho estão no Apêndice A).

Na Tabela 5.2 são apresentadas as ST da *NN GC1* agrupadas por base de dados, a frequência de aquisição dos dados e o tamanho das séries. Neste trabalho não foram consideradas as séries da base 1.A, bem como as séries 1.D-002, 1.E-003, 1.F-001 e 1.F-002, para a realização dos experimentos (maiores detalhes são apresentados na Seção 5.3).

Tabela 5.2: Características das ST disponíveis pela *NN GC1*.

Séries	**Base de Dados**	**Aquisição**	**Tamanho**
1.B-001 a 1.B-011	1.B	Quaternal	31 a 148
1.C-001 a 1.C-011	1.C	Mensal	48 a 228
1.D-001 a 1.D-011	1.D	Semanal	437 a 618
1.E-001 a 1.E-011	1.E	Diária	377 a 747
1.F-003 a 1.F-011	1.F	Horária	902 a 1742

Todas as séries são mensurações de dados relacionados ao transporte. Essas mensurações incluem tráfego em rodovias, em túneis, em estações de pedágio automatizadas, em ferrovias, tráfego de pessoas em metrôs, voôs domésticos, quantidades de navios de importação entre outros.

Devido às diferentes frequências de aquisição e combinação de forças casuais (mudanças de calendários, feriados e fatores não-observados), as séries podem possuir nenhum ou vários dos componentes básicos das ST, tais como tendências locais, sazonalidades, valores fora do padrão esperado ou valores faltantes. Esses componentes casuais podem exercer diferentes influências em cada série. Valores faltantes foram disponibilizados como zero (0), assim, em algumas séries, a presença desse valor pode representar a ausência de determinado ponto, bem como ser um valor zero real.

[1] http://www.neural-forecasting-competition.com/

A Figura 5.3 apresenta a primeira ST de cada uma das bases disponibilizadas pela *NN GC1*.

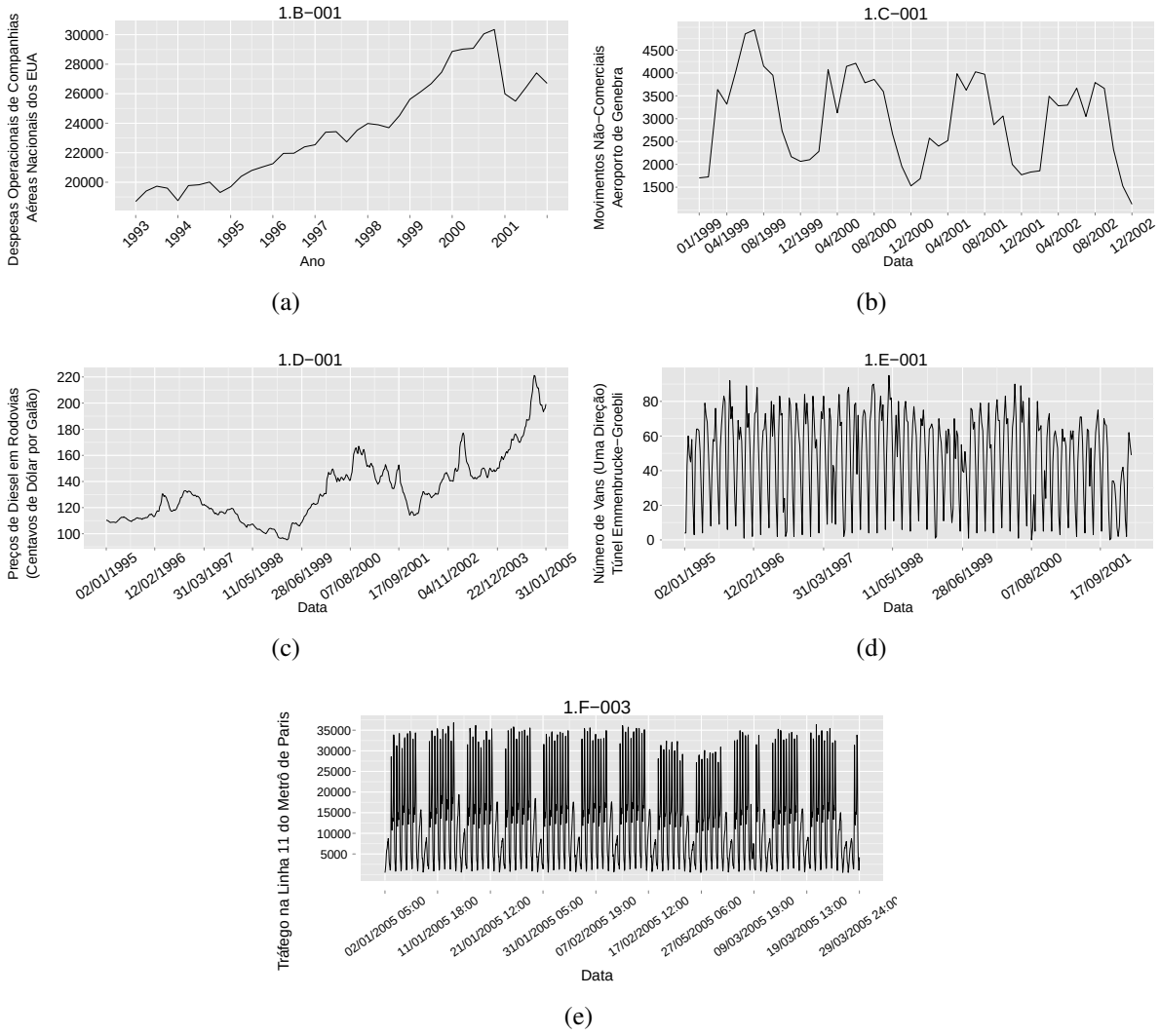


Figura 5.3: Séries reais disponibilizadas pela *NN GC1*: (a) 1.B-001, (b) 1.C-001, (c) 1.D-001, (d) 1.E-001 e (e) 1.F-003.

5.3 Configuração Experimental

Como mencionado, o algoritmo *kNN-TSP* necessita da configuração de alguns parâmetros, dentre eles a medida de similaridade, o tamanho da janela de busca, a quantidade de pontos a serem previstos e a quantidade de vizinhos mais próximos (k). Neste trabalho, a quantidade de vizinhos mais próximos considerada foi de $k = 1$, $k = 5$ e $k = 10$ para todas as séries e para as medidas de similaridade da Norma L_p (inteiras e fracionárias), Canberra, Geodésica e *DTW*. Essa quantidade de vizinhos foi selecionada devido ao fato de, em estudos preliminares, representarem os momentos de variação de comportamento da curva de erro do algoritmo. Valores de k maiores que um auxiliam também a lidar com alguns dos pontos negativos de determinadas

medidas, como a sensibilidade aos *outliers* da Norma L_p, uma vez que mais de uma sequência é considerada.

O valor do tamanho de janela das séries artificiais, *w*, foi selecionado baseado em Kulesh et al. (2008). Na Tabela 5.3 são apresentados o tamanho da série (*m*), os valores de *w* e o número de valores a serem previstos em cada ST artificial.

Tabela 5.3: Configuração dos parâmetros *w*, *m* e número de valores previstos para as ST artificiais.

Id	Série Temporal	*m*	*w*	Valores Previstos
STS1	Dependência sazonal	2200	100	220
STS2	Sazonalidade multiplicativa	590	15	88
STS3	Alta frequência	550	70	55
STC1	Lorenz	551	25	100
STC2	Mackey-Glass	551	7	100

O tamanho da janela para as séries reais foi selecionado de acordo com a sazonalidade sugerida por Lemke and Gabrys (2010) para cada uma das bases de dados. A quantidade de valores previstos é relativa à solicitação de previsão da competição *NN GC1* para cada uma das séries. Ambas informações podem ser observadas na Tabela 5.4.

As séries da *NN GC1* da base 1.A, não apresentam sazonalidade sugerida por Lemke and Gabrys (2010), não havendo então uma recomendação de sazonalidade para o parâmetro de tamanho da janela. Assim, os experimentos foram conduzidos sem utilizar as séries da base 1.A. A série 1.D-002 apresenta inconsistência de dados entre a descrição das datas de aquisição das séries e a quantidade de observações, sendo que a última superava em grande quantidade o esperado, dessa forma essa série foi removida da lista de séries avaliadas nos experimentos. As séries 1.E-003, 1.F-001 e 1.F-002 foram removidas da lista de séries para os experimentos por apresentarem grande quantidade de valores faltantes, quando relacionadas às descrições das séries e sua respectiva quantidade de pontos, chegando a apresentar um total de valores faltantes próximos a 6% do tamanho esperado da série, como no caso da série 1.F-001.

Tabela 5.4: Sazonalidade e quantidade de pontos previstos para as ST da *NN GC1*.

Base de Dados	Sazonalidade	Pontos Previstos
1.B	4	8
1.C	12	12
1.D	52	26
1.E	7	14
1.F	24	48

O algoritmo *kNN-TSP* foi inicialmente proposto utilizando a distância Euclidiana como medida para definir a similaridade entre as sequências (Ferrero et al., 2009). Essa distância pertence a um conjunto de medidas, conhecidas como Norma L_p, sendo que essas medidas

constituem uma das abordagens mais aplicadas para determinar a distância com custo computacional linear.

De maneira a ampliar os critérios de similaridade adotados, a implementação do algoritmo *kNN-TSP* foi expandida para utilizar outros valores de p da Norma L_p, inclusive fracionários, bem como medidas de similaridade não pertencentes à família dessa norma. Assim, foram utilizadas medidas de similaridade da Norma L_p, com valores de p inteiros variando de um à três; valores fracionários, sendo $p = 0.1$, $p = 0.3$, $p = 0.5$ e $p = 0.7$; a distância *Dynamic Time Warping*; a distância Canberra; e a distância Geodésica.

Neste trabalho, o algoritmo *kNN-TSP* foi utilizado para realizar previsões de um único valor futuro, considerando, a cada estimativa, todos os valores observados do passado.

A função de previsão utilizada em conjunto com o *kNN-TSP* neste trabalho foi a Média de Valores Relativos (MVR), apresentada em Ferrero (2009). Essa função foi escolhida por apresentar melhores resultados na previsão de ST, se comparadas com abordagens tradicionais da literatura, tanto em séries com dependência sazonal e tendência conhecidas, como em séries caóticas (Ferrero et al., 2009).

Uma das vantagens dessa função de previsão sobre as abordagens tradicionais na literatura, é o fato de permitir a estimativa de valores futuros levando em consideração padrões em níveis diferentes de tendência.

A MVR é definida pela Equação 5.1 (Ferrero et al., 2009):

$$f_{MVR}(S') = z_n + \frac{\sum_{i=1}^{k} \Delta s'_{i,w+1}}{k} = \hat{z}_{n+1} \tag{5.1}$$

onde $S' = \{s'_1, s'_2, \ldots, s'_k\}$ representa o conjunto de sequências mais similares do passado; z_n representa o último valor observado; $\hat{z}_{n+1}$ representa o valor futuro estimado; w define o tamanho da janela; k determina o número de vizinhos próximos; e $\Delta s'_{i,w+1} = s'_{i,w+1} - s'_{i,w}$ define a diferença entre o valor futuro e o último valor observado da i-ésima sequência similar s'_i pertencente a S'.

Os erros de predição foram medidos por meio do *Mean Absolute Percentage Error* (*MAPE*), o qual é calculado conforme a Equação 5.2 (Hyndman and Koehler, 2006):

$$MAPE = \frac{1}{m} \sum_{t=1}^{m} \left| \frac{z_t - a_t}{z_t} \right| \times 100 \tag{5.2}$$

onde a_t indica um valor previsto no intervalo de tempo t.

O resultado da aplicação dessa equação é um valor percentual que relaciona o valor previsto e o valor real da série. O *MAPE* foi escolhido devido a sua característica de independência da escala da ST, permitindo a comparação de erros entre séries de diferentes escalas. Assim,

quanto menor o valor de *MAPE* mais precisa foi a previsão.

Para a análise quanto à existência de diferença estatisticamente significativa (**d.e.s**) foi utilizado o teste estatístico não-paramétrico de Friedman para dados emparelhados e comparações múltiplas, considerando nível de significância de 5% ($p\text{-}valor < 0,05$), com pós-teste de Dunn (Motulsky, 1995).

Os algoritmos e os *scripts* deste trabalho foram desenvolvidos utilizando-se a linguagem de programação R[2] na versão 2.13.0. A linguagem R é uma linguagem interpretada, multiplataforma, voltada para aplicações estatísticas e matemáticas. É uma linguagem gratuita e de código aberto que possui uma grande variedade de ferramentas estatísticas e gráficas, objetivando ser uma alternativa livre para a conhecida linguagem comercial S.

Para apoio ao desenvolvimento dos *softwares* foi utilizado o ambiente de desenvolvimento RStudio[3] na versão 0.94.110. A ferramenta RStudio foi selecionada por apresentar uma série de facilitadores para manipulação de códigos e depuração, além de ser multiplataforma e fornecer uma interface *web* onde é possível a execução de códigos diretamente em um servidor. A análise dos dados, bem como os gráficos foram desenvolvidos utilizando-se a interface para R Deducer[4], que apresenta uma série de facilitadores gráficos para rotinas comuns da linguagem R, além de ser multiplataforma. Ambas as ferramentas são gratuitas e de código aberto.

Os experimentos foram executados em um servidor Dell OPTIPLEX 780, o qual foi isolado de maneira a não receber nenhuma outra tarefa enquanto os experimentos eram executados. O servidor possui a seguinte configuração:

- Processador Intel®Core™2 Quad Q9550 de 2.86 Ghz;
- 2×2 Gb de memória RAM DDR3 1.066 Mhz;
- Sistema operacional Ubuntu 10.10 (Maverick), Kernel Linux 2.6.35-30 generic.

A implementação do algoritmo *kNN-TSP* foi alterada de maneira a se tornar mais concisa, otimizando diversas rotinas de maneira a eliminar as redundâncias existentes no código, ampliando a modularidade e facilitando a manutenção futura dos códigos. Foi alterada a topologia de organização dos experimentos, de maneira a melhor organizar e facilitar a preparação dos experimentos.

Foram também desenvolvidas camadas de mais alto nível, de maneira a permitir que todas as configurações do algoritmo pudessem ser realizadas em apenas um *script* principal, sendo que esse *script* também permitisse a execução de várias configurações diferentes em várias ST de maneira automatizada. Foram adicionadas novas funcionalidades como controle sobre o

[2] http://www.r-project.org/
[3] http://rstudio.org/
[4] http://www.deducer.org

tempo de execução do algoritmo, bem como novos *scripts* para tabulação de dados e impressão de gráficos.

Assim como a implementacão original do algoritmo, todas essas novas funcionalidades ou adaptações de funcionalidades foram desenvolvidas em ambiente de linha de comando (modo console), exigindo que o usuário tenha um conhecimento intermediário tanto de Linguagem R e de execução de *scripts*, quanto do sistema operacional que está executando. De maneira a facilitar a utilização da solução para usuários sem esse conhecimento técnico, também foi desenvolvida uma camada de interação com o usuário em ambiente *web*, no qual usuários sem conhecimentos de Linguagem R podem realizar experimentos, utilizando um navegador através de uma interface mais amigável.

Na Figura 5.4 podem ser visibilizadas duas telas do protótipo da interface *web* para a execução de experimentos utilizando o algoritmo *kNN-TSP* para a previsão de ST. Em (a) o usuário pode realizar o envio da série para o servidor e observar um gráfico da série enviada, e em (b) o usuário pode definir os parâmetros do algoritmo *kNN-TSP* a serem utilizados para a previsão da série.

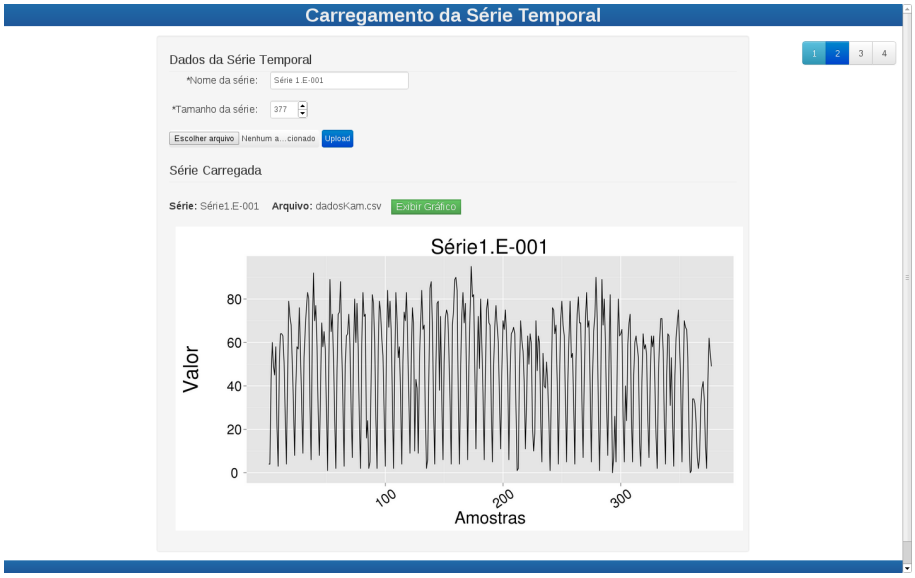
5.4 Resultados e Discussão

De modo a avaliar a influência das diversas medidas de similaridade, bem como o número de vizinhos próximos, na previsão de ST utilizando o algoritmo *kNN-TSP*, uma avaliação empírica foi realizada usando séries artificiais (McNames, 1999; Kulesh et al., 2008) e séries reais[5]. Os resultados foram analisados separadamente para cada um desses grupos de ST, e após foi realizada uma análise comparativa entre essas séries considerando cada medida separadamente e posteriormente em conjunto. Foram analisados inicialmente os valores de *MAPE* dos resultados, na sequência foram realizados testes estatísticos de maneira a verificar a existência de **d.e.s** entre eles. Ao final são discutidas as características das medidas para os conjuntos de dados avaliados.

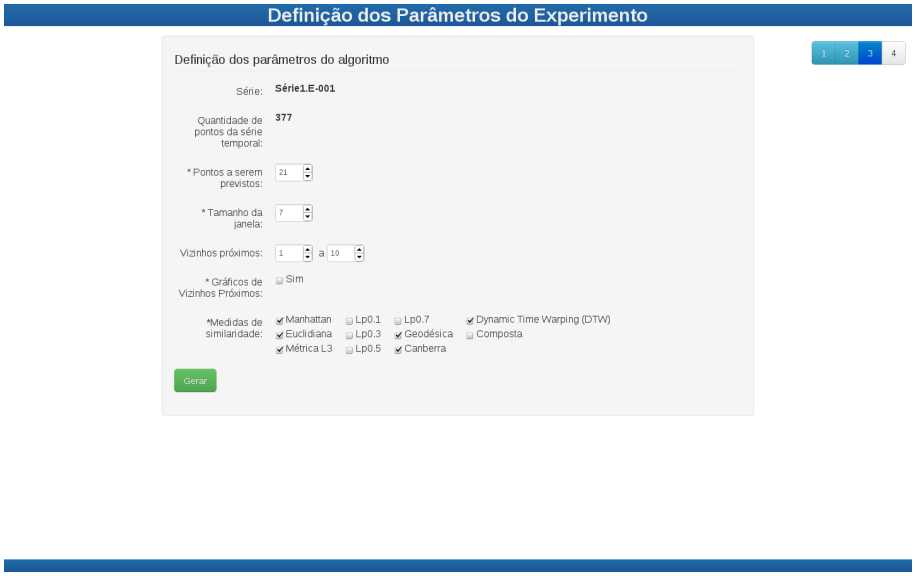
5.4.1 Séries Artificiais

Na Tabela 5.5 são apresentados os valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries artificiais, agrupando todos os resultados para um, cinco e dez vizinhos mais próximos, conforme mencionado na Seção 5.3. Valores que apresentam sombreamento de célula em cor verde representam o menor valor na respectiva linha, e valores com sombreamento de célula em cor vermelha, o maior valor da linha. Os tons entre as cores verde e vermelha

[5]http://www.neural-forecasting-competition.com/



(a)



(b)

Figura 5.4: Telas do protótipo de execução de experimentos de previsão de séries temporais. (a) Tela de envio da ST para o servidor e (b) definição dos parâmetros dos experimentos.

representam valores intermediários.

Nessa tabela, as medidas L_p Inteiras (Manhattan, Euclidiana e Métrica L_3) apresentaram, em geral, os menores valores de *MAPE* para as séries artificiais. Para a série STS1, a Métrica L_3 apresentou 00,0044% de erro médio[6] enquanto que para a série STS2, 20,4702%. As distâncias Euclidiana e Métrica L_3 foram as duas com menores valores médios de erro para a série STS3, com 13,8087% e 13,9129% respectivamente. Para a série STC1, as medidas L_p Inteiras assumiram valores intermediários de erro médio. Para a série STC2, essas medidas apresentaram os três menores erros médios, tendo sido 00,5856% para a distância Euclidiana, 00,5895% para a

[6]Neste trabalho os termos *MAPE*, erro médio e média dos erros são usados indistintamente.

Tabela 5.5: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries artificiais, agrupando os valores de um, cinco e dez vizinhos mais próximos.

Artificiais											
Série	**Estatística**	**Manhattan**	**Euclidiana**	**Métrica L_3**	**$Lp_{0.1}$**	**$Lp_{0.3}$**	**$Lp_{0.5}$**	**$Lp_{0.7}$**	***DTW***	**Canberra**	**Geodésica**
STS1	**Média**	00,0069	00,0053	00,0044	00,0124	00,0104	00,0087	00,0077	00,0118	00,0356	00,0052
	Desv. Padrão	00,0200	00,0159	00,0119	00,0281	00,0257	00,0238	00,0224	00,0223	00,0601	00,0158
	Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000
	Máximo	00,1349	00,1284	00,0910	00,1476	00,1476	00,1476	00,1412	00,1148	00,4102	00,1284
STS2	**Média**	20,6876	20,9168	20,4702	21,0780	20,5555	20,7540	20,5822	37,6251	20,5802	20,4878
	Desv. Padrão	16,3812	15,8899	16,5301	17,3218	16,1841	15,9566	16,2992	27,4065	17,0013	17,1991
	Mínimo	00,6641	00,6885	00,6885	00,6885	00,6885	00,6885	00,6641	01,7333	00,6885	00,6885
	Máximo	75,2234	68,4527	70,2030	73,1782	71,5481	68,7250	75,2234	120,0507	79,2077	83,0236
STS3	**Média**	15,0882	13,8087	13,9129	15,4774	15,0226	14,9730	14,9922	60,0798	17,3565	18,8027
	Desv. Padrão	43,8497	35,7605	37,4132	44,6139	43,7673	43,7641	43,7503	125,5721	43,0448	46,0285
	Mínimo	00,0224	00,0471	00,0241	00,0415	00,0415	00,0224	00,0224	00,0621	00,0547	00,2338
	Máximo	356,5351	317,6741	317,6741	356,5351	356,5351	356,5351	356,5351	1133,0019	426,2583	426,0089
STC1	**Média**	11,5661	11,0808	11,0241	10,8067	11,9341	11,8993	11,8691	14,8320	12,8268	11,3049
	Desv. Padrão	23,9709	21,3158	21,1672	21,4404	25,9322	24,6246	24,6681	38,8731	18,1264	21,2057
	Mínimo	00,0770	00,0015	00,0643	00,0668	00,0668	00,0770	00,0770	00,0332	00,0719	00,0015
	Máximo	256,9434	171,9324	187,0545	172,5227	256,9434	256,9434	256,9434	419,8998	153,9658	187,0545
STC2	**Média**	00,6056	00,5856	00,5895	00,8950	00,7303	00,6660	00,6284	00,6179	00,8736	00,6605
	Desv. Padrão	00,8265	00,8826	00,9317	01,2130	00,9513	00,8485	00,8165	00,8431	01,1861	00,6915
	Mínimo	00,0007	00,0007	00,0032	00,0032	00,0032	00,0032	00,0032	00,0032	00,0032	00,0032
	Máximo	06,9189	08,9251	08,9251	08,1749	07,1601	06,4179	06,7833	06,3156	09,5243	05,4541

Métrica L_3 e 00,6056% para a distância Manhattan.

Quando avaliados os valores de k individualmente, isto é, quando observadas as tabelas de *MAPE* para um, cinco e dez vizinhos mais próximos, as medidas L_p Inteiras permaneceram, em geral, como as de menores valores de erro médio. Para $k = 10$, a Métrica L_3 apresentou os menores valores de *MAPE* para as séries STS1 (00,0107%), STS2 (29,1318%) e STS3 (13,4997%). Para a série STC1, essas medidas assumiram valores intermediários. Já para a série STC2, as medidas Euclidiana, Manhattan e Métrica L_3 foram, nessa ordem, as três com os menores valores de erro, com 00,7617%, 00,7770% e 00,7780%, respectivamente. Quando $k = 5$, as três medidas L_p Inteiras apresentaram o mesmo valor de erro médio para a série STS1, tendo sido esse o menor erro médio dessa série (00,0010%), as distâncias Manhattan (21,8642%) e Euclidiana (21,8931%) as duas de menor erro médio para a série STS2, bem como as distâncias Euclidiana (12,4930%) e Métrica L_3 (12,8972%) as duas de menor erro médio para a série STS3. A Métrica L_3 apresentou o menor *MAPE* para a série STC1, 11,9253%, e para a série STC2, as distâncias Euclidiana, Manhattan e Métrica L_3 foram, nessa ordem, as três de menor erro médio, tendo sido 00,5224%, 00,5559% e 00,5568%, respectivamente, sendo que essas medidas assumiram valores intermediários nos demais casos. Para $k = 1$, as três medidas L_p Inteiras apresentaram os mesmos valores de média de erro para as séries STS1 (00,0015%) e STS2 (09,6758%), sendo que esses valores foram os menores para essas séries. Para as demais séries essas medidas assumiram valores intermediários. Os resultados detalhados para todas as medidas, para todas as ST e cada valor de k analisados na avaliação experimental são apresen-

tados no Apêndice B.

As medidas L_p Fracionárias ($L_{p0.1}$, $L_{p0.3}$, $L_{p0.5}$ e $L_{p0.7}$), apresentaram maiores valores de erros médios do que as medidas L_p Inteiras, entretanto, geralmente abaixo de outras medidas como a *DTW* e a Canberra. Uma exceção ocorreu na série STC2, na qual a distância $L_{p0.1}$ apresentou o maior valor de erro médio (00,8950%) entre todas as medidas de similaridade para essa série. Esse padrão de valores de erro médio intermediários para as medidas L_p Fracionárias se repetiu para $k = 10$ e para $k = 5$, sendo que, para $k = 5$, a medida $L_{p0.5}$ apresentou o maior valor de erro médio para a série STC1 (13,5858%) e a $L_{p0.1}$ a maior média de erro para a série STC2 (00,8540%). Quando $k = 1$, as medidas L_p Fracionárias apresentaram os mesmos valores que as séries L_p Inteiras para as séries STS1 (00,0015%) e STS2 (09,6758%), sendo que esses foram os menores erros médios dessas séries, tendo permanecido como medidas intermediárias nas demais séries. Novamente, a exceção ocorreu com a medida $L_{p0.1}$, que apresentou o maior valor de *MAPE* para a série STC2 (00,6022%).

Os valores médios de erro obtidos com a distância *DTW* foram os mais altos para as séries STS2 (37,6251%), STS3 (60,0798%) e STC1 (14,8320%), sendo que na série STS3 chegou a ser mais que o dobro da distância que apresentou o segundo maior valor para essa série, e foi o maior erro médio entre todas as séries. Para as séries STS1 e STC2, essa distância apresentou valores intermediários. Quando $k = 10$, apenas para as séries STS2 e STS3, a *DTW* continuou a apresentar os valores de erro médio mais elevados, 39,5033% e 57,6690% respectivamente, sendo que para as demais apresentou valores intermediários. Esse comportamento é observado também para $k = 5$, na qual apresentou novamente para as séries STS2 e STS3 os maiores erros médios, tendo sido 35,3455% e 67,6721%, respectivamente. Para $k = 1$, verifica-se que a *DTW* apresentou os maiores valores de erros médios para as séries STS1 (0,0037%), STS2 (38,0265%), STS3 (54,8981%) e STC1 (20,5628%), tendo sido classificada apenas para a série STC2 como medida de erro médio intermediário.

A distância Canberra apresentou o maior erro médio para a série STS1 (00,0356%), sendo que para as demais séries apresentou valores intermediários. Para $k = 10$, essa distância permaneceu apresentando o maior valor de erro médio para a série STS1 (0,0754%), bem como para as séries STC1 (13,7677%) e STC2 (01,2811%), e intermediário para as demais ST. Para $k = 5$, novamente essa distância foi a de maior valor de erro médio apenas para a série STS1 (00,0297%). Os valores de erros médios da distância Canberra para $k = 1$ foram intermediários para todas as séries, exceto para a série STS2, onde foi o mesmo das distâncias L_p Inteiras e L_p Fracionárias (09,6758%), sendo que esse foi o menor erro médio para essa série.

Para a média dos três valores de k agrupados, a distância Geodésica apresentou valores intermediários de erro médio para todas as séries. Esse mesmo comportamento dessa distância pôde ser observado quando assumiu-se $k = 10$. Quando $k = 5$, essa distância apresentou, para a série STS1, o mesmo valor de *MAPE* das medidas L_p Inteiras (00,0010%), tendo sido esse o

menor valor de média de erro para essa série, e para as demais séries, valores intermediários. Para $k = 1$, novamente a distância Geodésica apresentou os mesmos valores de *MAPE* que as medidas L_p Inteiras (00,0015%), tendo sido esse o menor erro médio. O mesmo valor de erro médio das medidas L_p Inteiras, L_p Fracionárias e Canberra foi obtido com a distância Geodésica para a série STS2 (09,6758%), sendo que novamente esse foi o menor valor para essa série. Para as séries STS3 e STC1, essa medida apresentou valores intermediários de erro médio, sendo que para a série STC2, apresentou o menor valor de *MAPE* (00,3966%).

Na Figura 5.5 são apresentadas a variação da média, em forma de pontos, e do desvio padrão, representado pela área sombreada, do *MAPE* das medidas de similaridades avaliadas para as séries artificiais. As medidas Manhattan, Euclidiana e Métrica L_3 foram agrupadas em L_p Inteiras e as medidas $L_{p0.1}$, $L_{p0.3}$, $L_{p0.5}$ e $L_{p0.7}$ foram agrupadas em L_p Fracionárias.

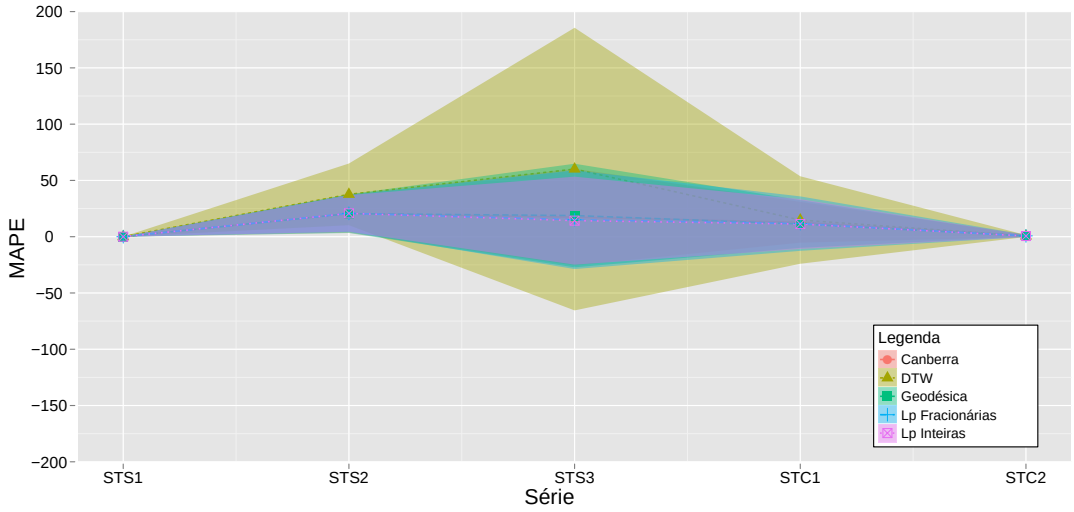


Figura 5.5: Média e desvio padrão de *MAPE* das séries artificiais para todas as medidas avaliadas.

Nessa figura, é possível observar que as medidas L_p Inteiras apresentaram valores médios de *MAPE* e desvio padrão abaixo ao das demais medidas para as séries artificiais. A segunda medida que apresentou os menores valores de erro médio e desvio padrão foi a medida Canberra. As medidas L_p Fracionárias apresentaram o terceiro menor valor de erro médio e de desvio padrão seguidas pela medida Geodésica. A medida *DTW* apresentou valores médios de *MAPE* e de desvio padrão expressivamente maiores que os das demais medidas para a maioria das séries.

É possível observar ainda, nessa figura, que a série STS3 foi a que apresentou os maiores valores de desvio padrão, especialmente para a medida *DTW*. Isso evidencia que todas as medidas avaliadas apresentaram a maior variabilidade na precisão da previsão na série que possui característica de tendência linear e sazonalidade multiplicativa, apontando essa combinação de característica como a de maior dificuldade de previsão entre as ST artificiais.

Já para a série STS1, que apresenta tendência linear e sazonalidade constante, verificou-se os menores valores de *MAPE* e de desvio padrão para todas as medidas. Isso reflete a maior facilidade de previsão de séries com essas características. Os valores de erros médios para a série STC2 também se apresentaram inferiores a 1%, com valores de desvio padrão pequenos para todas as medidas. Apesar da característica caótica dessa série, com os parâmetros utilizados, observa-se uma certa regularidade na faixa de valores bem como na sazonalidade da série (Figura 5.2), o que favorece o algoritmo *kNN-TSP* utilizando a MVR como função de previsão para todas as medidas.

Como mencionado, para verificar a existência de **d.e.s** entre os valores de *MAPE* das medidas avaliadas, os resultados foram analisados utilizando o teste estatístico não-paramétrico de Friedman para dados emparelhados e comparações múltiplas, considerando nível de significância de 5% ($p\text{-}valor < 0,05$), com pós-teste de Dunn (Motulsky, 1995).

Na Tabela 5.6 é apresentado o resultado da aplicação dos testes estatísticos para as séries artificiais. Nessa tabela:

*****:** Representa a existência de **d.e.s** com $p\text{-}valor < 0,001$;

****:** Representa a existência de **d.e.s** com $p\text{-}valor < 0,01$;

***:** Representa a existência de **d.e.s** com $p\text{-}valor < 0,05$;

Célula em Branco: Representa que não foi possível constatar a existência de **d.e.s**;

t.d.f: Representa o total de **d.e.s** favorável (t.d.f) para a medida da respectiva linha;

t.d.d: Representa o total de **d.e.s** desfavorável (t.d.d) para a medida da respectiva coluna.

Assim, a célula preenchida representa a existência de **d.e.s** entre a medida da linha e da coluna de encontro da célula, sendo que a medida da linha apresentou menores valores de *MAPE*. Assim, a notação *** na célula de encontro entre a linha da medida Manhattan e a coluna da medida $L_{p0.1}$, representa a existência de **d.e.s** com $p\text{-}valor < 0,001$ entre essas medidas, sendo que a distância Manhattan apresentou um valor médio de *MAPE* menor que a $L_{p0.1}$.

Observa-se, nessa tabela, que as medidas L_p Inteiras apresentaram **d.e.s** para com a maioria das medidas, sendo que suas médias foram sempre inferiores às das demais, em especial a Métrica L_3 que apresentou **d.e.s** favorável para com todas as medidas não pertencentes ao grupo das L_p Inteiras. A distância Euclidiana apresentou **d.e.s** favorável com todas as medidas não pertencentes ao grupo das L_p Inteiras, exceto $L_{p0.7}$; e a distância Manhattan apresentou **d.e.s** favorável para com as medidas $L_{p0.1}$, *DTW*, Canberra e Geodésica.

Tabela 5.6: Comparativo sobre a existência de **d.e.s** entre as medidas de similaridade para as séries artificiais.

Artificiais											
	Manhattan	**Euclidiana**	**Métrica L_3**	$L_{p0.1}$	$L_{p0.3}$	$L_{p0.5}$	$L_{p0.7}$	***DTW***	**Canberra**	**Geodésica**	**t.d.f**
Manhattan	——			***				***	***	**	**4**
Euclidiana		——		***	***	**		***	***	***	**6**
Métrica L_3			——	***	***	***	**	***	***	***	**7**
$L_{p0.1}$				——		*	***	***	***		**4**
$L_{p0.3}$					——			***	***		**2**
$L_{p0.5}$						——		***	***		**2**
$L_{p0.7}$							——	***	***		**2**
DTW								——			**0**
Canberra								***	——		**1**
Geodésica								***	***	——	**2**
t.d.d	**0**	**0**	**0**	**3**	**2**	**3**	**2**	**9**	**8**	**3**	——

Pode ser observado ainda que não foi possível constatar existência de **d.e.s** entre as medidas L_p Inteiras, indicando que não há evidência estatística que comprove melhor qualidade em termos de exatidão de alguma delas. Assim, devido ao menor custo computacional, a distância Manhattan pode ser determinada como a melhor escolha entre essas três. Esse resultado corrobora com o alcançado em Aikes Junior et al. (2011) para essas mesmas séries, considerando valores de k variando entre um e cinco.

O custo computacional relacionado à complexidade para se calcular cada medida, neste caso, é um dos fatores que exercem influência no tempo de execução do algoritmo. Como exemplo dessa influência, para a série STS1 utilizando dez vizinhos mais próximos, os tempos de execução do algoritmo *kNN-TSP* para as medida L_p Inteiras foram: 263,432 segundos para a distância Manhattan; 263,889 segundos para a distância Euclidiana e 271,521 segundos para a Métrica L_3. É interessante observar que em séries que apresentam maior quantidade de observações, essa diferença entre os tempos de execução pode ser maior.

Todas as medidas L_p Fracionárias apresentaram **d.e.s** favorável apenas com as medidas *DTW* e Canberra. Foi possível constatar também **d.e.s** favorável da medida $L_{p0.1}$ para com as medidas $L_{p0.5}$ e $L_{p0.7}$. Dessa maneira, considerando apenas as medidas L_p Fracionárias, a medida $L_{p0.1}$ apresentou-se, neste trabalho, como sendo a mais adequada para as ST artificiais.

As medidas *DTW* e Canberra apresentaram, para as séries artificiais, **d.e.s** desfavorável para com todas as demais medidas avaliadas, sendo que, entre as duas, a medida Canberra apresentou **d.e.s** favorável. Assim, para as séries artificiais, essas medidas mostraram-se, neste trabalho, como as menos adequadas para as séries artificiais. Para a medida Geodésica foi constatada **d.e.s** desfavorável para com as medidas L_p Inteiras e favorável apenas para com as medidas *DTW* e Canberra.

É possível observar ainda, que a medida Métrica L_3 apresentou a maior quantidade de **d.e.s** favoráveis entre todas as medidas, tendo sido sete seu t.d.f, sendo seguida pela distância

Euclidiana com seis **d.e.s** favoráveis. A distância Manhattan e $L_{p0.1}$ apresentaram o mesmo t.d.f, quatro, tendo sido esse o terceiro maior. Isso evidência, neste trabalho, uma possível vantagem das medidas L_p Inteiras sobre as demais medidas avaliadas nas séries artificiais.

Por outro lado, a medida *DTW* foi a única em que foi constatada **d.e.s** desfavorável para com todas as medidas avaliadas, o que demonstra sua possível desvantagem quando comparada às demais medidas avaliadas. A medida Canberra apresentou t.d.d igual a oito, não tendo sido evidenciada **d.e.s** desfavorável apenas para com a medida *DTW*. Já as três medidas L_p Inteiras, foram as únicas que não apresentaram **d.e.s** desfavoráveis.

Na Tabela 5.7 são apresentadas as diferenças entre o t.d.f e o t.d.d das medidas de similaridade avaliadas para as séries artificiais. A utilização dessa diferença objetiva sintetizar os resultados, de maneira a se ter uma visão global dos resultados de maneira simplificada. Assim, quanto maior essa diferença, mais ocorrências de **d.e.s** favoráveis a medida apresentou em relação às demais.

Tabela 5.7: Diferença entre t.d.f e t.d.d para as séries artificiais.

	Manhattan	**Euclidiana**	**Métrica L_3**	$L_{p0.1}$	$L_{p0.3}$	$L_{p0.5}$	$L_{p0.7}$	***DTW***	**Canberra**	**Geodésica**
(t.d.f-t.d.d)	4	6	7	1	0	-1	0	-9	-7	-1

Nessa tabela, é possível observar que as medidas L_p Inteiras apresentaram as maiores diferenças positivas entre t.d.f e t.d.d, em especial as medidas Métrica L_3 e Euclidiana, apontando possível vantagem dessas medidas sobre as demais. Já as medidas *DTW* e Canberra foram as que apresentaram as diferenças com os menores valores, indicando possível desvantagem para as demais medidas.

Apesar da medida Métrica L_3 ter apresentado uma diferença positiva maior que as demais medidas L_p Inteiras, cabe relembrar que não foi possível constatar **d.e.s** entre essas medidas (Tabela 5.6). Assim, a medida Manhattan, apesar de ter apresentado uma diferença entre t.d.f e t.d.d menor, pode ser vantajosa devido ao seu menor custo computacional.

5.4.2 Séries Reais

Na Tabela 5.8 são apresentados os valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries da *NN GC1* agrupando todos os resultados para um, cinco e dez vizinhos mais próximos, de modo similar à Tabela 5.5. Para facilitar a análise, as séries foram agrupadas de acordo com sua frequência de aquisição (Tabela 5.2), ou seja, os resultados estão agrupados por base de ST. Valores que apresentam sombreamento de célula em cor verde representam o menor valor na respectiva linha, e valores com sombreamento de célula em cor vermelha, o maior valor da linha. Os tons entre as cores verde e vermelha representam valores intermediários.

Tabela 5.8: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries da *NN GC1*, agrupando os valores de um, cinco e dez vizinhos mais próximos.

		Reais									
Base	**Cálculo**	**Manhattan**	**Euclidiana**	**Métrica L_3**	**$Lp_{0.1}$**	**$Lp_{0.3}$**	**$Lp_{0.5}$**	**$Lp_{0.7}$**	***DTW***	**Canberra**	**Geodésica**
1.B	**Média**	04,3305	04,3754	04,3472	04,6831	04,4638	04,3348	04,3491	04,4130	04,4295	04,3553
	Desv. Padrão	03,3818	03,5765	03,5881	03,6656	03,6290	03,3845	03,3857	03,5387	03,5808	03,4928
	Mínimo	00,0000	00,0000	00,0000	00,0291	00,0358	00,0000	00,0000	00,0000	00,0000	00,0000
	Máximo	17,6832	18,2254	18,2254	21,5812	21,5812	17,6832	17,6832	18,4245	18,1825	20,0571
1.C	**Média**	09,3340	09,2817	09,0492	11,0251	10,4069	10,1055	09,8027	11,6188	11,5049	09,1323
	Desv. Padrão	10,3900	10,1093	09,7311	13,9463	12,2154	11,7431	11,4373	13,0288	13,1303	10,1515
	Mínimo	00,0027	00,0272	00,0287	00,0127	00,0027	00,0027	00,0027	00,0021	00,0000	00,0038
	Máximo	57,0833	57,0833	49,0141	127,3973	82,4205	82,4205	82,4205	82,4205	125,9717	61,9282
1.D	**Média**	11,0933	10,8341	10,9914	11,5462	11,2017	10,6916	10,7640	12,5904	11,9959	10,5170
	Desv. Padrão	22,6905	23,2732	22,9447	23,9333	22,4598	21,4961	21,6027	25,7694	25,1886	21,0774
	Mínimo	00,0213	00,0000	00,0000	00,0000	00,0054	00,0000	00,0175	00,0040	00,0000	00,0000
	Máximo	181,9672	172,5000	174,4643	273,2558	181,9672	181,9672	181,9672	234,4398	182,1429	168,1034
1.E	**Média**	35,0619	29,1431	27,2483	36,1532	36,4985	37,0178	36,0344	29,9580	43,6634	54,7279
	Desv. Padrão	127,1498	85,1116	74,2007	144,9975	131,8605	138,5429	132,8563	86,1792	153,5974	134,2987
	Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0152	00,0000
	Máximo	1600,0000	950,0000	810,0000	2000,0000	1420,0000	1600,0000	1600,0000	960,0000	1600,0000	850,0000
1.F	**Média**	21,5107	18,6968	17,9128	23,2658	23,1804	22,9243	22,3257	22,5253	16,5189	11,4941
	Desv. Padrão	65,3133	54,0929	49,6660	73,4028	74,4874	73,5961	68,5316	36,8095	47,4482	23,3068
	Mínimo	00,0000	00,0000	00,0071	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000
	Máximo	655,0980	655,0980	602,2274	1070,7317	1070,7317	1070,7317	655,0980	279,1304	613,6577	246,2036

Nessa tabela, pode ser observado que as medidas L_p Inteiras encontraram-se novamente ou como as medidas com os menores valores de *MAPE*, caso da base 1.B, na qual a distância Manhattan apresentou 04,3305% de erro médio, da base 1.C, na qual a Métrica L_3 apresentou 09,0492% e da base 1.E, na qual a Métrica L_3 apresentou 27,2483%, ou como medidas intermediárias, casos da base 1.D e 1.F.

Quando avaliados os erros médios para os valores de $k = 1$, 5 e 10 individualmente, foi possível verificar novamente que as medidas L_p Inteiras permaneceram sendo as medidas de menor erro médio, ou nos piores casos, medidas intermediárias. Quando $k = 10$, as medidas Métrica L_3, Euclidiana e Manhattan foram, nessa ordem, as três medidas com os menores valores de erro médio da base 1.B, tendo alcançado 03,9998%, 04,0456% e 04,0987%, respectivamente, e para as demais bases essas medidas assumiram valores intermediários. Para $k = 5$, a Métrica L_3 e a Euclidiana apresentaram os dois menores valores médios de *MAPE* da base 1.B, 03,9896% e 04,1821%, respectivamente, e a Métrica L_3 foi a de menor média de erro da base 1.E (28,0565%), sendo que para os demais casos essas medidas apresentaram valores intermediários. Assumindo $k = 1$, a Métrica L_3 foi a de menor erro médio para a base 1.C (09,0316%), e a Métrica L_3 e a distância Euclidiana foram as duas medidas de menor erro médio da base 1.E, 26,7985% e 27,4373%, respectivamente. Nos demais casos, as medidas L_p Inteiras apresentaram valores intermediários de erro médio. Os resultados detalhados para todas as medidas, para todas as ST e cada valor de k analisados na avaliação experimental são apresentados no

As medidas L_p Fracionárias apresentaram, em geral, valores de erro médio intermediários para a maioria das bases. Exceções ocorreram na base 1.B para a medida $L_{p0.1}$, que apresentou o maior valor de *MAPE* para essa base, 04,6831%, e na base 1.F, na qual as medidas $L_{p0.1}$, $L_{p0.3}$ e $L_{p0.5}$ apresentaram os três maiores erros médios, 23,2658%, 23,1804% e 22,9243%, respectivamente.

Ainda para as medidas L_p Fracionárias, assumindo $k = 10$, essas medidas alcançaram novamente valores de erro médio intermediários, exceto a medida $L_{p0.3}$ para a base 1.B, que apresentou o maior valor de média de erro dessa base (04,2273%) e para a medida $L_{p0.1}$, que apresentou o menor valor de erro médio para a base 1.E (19,7046%). Quando $k = 5$, a medida $L_{p0.1}$ possui o maior erro médio para a base 1.B (04,5679%), a $L_{p0.7}$ o menor erro médio para a base 1.C (08,7069%) e as medidas $L_{p0.1}$, $L_{p0.3}$, $L_{p0.5}$ e $L_{p0.7}$ foram, na média, as quatro medidas que apresentaram os maiores valores de erro médio da base 1.F com 23,4491%, 23,3005%, 22,8535% e 22,8453%, respectivamente. Nos demais casos, as medidas L_p Fracionárias apresentaram valores intermediários para esse valor de k. Para $k = 1$, novamente a medida $L_{p0.1}$ foi a que apresentou o maior erro médio para a base 1.B, tendo sido 05,3490%, entretanto, as medidas $L_{p0.5}$ e $L_{p0.7}$ foram as duas com os menores valores de *MAPE* nessa base, com 04,5038% e 04,5484%, respectivamente. Para a base 1.D, as medidas $L_{p0.7}$ e $L_{p0.5}$ foram as que possuíram o menor valor de erro médio, 11,5568% e 11,7749%, respectivamente. Já para a base 1.F, as medidas $L_{p0.1}$, $L_{p0.3}$ e $L_{p0.5}$ foram as três medidas de maior valor de *MAPE*, com 22,2511%, 22,1089% e 21,9252%, respectivamente.

A medida *DTW* apresentou valores de *MAPE* intermediários para as bases 1.B, 1.E e 1.F, sendo que para as bases 1.C e 1.D ela apresentou os maiores valores de erro médio, com 11,6188% e 12,5904%, respectivamente. Para $k = 10$, essa medida apresentou novamente valores de erro médio intermediários, exceto para a base 1.D, na qual seu valor de *MAPE* foi o mais elevado com 10,6857%, e para a base 1.F, na qual também apresentou o maior valor de erro médio com 25,1576%. Assumindo $k = 5$, a distância *DTW* apresentou o maior valor de média de erro apenas para a base 1.D (11,2532%), sendo que para as demais bases apresentou valores de erro médio intermediários. Já para $k = 1$, essa distância apresentou os maiores valores de erro médio para as bases 1.C e 1.D, tendo sido 14,8719% e 15,8322%.

Para a média assumindo os pontos de todos os valores de k avaliados neste trabalho, a medida Canberra apresentou valores intermediários de erro médio para todas as bases de ST da *NN GC1*. Quando $k = 10$, essa distância apresentou os maiores valores de erro médio para as bases 1.B e 1.C com 04,2339% e 10,8287%, respectivamente, tendo se mantido como medida intermediária para as demais bases. Já para $k = 5$, a Canberra foi a medida com o maior valor de erro médio apenas para a base 1.C (10,6589%), tendo sido novamente uma medida intermediária para as demais bases. Assumindo $k = 1$, essa distância novamente apresentou valores de erro

médio intermediários para todas as bases.

A distância Geodésica apresentou o menor valor médio de *MAPE* para as bases 1.D e 1.F, tendo sido 10,5170% e 11,4941%, respectivamente. Já para a base 1.E, essa medida apresentou o maior erro médio com 54,7279%, e para as demais bases se apresentou como medida intermediária. Quando $k = 10$, a distância Geodésica apresentou os menores valores de erro médio para as bases 1.C, 1.D e 1.F, com 08,9714%, 09,9163% e 11,0491%, respectivamente. Novamente apresentou para a base 1.E o maior erro médio (53,3438%) e para a base 1.B apresentou erro médio intermediário. Assumindo $k = 5$, essa distância apresentou os menores valores de *MAPE* para as bases 1.D (09,1940%) e 1.F (11,0063%), o maior para a base 1.E (51,0474%) e intermediários para as demais bases. Para $k = 1$, essa distância apresentou valores de erros médios intermediários para as bases 1.B, 1.C e 1.D, tendo sido novamente a de maior erro médio para a base 1.E (59,7926%) e a de menor erro médio para a base 1.F, com 12,4269%.

Na Figura 5.6 são apresentadas a variação da média, em forma de pontos, e do desvio padrão, representado pela área sombreada, do *MAPE* das medidas de similaridades avaliadas para as séries da *NN GC1*. As medidas Manhattan, Euclidiana e Métrica L_3 foram agrupadas em L_p Inteiras e as medidas $L_{p0.1}$, $L_{p0.3}$, $L_{p0.5}$ e $L_{p0.7}$ foram agrupadas em L_p Fracionárias.

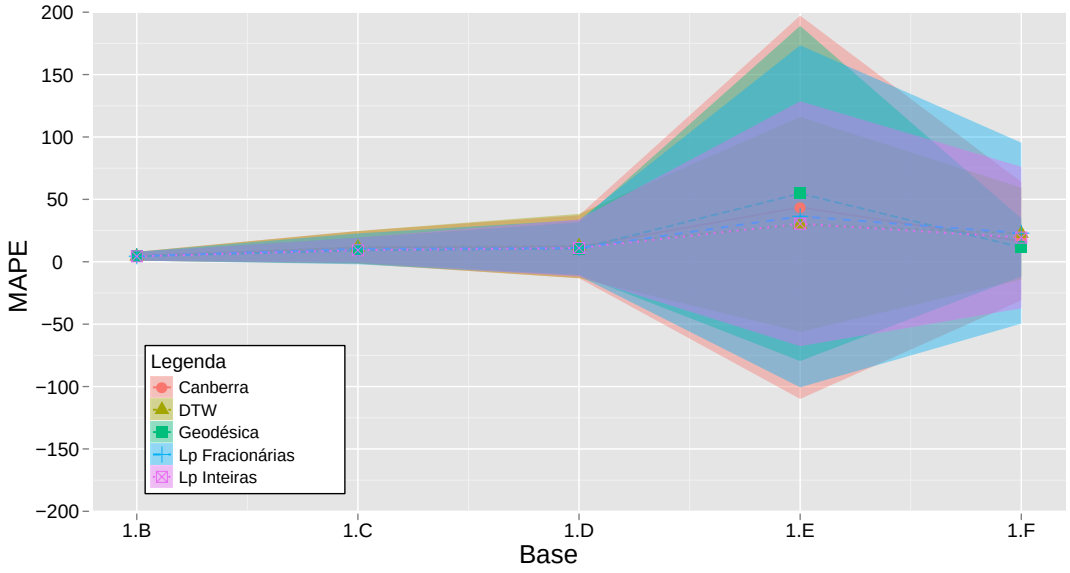


Figura 5.6: Média e desvio padrão para todas as medidas avaliadas e séries da *NN GC1*.

Nessa figura, é possível observar que as medidas L_p Inteiras apresentaram, em geral, os menores valores médios de *MAPE* e de desvio padrão para a maioria das bases de ST da *NN GC1* avaliadas. A medida Geodésica apresentou o segundo menor valor de erro médio, e o terceiro menor desvio padrão médio, ocorrendo exceção na base 1.E, na qual essa medida apresentou o maior erro médio e o segundo maior desvio padrão. A medida *DTW* apresentou a segunda maior média de erro, entretanto, para essas séries ela apresentou o menor desvio padrão. Já as medidas L_p Fracionárias apresentaram tanto os maiores valores de erros médios como de desvio

padrão.

A medida Canberra apresentou o terceiro menor erro médio, entretanto, devido ao fato de ter apresentado o maior desvio padrão para a base 1.E, seu desvio padrão acabou por ser o segundo maior para as séries da *NN GC1*. É possível observar ainda que a única base em que essa medida apresentou *MAPE* consideravelmente baixo, inclusive inferior aos das medidas L_p Inteiras, foi na base 1.F. Dessa forma, a característica de normalização da medida Canberra pode ser vantajosa nas situações de ST semelhantes às que compõem a base 1.F, as quais apresentam quantidade considerável de observações (nessa base todas as séries possuem mais de 900 observações), sazonalidade e tendência suave ou nula. Entretanto, essa medida possivelmente não apresentou vantagem para as demais situações.

É possível observar ainda nessa figura, que todas as medidas apresentaram os maiores valores médios de *MAPE* e desvio padrão para a base 1.E. Esse comportamento deve-se às características das séries dessa base, que apresentaram a sazonalidade e a tendência mais irregulares entre todas as bases. Destaca-se ainda, que as medidas L_p Inteiras e *DTW* foram as que apresentaram os menores erros médios e desvios padrão médios para essa base, demonstrando uma possível vantagem na utilização dessas medidas em situações de ST irregulares.

Na Tabela 5.9 é apresentado o resultado da aplicação dos testes estatísticos para as séries da *NN GC1*. Assim como na Tabela 5.6, nessa tabela:

*****:** Representa a existência de **d.e.s** com $p\text{-}valor < 0,001$;

****:** Representa a existência de **d.e.s** com $p\text{-}valor < 0,01$;

***:** Representa a existência de **d.e.s** com $p\text{-}valor < 0,05$;

Célula em Branco: Representa que não foi possível constatar a existência de **d.e.s**;

t.d.f: Representa o total de **d.e.s** favorável (t.d.f) para a medida da respectiva linha;

t.d.d: Representa o total de **d.e.s** desfavorável (t.d.d) para a medida da respectiva coluna.

Desse modo, a notação *** na célula de encontro entre a linha da medida Manhattan e a coluna da medida *DTW*, representa a existência de **d.e.s** com $p\text{-}valor < 0,001$ entre essas medidas, sendo que a distância Manhattan apresentou um valor médio de *MAPE* menor que a *DTW*.

É possível observar, nessa tabela, que não foi possível constatar **d.e.s** entre as medidas L_p Inteiras. As medidas Euclidiana e Métrica L_3 apresentaram **d.e.s** favorável para as medidas $L_{p0.1}$, $L_{p0.3}$, *DTW* e Canberra, e a distância Manhattan apenas para com a *DTW* e a Canberra. Dentre as medidas L_p Inteiras, a distância Manhattan foi a única a ter apresentado **d.e.s** desfavorável, sendo que essa diferença foi com a medida Canberra.

Tabela 5.9: Comparativo sobre a existência de **d.e.s** entre as medidas de similaridade para as séries da *NN GC1*.

Reais											
	Manhattan	**Euclidiana**	**Métrica L_3**	$L_{p0.1}$	$L_{p0.3}$	$L_{p0.5}$	$L_{p0.7}$	***DTW***	**Canberra**	**Geodésica**	**t.d.f**
Manhattan	——							***			**1**
Euclidiana		——		***	**			***	***		**4**
Métrica L_3			——	***	***			***	***		**4**
$L_{p0.1}$				——							**0**
$L_{p0.3}$					——						**0**
$L_{p0.5}$						——					**0**
$L_{p0.7}$							——				**0**
DTW				***	***	***	***	——			**4**
Canberra	***					**	**	**	——		**4**
Geodésica								***	**	——	**2**
t.d.d	**1**	**0**	**0**	**3**	**3**	**2**	**2**	**5**	**3**	**0**	——

As medidas L_p Fracionárias não apresentaram **d.e.s** favorável para com nenhuma das demais medidas. Elas exibiram ainda médias de erro superiores às demais, sugerindo assim ser desfavorável a sua utilização. Diferentemente do que ocorreu nas séries artificiais, não foi possível constatar **d.e.s** entre essas medidas nas séries da *NN GC1*.

A medida *DTW* apresentou **d.e.s** favorável para com as medidas L_p Fracionárias, ao contrário do que ocorreu nas séries artificiais, onde todas as medidas se apresentaram superiores à *DTW*. Além das medidas L_p Inteiras, a *DTW* apresentou **d.e.s** desfavorável também para as medidas Canberra e Geodésica.

A medida Canberra apresentou **d.e.s** favorável para a distância Manhattan e desfavorável para as demais medidas L_p Inteiras. Exibiu **d.e.s** favorável também para as medidas $L_{p0.5}$ e $L_{p0.7}$, bem como para a medida *DTW*. Assim, nas séries da *NN GC1*, para a Canberra foi constatada menor quantidade de **d.e.s** desfavorável quando comparadas com as séries artificiais.

A medida Geodésica apresentou **d.e.s** favorável para as medidas *DTW* e Canberra, da mesma forma ocorrida nas séries artificiais. Entretanto, ao contrário do que ocorreu nas séries artificiais, nas séries da *NN GC1* não foi possível constatar **d.e.s** desfavorável para a medida Geodésica.

Ainda nessa tabela, é possível observar que as medidas Métrica L_3, Euclidiana, *DTW* e Canberra apresentaram o mesmo t.d.f, quatro, tendo sido essa a maior quantidade de **d.e.s** favorável. Já as quatro medidas L_p Fracionárias foram as únicas a não ter apresentado nenhuma **d.e.s** favorável. Isso aponta possível vantagem, nas séries da *NN GC1*, para as medidas L_p Inteiras, juntamente com a *DTW* e Canberra, e possível desvantagem das medidas L_p Fracionárias.

Observando o t.d.d, verifica-se que a medida *DTW* apresentou a maior quantidade de **d.e.s** desfavorável entre todas as medidas avaliadas, seguida das medidas Canberra e L_p Fracionárias. Já as medidas Euclidiana, Métrica L_3 e Geodésica não apresentaram nenhuma **d.e.s**

desfavorável.

Desse modo, verifica-se a possível vantagem das medidas L_p Inteiras, em especial as medidas Euclidiana e Métrica L_3, por apresentarem os maiores valores de t.d.f, juntamento com a *DTW* e a Canberra, porém sem ter apresentado t.d.d, como ocorreu com as duas últimas. Já as medidas L_p Fracionárias podem ser desfavoráveis, devido a não ter apresentado **d.e.s** favorável para com nenhuma medida, e ter apresentado **d.e.s** desfavorável com as medidas Euclidiana, Métrica L_3, *DTW* e Canberra.

Na Tabela 5.10 são apresentadas as diferenças entre o t.d.f e o t.d.d das medidas de similaridade avaliadas para as séries da *NN GC1*.

Tabela 5.10: Diferença entre t.d.f e t.d.d para as séries reais.

	Manhattan	**Euclidiana**	**Métrica L_3**	$L_{p0.1}$	$L_{p0.3}$	$L_{p0.5}$	$L_{p0.7}$	***DTW***	**Canberra**	**Geodésica**
(t.d.f-t.d.d)	0	4	4	-3	-3	-2	-2	-1	1	2

Nessa tabela, pode-se observar novamente a vantagem das medidas L_p Inteiras, tendo apresentado para as medidas Euclidiana e Métrica L_3 as maiores diferenças positivas entre t.d.f e t.d.d. Além dessas medidas, apenas a Canberra e a Geodésica apresentaram diferenças positivas. Já as medidas L_p Fracionárias apresentaram os menores valores de diferenças entre t.d.f e t.d.d, seguidas pela medida *DTW*, indicando sua possível desvantagem para com as demais medidas avaliadas.

5.4.3 Comparação Geral

Na Tabela 5.11 são apresentados os valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries artificiais e as séries da *NN GC1* agrupando todos os resultados para um, cinco e dez vizinhos mais próximos. De maneira a facilitar a análise, as séries foram agrupadas em séries artificiais e reais (*NN GC1*), isto é, todos os resultados das séries artificiais estão em um grupo e todos os resultados das séries da *NN GC1* em outro. Valores que apresentaram sombreamento de célula em cor verde representam o menor valor na respectiva linha, e valores com sombreamento de célula em cor vermelha, o maior valor da linha. Os tons entre as cores verde e vermelha representam valores intermediários.

É possível observar, nessa tabela, que para ambos os conjuntos de séries, as medidas L_p Inteiras apresentaram, em geral, os menores valores de *MAPE*. Para as séries artificiais, as medidas Métrica L_3, Euclidiana e Manhattan apresentaram os três menores erros médios, 06,6233%, 06,6926% e 06,8722%, respectivamente. Para as séries da *NN GC1*, a Métrica L_3 e a distância Euclidiana apresentaram os dois menores valores de média de erro, tendo sido 15,1976% e 15,7644%, respectivamente.

Tabela 5.11: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para o agrupamento das séries artificiais e das séries da *NN GC1*, incluindo os resultados de um, cinco e dez vizinhos mais próximos.

Série	Estatística	Manhattan	Euclidiana	Métrica L_3	$Lp_{0.1}$	$Lp_{0.3}$	$Lp_{0.5}$	$Lp_{0.7}$	*DTW*	Canberra	Geodésica
Artificiais	Média	06,8722	06,6926	06,6233	06,8899	06,9340	06,9420	06,9045	14,4991	07,3598	07,1665
	Desv. Padrão	19,9072	17,5594	17,8972	19,7006	20,2889	19,9898	20,0224	48,1651	18,8247	20,0816
	Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000
	Máximo	356,5351	317,6741	317,6741	356,5351	356,5351	356,5351	356,5351	1133,0019	426,2583	426,0089
Reais	Média	17,7744	15,7644	15,1976	18,9940	18,8238	18,6131	18,2175	18,1754	17,3730	16,1127
	Desv. Padrão	64,1890	48,5814	43,9081	72,4989	69,4313	70,7035	67,0249	42,2367	65,9875	54,5395
	Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000
	Máximo	1600,0000	950,0000	810,0000	2000,0000	1420,0000	1600,0000	1600,0000	960,0000	1600,0000	850,0000

Para $k = 10$, novamente a Métrica L_3 e a distância Euclidiana foram as duas medidas que obtiveram o menor erro médio para as séries artificiais, e a distância Manhattan assumiu valor intermediário. Para as séries da *NN GC1*, as medidas L_p Inteiras apresentaram valores intermediários de erro médio. Quando $k = 5$, para as séries artificiais, as medidas Euclidiana, Métrica L_3 e Manhattan foram, nessa ordem, as três que apresentaram os menores valores de erro médio. Para as séries da *NN GC1*, as medidas L_p Inteiras apresentaram valores intermediários. Para $k = 1$, essas medidas assumiram valores intermediários para as séries artificiais, e para as séries da *NN GC1* a Métrica L_3, Euclidiana e Manhattan, foram as três medidas de menor *MAPE*. Os resultados detalhados para todas as medidas, para todas as ST e cada valor de k analisados na avaliação experimental são apresentados no Apêndice B.

As medidas L_p Fracionárias apresentaram valores de erro médio intermediários para as séries artificiais. Já para as séries da *NN GC1*, as medidas $L_{p0.1}$, $L_{p0.3}$, $L_{p0.5}$ e $L_{p0.7}$ apresentaram os quatro maiores valores de *MAPE* dentre todas as medidas, tendo sido 18,9940%, 18,8238%, 18,6131% e 18,2175%, respectivamente. Quando $k = 10$, novamente essas medidas assumiram valores intermediários para as séries artificiais, enquanto que para as séries da *NN GC1* as medidas $L_{p0.3}$, $L_{p0.5}$ e $L_{p0.7}$ foram as três com os maiores valores de erro médio e a medida $L_{p0.1}$ apresentou erro médio intermediário. Quando $k = 5$, para as séries artificiais, as medidas L_p Fracionárias apresentaram valores intermediários de *MAPE* e para as séries da *NN GC1*, as medidas $L_{p0.3}$, $L_{p0.5}$, $L_{p0.7}$ e $L_{p0.1}$ foram, nessa ordem, as quatro com os maiores erros médios. Assumindo $k = 1$, as medidas L_p Fracionárias foram as quatro que alcançaram os menores valores de *MAPE* para as séries artificiais. Para as séries da *NN GC1*, a medida $L_{p0.1}$ foi a que apresentou o maior erro e as demais medidas L_p Fracionárias apresentaram erro médio intermediário.

A *DTW* apresentou, para as séries artificiais, maior valor de *MAPE*, tendo sido 14,4991% e para as séries da *NN GC1*, seu valor de erro médio foi intermediário. Quando avaliados individualmente os valores de $k = 1$, 5 e 10, essa medida foi novamente a de maior erro médio para as séries artificiais. Para as séries de *NN GC1*, essa medida apresentou erro médio intermediário

para todos os valores de k avaliados.

A medida Canberra apresentou valores intermediários de *MAPE*, tanto para as séries artificiais quanto para as séries da *NN GC1*. Assim como ocorreu com a medida *DTW*, a medida Canberra apresentou o mesmo padrão de erro médio para todos os valores de k avaliados neste trabalho, ou seja, para os valores de $k = 1, 5$ e 10, essa medida apresentou valores de erro médio intermediários para ambos os conjuntos de séries.

A medida Geodésica, assim como a Canberra, apresentou valores de erro médio intermediários para ambas as séries considerando todos os valores de k agrupados. Para $k = 10$, essa medida apresentou valor de erro médio intermediário para as séries artificiais e o menor valor para as séries da *NN GC1*. Quando $k = 5$, novamente para as séries artificiais o valor de *MAPE* da Geodésica foi intermediário, sendo que para as séries da *NN GC1* ela apresentou o menor valor. Para $k = 1$, essa medida apresentou novamente valores de erro médio intermediários para ambos os conjuntos de séries.

Na Figura 5.7 são apresentadas a variação da média, em forma de pontos, e do desvio padrão, representado pela área sombreada, do *MAPE* das medidas de similaridades avaliadas para as séries artificiais e para as séries da *NN GC1*. As medidas Manhattan, Euclidiana e Métrica L_3 foram agrupadas em L_p Inteiras e as medidas $L_{p0.1}$, $L_{p0.3}$, $L_{p0.5}$ e $L_{p0.7}$ foram agrupadas em L_p Fracionárias.

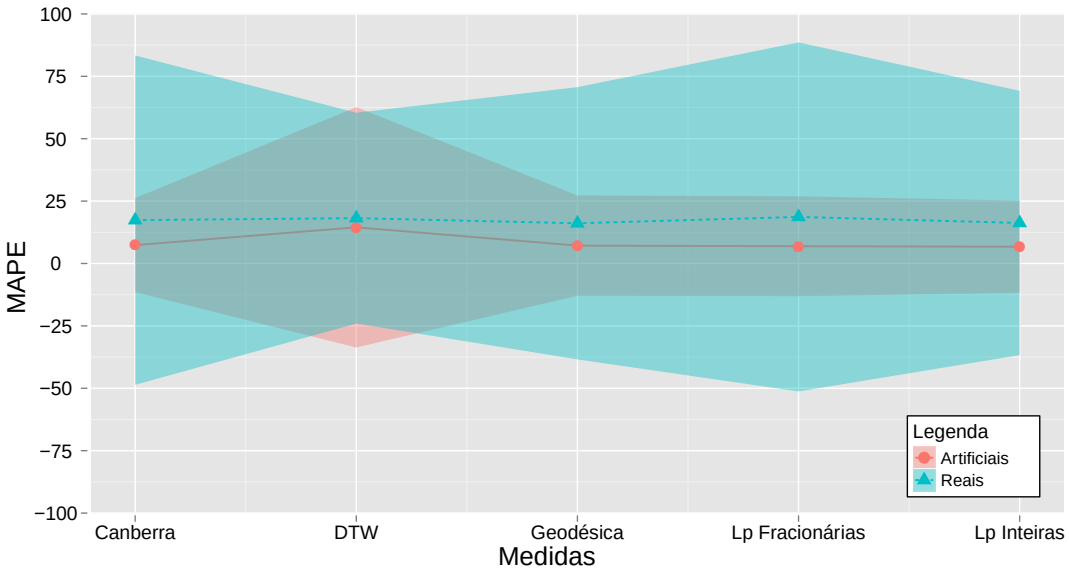


Figura 5.7: Média e desvio padrão para todas as medidas avaliadas agrupadas por séries artificiais e da *NN GC1*.

Nessa figura, é possível observar que para todas as medidas, tanto o valor médio de *MAPE* quanto o valor do desvio padrão para as séries artificiais foi menor do que para as séries da *NN GC1*, exceto para a medida *DTW*. Esse comportamento deve-se ao fato das séries da *NN GC1*, além de serem mais numerosas, apresentarem muito mais variabilidade de caracterís-

ticas do que as séries artificiais. Além disso, para as séries artificiais, é possível perceber que foi menor a diferença entre os resultados das medidas quando comparados às séries da *NN GC1*.

Destaca-se ainda nessa figura, o comportamento contrário do desvio padrão da medida *DTW* entre as séries artificiais e as ST da *NN GC1*. Nas séries artificiais, o desvio padrão dessa medida apresentou grande crescimento quando comparado às demais medidas, e nas séries da *NN GC1* tem-se o inverso. Isso pode indicar que, apesar do valor médio de erro da distância *DTW* não ter se apresentado entre os menores, essa medida pode ser preferencial em situações em que se buscam menor variabilidade dos resultados em séries reais.

Na Tabela 5.12 é apresentado o resultado da aplicação dos testes estatísticos para as séries artificiais e da *NN GC1*. Nessa tabela:

*****:** Representa a existência de **d.e.s** com $p\text{-}valor < 0,001$;

****:** Representa a existência de **d.e.s** com $p\text{-}valor < 0,01$;

***:** Representa a existência de **d.e.s** com $p\text{-}valor < 0,05$;

Célula em Branco: Representa que não foi possível constatar a existência de **d.e.s**;

t.d.f: Representa o total de **d.e.s** favorável (t.d.f) para a medida da respectiva linha;

t.d.d: Representa o total de **d.e.s** desfavorável (t.d.d) para a medida da respectiva coluna.

Assim, a notação *** na célula de encontro entre a linha da medida Manhattan e a coluna da medida $L_{p0.1}$, representa a existência de **d.e.s** com $p\text{-}valor < 0,001$ entre essas medidas, sendo que a distância Manhattan apresentou um valor médio de *MAPE* menor que a *DTW*.

Tabela 5.12: Comparativo sobre a existência de **d.e.s** entre as medidas de similaridade para as séries artificiais e para as séries da *NN GC1*.

	Manhattan	**Euclidiana**	**Métrica** L_3	$L_{p0.1}$	$L_{p0.3}$	$L_{p0.5}$	$L_{p0.7}$	***DTW***	**Canberra**	**Geodésica**	**t.d.f**
Manhattan	——			***	**			***			**3**
Euclidiana		——		***	***	**	*	***	***	***	**7**
Métrica L_3			——	***	***	**	*	***	***		**6**
$L_{p0.1}$				——				***			**1**
$L_{p0.3}$					——			***			**1**
$L_{p0.5}$				***		——		***			**2**
$L_{p0.7}$				***			——	***			**2**
DTW								——			**0**
Canberra	***			***	***	***	***		——		**5**
Geodésica			***	**				***	***	——	**4**
t.d.d	**1**	**0**	**1**	**7**	**4**	**3**	**3**	**8**	**3**	**1**	——

Nessa tabela, pode ser observado que novamente não foi possível constatar **d.e.s** nas medidas L_p Inteiras entre si, para os dados das séries artificiais e da *NN GC1* agrupados. A distância

Manhattan apresentou **d.e.s** favorável para com as medidas $L_{p0.1}$, $L_{p0.3}$ e *DTW*. A distância Euclidiana apresentou **d.e.s** favorável com as mesmas medidas que a distância Manhattan, acrescidas ainda da $L_{p0.5}$, $L_{p0.7}$, Canberra e Geodésica. Já a Métrica L_3 apresentou **d.e.s** favorável com todas as medidas L_p Fracionárias, *DTW* e Canberra. Foi constatada **d.e.s** desfavorável para a medida Manhattan em relação à Canberra e da Métrica L_3 em relação à Geodésica.

Todas as medidas L_p Fracionárias apresentaram **d.e.s** favorável em relação à medida *DTW*, sendo que a $L_{p0.1}$ também apresentou-se favorável à Geodésica. Para todas as medidas L_p Fracionárias foi possível constatar **d.e.s** desfavorável para com as medidas Euclidiana, Métrica L_3 e Canberra, sendo ainda que a $L_{p0.1}$ e a $L_{p0.5}$ apresentaram **d.e.s** desfavorável também em relação à distância Manhattan. A $L_{p0.1}$ também apresentou **d.e.s** desfavorável para as medidas $L_{p0.5}$, $L_{p0.7}$ e Geodésica.

A medida *DTW* não apresentou **d.e.s** favorável para com nenhuma medida, e desfavorável para todas as demais medidas exceto para com a Canberra.

Para a medida Canberra, foi possível constatar **d.e.s** favorável para com a distância Manhattan e para com todas as medidas L_p Fracionárias. Foi evidenciada **d.e.s** desfavorável para com as medidas Euclidiana, Métrica L_3 e Geodésica.

A medida Geodésica apresentou **d.e.s** favorável em relação à Métrica L_3, à $L_{p0.1}$, à *DTW* e à Canberra. Foi constatada **d.e.s** desfavorável da medida Geodésica apenas para com a medida Manhattan.

Nessa tabela, é possível ainda observar que a distância Euclidiana foi a que apresentou a maior quantidade de **d.e.s** favorável, seguida pela Métrica L_3. Já a medida *DTW* foi a única a não ter apresentado **d.e.s** favorável para com nenhuma das medidas avaliadas.

Observando o t.d.d, as medidas *DTW* e $L_{p0.1}$ apresentaram desvantagem, tendo sido as duas medidas que apresentaram a maior quantidade de **d.e.s** desfavorável. Já a distância Euclidiana foi a única a não ter apresentado **d.e.s** desfavorável.

Na Tabela 5.13 são apresentadas as diferenças entre o t.d.f e o t.d.d das medidas de similaridade avaliadas para as séries artificiais e da *NN GC1* agrupadas.

Tabela 5.13: Diferença entre t.d.f e t.d.d para as séries artificiais e da *NN GC1*.

	Manhattan	**Euclidiana**	**Métrica** L_3	$L_{p0.1}$	$L_{p0.3}$	$L_{p0.5}$	$L_{p0.7}$	***DTW***	**Canberra**	**Geodésica**
(t.d.f-t.d.d)	2	7	5	-6	-3	-1	-1	-8	2	3

Nessa tabela, é possível observar que as medidas Euclidiana e Métrica L_3 apresentaram as maiores diferenças positivas entre o t.d.f e o t.d.d, indicando sua possível vantagem sobre as demais medidas, seguidas pelas medidas Geodésica, Manhattan e Canberra. Em contrapartida, todas as medidas L_p Fracionárias apresentaram diferenças negativas, demonstrando possível

sua desvantagem. Já a medida *DTW* foi a que apresentou o menor valor de diferença entre t.d.f e t.d.d, indicando sua possível desvantagem em relação a maioria das medidas avaliadas.

As medidas L_p Inteiras estão entre as medidas de similaridade mais utilizadas na literatura, em grande parte devido a sua simplicidade de compreensão e interpretação, além do custo computacional, usualmente, inferior. Seu desempenho na previsão de ST, utilizando o algoritmo *kNN-TSP*, demonstra que além dessas características, as medidas L_p Inteiras apresentaram baixos valores de *MAPE*, quando comparadas às demais medidas avaliadas neste trabalho. Em geral, essas medidas apresentaram os menores erros médios, ou, nos piores casos, erros médios intermediários para a maioria das séries. Assim, é possível reafirmar sua adoção preferencial para a maioria das séries.

Na literatura, observa-se que na maior parte das vezes, dentre as medidas L_p Inteiras, a distância Euclidiana é adotada como medida de similaridade a ser utilizada. Entretanto, devido a não constatação de **d.e.s** para nenhum dos conjuntos de dados entre as medidas L_p Inteiras e o menor custo computacional, a distância Manhattan pode ser candidata interessante como medida de similaridade para previsão de dados temporais, reafirmando assim, resultados anteriores obtidos, com k variando de um à cinco (Aikes Junior et al., 2011).

Apesar das medidas L_p Fracionárias apresentarem vantagens quanto à exatidão nas tarefas de agrupamento e classificação, quando comparadas às medidas L_p Inteiras em dados com várias dimensões (Aggarwal et al., 2001), essa vantagem não foi constatada para a previsão de ST. As medidas L_p Fracionárias apresentaram maiores valores de *MAPE*, sendo que foi constatada **d.e.s** favorável apenas quando comparadas às distâncias *DTW* e Canberra nas séries artificiais. Entretanto, não houve **d.e.s** favorável para essas medidas nas séries da *NN GC1*, inclusive havendo **d.e.s** desfavorável para algumas medidas L_p Fracionárias quando comparadas a *DTW* e a Canberra. Quando analisados os dados das séries artificiais e da *NN GC1* em conjunto, essas medidas apresentaram vantagens apenas para com a *DTW*. Assim, a característica das medidas L_p Fracionárias de ressaltar a percepção de pequenas diferenças em conjuntos de alta dimensionalidade não foi vantajosa para a seleção dos melhores vizinhos mais próximos nas séries avaliadas neste trabalho. Desse modo, verifica-se que as distâncias L_p Fracionárias podem não ser as mais adequadas para a previsão de dados temporais utilizando o algoritmo *kNN-TSP*, considerando um cenário amplo.

Para a maioria das situações, a distância *DTW* apresentou **d.e.s** com todas as medidas, sendo que, na maior parte dos casos, seus valores de erro médio foram superiores aos das demais medidas. Assim, ao contrário das tarefas de classificação na qual a *DTW* apresentou desempenho superior às L_p Inteiras (Ratanamahatana and Keogh, 2005), essa medida aparenta não ser adequada como medida de similaridade empregada na tarefa previsão de dados temporais utilizando o algoritmo *kNN-TSP*. A razão dessa degradação de desempenho pode estar relacionada com a característica dessa medida de buscar o melhor alinhamento entre as séries,

o que acaba por diminuir a influência de pequenas diferenças locais durante a etapa de busca dos vizinhos mais próximos. Assim, diferenças locais que poderiam ser verdadeiras e desejadas para o cálculo do valor futuro acabam recebendo menor influência ou mesmo não sendo consideradas. Isso possibilita que a escolha dos vizinhos mais próximos acabe por não ser a mais adequada para a função de previsão.

A medida Canberra permaneceu como medida de desempenho intermediário tanto para as séries artificiais quanto para as séries da *NN GC1*. Assim, a característica de normalização dessa medida demonstra não apresentar vantagens consideráveis quando empregada na previsão de ST com o algoritmo *kNN-TSP*.

Apesar da medida Geodésica ter apresentado valores de *MAPE* baixos, esses não foram os melhores. Essa medida apresentou **d.e.s** desfavorável apenas para as medidas L_p Inteiras e favorável apenas para a *DTW* e Canberra. Entretanto, seu elevado custo computacional, quando comparada às medidas L_p Inteiras e a não constatação de **d.e.s** favorável para a Geodésica nesses casos, indica que essa distância não apresentou vantagem considerável. Dessa forma, reforça-se a suposição de que medidas de menor custo computacional, como a Manhattan, podem ser empregadas para a previsão de ST.

5.5 Considerações Finais

Neste capítulo foram apresentadas as ST artificiais e reais utilizadas na avaliação experimental, bem como a configuração experimental e os métodos de avaliação dos resultados. A análise dos resultados também foi discutida.

No próximo capítulo é introduzido o ambiente de estudo de caso, bem como os dados do estudo de caso, resultados, discussão e a proposta de uma medida de similaridade composta.

Capítulo 6

Estudo de Caso

6.1 Considerações Iniciais

Neste capítulo é apresentado um estudo de caso comparativo, utilizando dados reais, realizado com o algoritmo *kNN-TSP* e as diversas medidas de similaridade apresentadas anteriormente. São detalhados os dados utilizados e os resultados obtidos pelo trabalho fonte desses dados, os quais foram comparados com o deste trabalho. A configuração dos experimentos, resultados obtidos e discussão são também apresentados. É ainda introduzida a proposta de uma medida de similaridade composta.

6.2 Previsão do Fluxo Diário de Pacientes

A Área de Emergência (AE) de um hospital tem como sua principal responsabilidade tratar casos que necessitam de atenção imediata, pois apresentam risco de morte ao paciente. A superlotação em AE é um dos principais problemas enfrentados por hospitais, devido a característica emergencial do atendimento. Como consequências dessa superlotação tem-se (Derlet and Richards, 2000):

- O risco à segurança do paciente, que em algumas situações acaba por não ser atendido com a rapidez necessária;
- O aumento da dor e do sofrimento do paciente;
- O tempo prolongado de espera acaba por gerar impaciência e, em alguns casos, até mesmo violência por parte dos pacientes;
- A diminuição da produtividade do médico;
- A diminuição da qualidade das aulas em hospitais universitários.

As causas dessas superlotações incluem (Derlet and Richards, 2000):

- O aumento no volume de pacientes;
- A falta de leitos, o que faz com que os pacientes aguardem na AE por vagas;
- A lentidão nos serviços de radiologias e laboratoriais;
- A pouca quantidade de médicos, enfermeiros e pessoal administrativo;
- As barreiras culturais e linguísticas.

Apesar de serem tomadas medidas com o intuito de controlar a superlotação, essas acabam por não acompanhar a demanda (Derlet and Richards, 2000). Uma ferramenta de análise, que pode auxiliar no controle e no gerenciamento da superlotação de AE, é a estimativa da quantidade de pacientes que serão atendidos diariamente nesse setor. De maneira a acompanhar a evolução da demanda, são necessários novos meios para estimar a quantidade diária de pacientes em AE, possibilitando assim um melhor planejamento da alocação de pessoas e recursos.

Observando essa necessidade, Kam et al. (2010) apresentaram um estudo sobre modelos de previsão da quantidade de pacientes atendidos pela AE em um hospital coreano. Como os dados utilizados consistem em dados temporais, foram empregados métodos de previsão de ST. Esses métodos demonstraram ser promissores para a previsão do fluxo diário de pacientes em AE.

Para a realização do estudo foram analisados dados provenientes do sistema de informação utilizado pelo hospital, incluindo 189.511 eventos que envolveram 169.375 pacientes atendidos pela AE entre janeiro/2007 e março/2009. A variação da quantidade de pacientes pode ser observada na Figura 6.1. Para a construção do modelo foram utilizados dados dos primeiros dois anos (janeiro/2007 à dezembro/2008), e para a validação foram utilizados dados dos três meses seguintes (janeiro/2009 à março/2009). A contagem da quantidade de pacientes atendidos por dia foi iniciada à meia-noite e encerrada à meia-noite seguinte. Além da quantidade de pacientes, foram coletados dados de quinze variáveis, conforme apresentado na Tabela 6.1, sendo que os dados meteorológicos utilizados foram adquiridos nos sites de agências de meteorologia.

O feriado Chuseok foi separado dos demais feriados por representar o feriado mais importante da Coreia. Ele é celebrado a partir do 15^{o} dia do 8^{o} mês do calendário lunar, tendo uma duração de três dias. Nesse feriado ocorre uma grande movimentação de pessoas em direção às suas cidades natais, onde prestam homenagens aos seus ancestrais. Desse modo, esse feriado representa um comportamento diferenciado quando comparado aos demais feriados. Os dois picos mais acentuados na Figura 6.1 correspondem aos dias do feriado de Chuseok. Dias subsequentes aos feriados também apresentaram comportamento diferenciado no fluxo de pacientes, dessa maneira também é importante identificá-los; são considerados pós-feriados, os dias úteis e sábados imediatamente seguidos de um feriado.

Tabela 6.1: Descrição das variáveis do estudo de caso (Kam et al., 2010).

Variável	**Detalhes**
Mês	Janeiro, Fevereiro, Março, Abril, Maio, Junho, Julho, Agosto, Setembro, Outubro, Novembro, Dezembro
Dia da Semana	Domingo, Segunda-feira, Terça-feira, Quarta-feira, Quinta-feira, Sexta-feira, Sábado
Quarto do Ano	1Q, 2Q, 3Q, 4Q
Feriado	Dias úteis = 0, Feriado = 1, Pós-feriado = 2
Chuseok	Sim = 1, Não = 0
Estações	Primavera, Verão, Outono, Inverno
Temperatura Média	Valor da temperatura média
Temperatura Mínima	Valor da temperatura mais baixa
Temperatura Máxima	Valor da temperatura mais alta
Diferença de Temperatura	Diferença entre a Temperatura Máxima e a Temperatura Mínima
Chuva	Sim (>= 10 mm) = 1, Não = 0
Neve	Sim = 1, Não = 0
Velocidade do Vento	O valor da velocidade do vento
Umidade Relativa	Valor da umidade relativa
Poeira Amarela	Sim (Tempestade de Areia) = 1, Não = 0

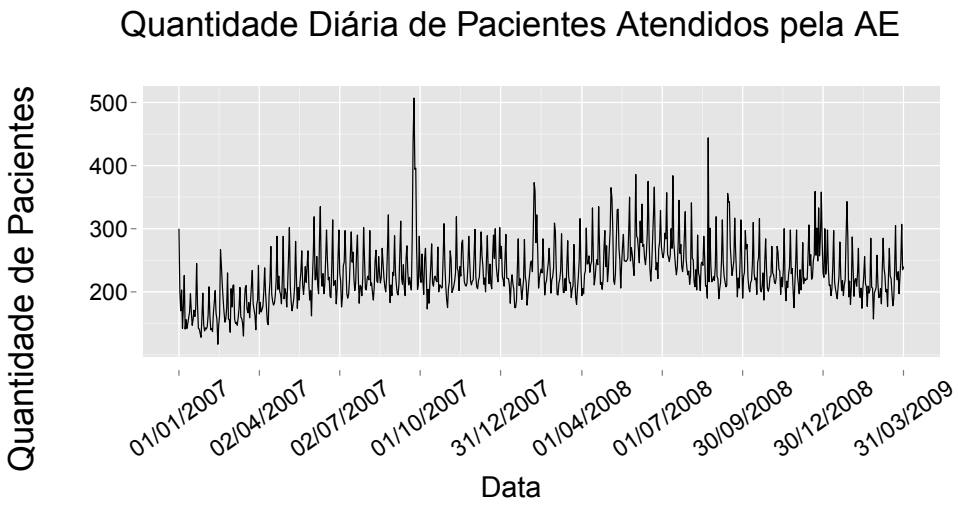


Figura 6.1: Quantidade de pacientes atendidos pela AE de um hospital coreano entre janeiro/2007 e março/2009 (Modificado de Kam et al. (2010)).

Com relação às variáveis ambientais, são considerados dias chuvosos apenas aqueles em que a quantidade de precipitação excede dez milímetros. Foram considerados dias com neve aqueles que apresentaram abundância de aglomeração de neve. A poeira amarela, também chamada de poeira asiática, representa o fenômeno meteorológico das tempestades de areia que afetam China, Coreia, Japão e leste da Rússia, chegando, em alguns casos, a afetar a qualidade do ar no leste dos Estados Unidos da América. Esse fenômeno afeta consideravelmente a qualidade do ar, aumentando a quantidade de pessoas que necessitam de atendimento médico.

Os métodos de previsão utilizados por Kam et al. (2010) consistiram nos modelos de MA, SARIMA univariado e SARIMA multivariado, sendo que o último adiciona variáveis explanatórias ao modelo, de maneira que alterações nas variáveis explanatórias exerçam influência na

variável dependente. A variável dependente consiste na quantidade de pacientes que visitam a AE, e as demais variáveis são consideradas variáveis explanatórias.

De maneira a verificar o ajuste e a exatidão dos modelos, os mesmos foram analisados por meio de *Akaike Information Criterion* (*AIC*), *Bayesian Information Criterion* (*BIC*) e *MAPE*. Foram realizados teste chi-quadrado, de maneira a verificar a existência de **d.e.s** entre o conjunto de dados de treinamento e testes, a qual não foi evidenciada. A estimação de parâmetros através de máxima verossimilhança indicou que apenas as variáveis Chuseok, Estações, Temperatura Média e Chuva são adequadas para serem selecionadas como variáveis explanatórias.

Como mencionado, três modelos foram desenvolvidos: MA(2), SARIMA(1,0,1)$(0,1,1)_7$ univariado e SARIMA(1,0,2)$(0,1,1)_7$ multivariado. O modelo MA demonstrou pouca exatidão da previsão, apresentando um valor de *MAPE* de 12,9%, indicando que a média dos valores observados é inadequada como modelo de previsão. Os dois modelos SARIMA apresentaram valor de *MAPE* inferiores a 10%, sendo que o modelo univariado apresentou erro médio de 7,8% e o multivariado 7,4% de *MAPE*.

O valor de erro médio inferior apresentado pelo modelo SARIMA multivariado demonstrou, para o trabalho realizado, que a inclusão de variáveis explanatórias, para a previsão da quantidade diária de pacientes atendidos pela AE, aumenta a exatidão da previsão quando utiliza-se modelos SARIMA.

6.3 Proposta de uma Medida Composta

Por meio do estudo experimental, apresentado no Capítulo 5, foi possível observar a diferença de desempenho do algoritmo *kNN-TSP* utilizando diferentes medidas de similaridade, frente a um grande conjunto de ST com características diversas. Observou-se também, vantagens e desvantagens de determinadas medidas para diversas situações. Com base nesses diferentes comportamentos observados, é proposto neste trabalho o desenvolvimento de uma medida de similaridade composta, que combine as características das medidas de similaridade individuais, buscando combinar as vantagens de suas características diversificadas.

A medida proposta inicialmente consiste em uma combinação linear das medidas individuais. De maneira a favorecer características específicas de cada medida, a distância calculada por cada uma das medidas individualmente é então ponderada através da atribuição de pesos, conforme Equação 6.1:

$$\sum_{i=1}^{I} \lambda_i \omega_i \tag{6.1}$$

onde λ representa a distância calculada pela medida de similaridade i, atribuindo o peso ω dessa

distância e I a quantidade de medidas a serem consideradas.

Neste trabalho considera-se o valor do peso ω como sendo inversamente proporcional ao erro *MAPE* da medida individual, obedecendo a Equação 6.2:

$$\omega = \frac{1}{MAPE_{(M_s,k)}} \tag{6.2}$$

onde $MAPE_{(M_s,k)}$ é o valor do *MAPE* da medida de similaridade M_s para a quantidade k de vizinhos próximos, para a série temporal sendo prevista.

Desse modo, para a definição da influência (peso) das medidas individuais, se faz necessário o conhecimento prévio do valor de *MAPE* dessas medidas. O conhecimento prévio do erro médio das medidas individuais pode ser adquirido de diversas maneiras, como:

- Por meio de experimentos com séries artificiais. Por exemplo, utilizando séries artificiais com características aproximadas às características da série a ser prevista, adotando o valor médio do *MAPE* resultante da aplicação do algoritmo nessas séries;
- Por meio de avaliações com séries temporais reais do mesmo domínio da série a ser prevista. Dessa maneira, o valor médio do *MAPE* de séries do mesmo domínio servem de estimadores para o erro médio da série a ser prevista;
- Por meio do passado da própria série a ser prevista. Para isso, remove-se uma quantidade definida de pontos do final da série e realiza-se previsões utilizando as medidas individuais na própria série, sem considerar a quantidade removida. Calcula-se então o *MAPE* das medidas individuais comparando os pontos previstos do final da série com seus pontos reais que foram previamente removidos, e o resultado é utilizado para a ponderação das medidas individuais.

Assim, as medidas que apresentam melhor desempenho para determinada série e para determinado valor de vizinhos mais próximos, exercem maior influência que aquelas que apresentam desempenho inferior. Nesse sentido, busca-se tomar vantagem das características individuais das medidas, favorecendo aquelas de melhor desempenho para determinada situação. Cabe-se observar que, nesta proposta inicial, os resultados das medidas individuais não foram normalizados, podendo ocorrer maior influência de determinadas medidas devido as suas diferentes escalas.

De modo a selecionar as medidas a serem utilizadas, foram empregados métodos de Seleção de Atributos (SA), buscando reduzir a quantidade de medidas candidatas adotadas para a composição da medida proposta. A taxa de redução do conjunto de medidas selecionada foi 50%, reduzindo assim o conjunto para cinco medidas. O processo de SA e os métodos utilizados são brevemente descritos a seguir.

6.3.1 Seleção de Atributos

A inserção da tecnologia em várias áreas do conhecimento está permitindo a criação de conjuntos de dados cada vez maiores e contendo mais atributos, dificultando o processo de MD e aumentando o tempo de processamento necessário. Nesse sentido, um dos objetivos da Seleção de Atributos (SA) é reduzir o tamanho do conjunto de dados, removendo atributos redundantes e/ou irrelevantes. Assim, a SA[1] busca encontrar um subconjunto mínimo de atributos de tal maneira que a probabilidade resultante das classes do subconjunto de dados seja a mais próxima possível do conjunto original (Han and Kamber, 2006).

As estratégias de SA podem ser classificadas em três categorias principais (Liu et al., 2010):

***Filter*:** Nessa abordagem, a SA ocorre como um processo separado do algoritmo de AM, que será usado para a construção do modelo, analisando as características gerais dos dados para a escolha dos atributos. Desse modo, os métodos *filter* são independentes do algoritmo de AM empregado, que receberá como entrada os atributos importantes fornecidos pelo filtro;

***Wrapper*:** Nessa abordagem, a SA ocorre com o auxílio do algoritmo de AM, que é empregado como uma caixa preta, analisando a cada iteração o subconjunto de atributos em questão. Assim, um subconjunto de atributos é gerado e utilizado como entrada para esse algoritmo, o qual analisa a exatidão resultante desse subconjunto de atributos como classificador, até que determinado critério de parada seja satisfeito;

***Embedded*:** Nessa abordagem, a SA ocorre como um processo incorporado ao algoritmo de AM, ou seja, a seleção dos atributos está embutida no algoritmo de AM empregado.

A SA não é o foco deste trabalho, porém, é uma ferramenta que foi utilizada para auxiliar na seleção de medidas para a proposta da medida composta. Assim, como mencionado, o conjunto de dados escolhido para a composição da entrada para a SA, são os valores de *MAPE* de todas as medidas de similaridade analisadas na avaliação experimental (Capítulo 5) para $k = 5$. Esse valor foi selecionado por representar a quantidade de vizinhos onde ocorre a maior parte da variação dos resultados. Os conjuntos foram avaliados individualmente, ou seja, foram desenvolvidos conjuntos utilizando apenas séries artificiais, apenas as séries da *NN GC1* e a composição de ambos os conjuntos. Desse modo, com esses dados foram desenvolvidas três tabelas atributo-valor, conforme características descritas na Tabela 6.2. Nessa tabela, cada exemplo corresponde a um ponto previsto, cada atributo representa uma medida de similari-

[1]Neste trabalho, o termo Seleção de Atributos será usado para se referir a seleção de um subconjunto de atributos.

dade avaliada e os valores desses atributos representam os erros percentuais para cada exemplo, como ilustrado na Tabela 6.3.

Tabela 6.2: Característica das tabelas atributo-valor utilizadas para a SA.

Conjunto	Exemplos	Atributos
Artificiais	563	10
NN GC1	1052	10
Artificiais + ***NN GC1***	1615	10

Tabela 6.3: Exemplo de tabela atributo-valor para a SA.

	Medida 1	**Medida 2**	$\cdots$	**Medida 10**
Ponto previsto 1	Erro percentual	Erro percentual	$\cdots$	Erro percentual
Ponto Previsto 2	Erro percentual	Erro percentual	$\cdots$	Erro percentual
Ponto Previsto 3	Erro percentual	Erro percentual	$\cdots$	Erro percentual
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Ponto Previsto N	Erro percentual	Erro percentual	$\cdots$	Erro percentual

Para a realização da SA utilizou-se duas abordagens: a primeira empregando Algoritmos Genéticos Multiobjetivo (AGM), e a segunda empregando a Dimensão Fractal (DF) para tratar os atributos redundantes. Ambas as abordagens caracterizam-se como pertencentes à classe *filter* e são brevemente apresentadas a seguir.

Algoritmos genéticos são algoritmos de busca e otimização inspirados em processos da natureza e na teoria de Darwin. Esses algoritmos objetivam encontrar soluções potencialmente otimizadas por meio de um processo iterativo, incorporando propriedades de busca heurística e otimizando uma função de aptidão que mensura a qualidade das soluções (Freitas, 2002). Neste trabalho, foi adotado o método empregado em Spolaôr (2010).

Assim, nessa abordagem é necessária a determinação de medidas de importância que serão utilizadas para avaliar a possível existência de redundância entre atributos. Neste trabalho, as seguintes medidas de importância de atributos foram empregadas para a SA:

***Laplacian Score* (*LS*):** Cria uma grafo de vizinhos mais próximos buscando identificar os atributos que favorecem as ligações entre as observações, construindo a estrutura geométrica local. Assim, essa medida de importância baseia-se no princípio de que se duas observações apresentam-se próximas, é provável que estejam relacionadas ao mesmo tópico, como ocorre em vários problemas de AM, por exemplo, classificação, em que observações de mesmo rótulo tendem a se encontrar próximas umas às outras. Desse modo, atributos que favorecem a construção dessas ligações são selecionados (He et al., 2006);

***Representation Entropy (RE)*:** Utiliza auto-valores extraídos de uma matriz de covariância como meio para mensurar a redundância de subconjuntos de atributos. Dessa maneira,

caso o valor de todos os auto-valores sejam iguais, verifica-se a existência de pouca redundância entre os dados, ou seja, que a informação é distribuída uniformemente nos dados. Já caso apenas um auto-valor seja diferente de zero, então toda a informação poderia ser representada por esse único atributo (Yan, 2007);

Intra Correlação (IC): Quantifica a relação entre os atributos de um subconjunto de dados por meio do estudo da correlação de Pearson. Atributos que apresentem altos valores de correlação podem ser considerados menos relevantes para descrever o subconjunto de dados (Wang and Huang, 2009).

A segunda abordagem de SA consiste em ordenar os atributos por importância utilizando a DF para tratar os atributos redundantes, conforme proposto por Lee (2005). Diferentemente da primeira abordagem, esta última não retorna apenas um subconjunto de atributos importantes, e sim uma ordem de importância dos atributos. Isso permite a seleção da quantidade de atributos mais adequada a cada problema, segundo essa medida de importância usando o valor da DF do conjunto de dados (Lee and Monard, 2006). De maneira a realizar a SA utilizando essa abordagem, foi adotada a variação do algoritmo que emprega métodos de aproximação para a determinação da DF (*MDE-Fast*) (Traina et al., 2003).

6.3.2 Composição da Medida

Como mencionado, para a composição da medida proposta, foram selecionadas as cinco medidas que apresentaram a maior quantidade de seleções pelos algoritmos de SA. A combinação dessas cinco medidas, ponderadas utilizando o inverso de seus valores de *MAPE*, constitui a medida Composta proposta inicialmente.

Na Tabela 6.4 é apresentado o resultado da aplicação das diversas configurações dos algoritmos de SA. Nessa tabela, L_1 representa a medida Manhattan, L_2 a medida Euclidiana e L_3 a Métrica L_3. As células sombreadas destacam as cinco medidas que foram selecionadas com maior frequência pelos algoritmos de SA.

As cinco medidas mais representativas para o conjunto de medidas avaliadas neste trabalho, selecionadas pelos algoritmos de SA, foram as medidas Métrica L_3, $L_{p0.1}$, *DTW*, Canberra e Geodésica. Desse modo, a combinação dessas cinco medidas, ponderadas de acordo com o inverso do *MAPE* individual em cada série e valor de k, compõem a medida nessa proposta inicial.

Conforme mencionado, o conhecimento do *MAPE* das medidas individuais é necessário para a aplicação da medida Composta. Neste trabalho, esse conhecimento prévio do erro médio foi adquirido utilizando-se o passado da série, ou seja, foram realizadas as previsões para os três últimos meses da série de fluxo diário de pacientes (Setembro/2008 à Dezembro/2008)

Tabela 6.4: Resultados da SA sobre os dados experimentais. L_1 representa a medida Manhattan, L_2 a medida Euclidiana e L_3 a Métrica L_3. As células sombreadas destacam as cinco medidas mais frequentes.

Algoritmo	Variação	Base	L_1	L_2	L_3	$Lp_{0.1}$	$Lp_{0.3}$	$Lp_{0.5}$	$Lp_{0.7}$	*DTW*	Canberra	Geodésica	DF
Algoritmo Genético	*IC*	Artificiais		*						*			
		Reais				*						*	
		Geral				*						*	
	LS	Artificiais								*			
		Reais									*		
		Geral									*		
	RE	Artificiais	*	*	*	*	*	*	*	*	*		
		Reais			*	*				*	*	*	
		Geral			*	*				*	*	*	
Fractal Dimension-Based Filter	*MDE*	Artificiais			*								00,310
		Reais		*	*	*					*		03,017
		Geral			*								00,823
		Total	**1**	**3**	**6**	**6**	**1**	**1**	**1**	**5**	**6**	**4**	

empregando cada uma das medidas individuais. Esses valores de *MAPE* adquiridos foram então utilizadas para a ponderação dos pesos da medida Composta.

Tabela 6.5: Composição da medida proposta para a série de fluxo diário de pacientes.

Vizinhos Próximos	Medida Composta
1-*NN*	Métrica $L_3 \times 0{,}0923$ + $L_{p0.1} \times 0{,}0677$ + *DTW* $\times 0{,}0916$ + Canberra $\times 0{,}0833$ + Geodésica $\times 0{,}0922$
5-*NN*	Métrica $L_3 \times 0{,}1132$ + $L_{p0.1} \times 0{,}0976$ + *DTW* $\times 0{,}1191$ + Canberra $\times 0{,}0997$ + Geodésica $\times 0{,}1208$
10-*NN*	Métrica $L_3 \times 0{,}1173$ + $L_{p0.1} \times 0{,}0953$ + *DTW* $\times 0{,}1162$ + Canberra $\times 0{,}1013$ + Geodésica $\times 0{,}1200$

Na Tabela 6.5 são apresentadas as formulações da medida Composta para a série de fluxo diário de pacientes, utilizando como peso o inverso do *MAPE* das medidas individuais para a previsão dos meses de Setembro/2008 à Dezembro/2008.

6.4 Configuração Experimental

Como mencionado, para fins de comparação da eficiência do algoritmo *kNN-TSP* para a previsão de dados temporais relacionados a saúde, foi realizado um estudo comparativo com o trabalho de Kam et al. (2010). Assim, a configuração dos parâmetros do algoritmo *kNN-TSP* foram ajustadas da seguinte maneira:

- Horizonte de Previsão – Os pontos previstos, assim como no trabalho em comparação, foram os três primeiros meses do ano de 2009;
- Tamanho da Janela – O tamanho da janela utilizado, w, foi de sete dias. Essa janela de previsão foi selecionada de acordo com a sazonalidade identificada por Kam et al. (2010), com picos aos domingos;

- Quantidade de Vizinhos Mais Próximos – Os valores de vizinhos mais próximos avaliados foram $k = 1$, $k = 5$ e $k = 10$;
- Medidas de Similaridade – As medidas analisadas foram as medidas da Norma L_p, com valores de p inteiros variando de 1 à 3; valores fracionários, sendo $p = 0.1$, $p = 0.3$, $p = 0.5$ e $p = 0.7$; a medida Canberra; a medida *DTW*; a medida Geodésica e a medida Composta, proposta inicialmente neste trabalho.

6.5 Discussão dos Resultados

Na Tabela 6.6 são apresentados os valores de média, desvio padrão, máximo e mínimo do *MAPE* para a série de fluxo diário de pacientes, agrupando os resultados para um, cinco e dez vizinhos mais próximos. Valores que apresentam sombreamento de célula em cor verde representam o menor valor na respectiva linha, e valores com sombreamento de célula em cor vermelha o maior valor da linha. Os tons entre as cores verde e vermelha representam valores intermediários.

Tabela 6.6: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para a série de fluxo diário de pacientes, agrupando os valores de um, cinco e dez vizinhos mais próximos.

Cálculo	Manhattan	Euclidiana	Métrica L_3	$Lp_{0.1}$	$Lp_{0.3}$	$Lp_{0.5}$	$Lp_{0.7}$	*DTW*	Canberra	Geodésica	Composta
Média	10,1744	09,8264	09,6122	12,4776	09,8835	09,8054	10,0572	10,1128	09,8344	09,6520	12,4677
Desv. Padrão	10,1490	09,5845	09,1452	11,6206	09,2289	08,8695	09,2488	09,4759	09,0714	09,7192	11,6333
Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000
Máximo	57,8125	57,8125	48,7685	76,5625	60,5263	40,0000	48,7685	57,8125	46,5700	48,4536	76,5625

Nessa tabela, é possível observar que a medida Métrica L_3 apresentou o menor valor de erro médio dentre todas as medidas avaliadas (09,6122%), e que as demais medidas L_p Inteiras assumiram valores intermediários para as médias de erro, agrupando as previsões para um, cinco e dez vizinhos mais próximos. Quando avaliados os valores de k individualmente, todas as medidas L_p Inteiras apresentaram valores de *MAPE* intermediários, exceto quando $k = 1$ em que a medida Métrica L_3 apresentou o menor valor de erro médio entre as medidas avaliadas, sendo 10,8145%. Os resultados detalhados para cada valor de k avaliado para todas as medidas, considerando a série de fluxo diário de pacientes, são apresentados no Apêndice C.

As medidas L_p Fracionárias apresentaram valores de erro médio intermediários, exceto a medida $L_{p0.1}$, que apresentou o maior valor de erro médio entre todas as medidas avaliadas para os valores de k agrupados (12,4776%). Quando avaliados os valores de k individualmente, a medida $L_{p0.1}$ apresentou novamente os maiores valores de *MAPE*, sendo 10,4977% para $k = 10$, 11,6069% para $k = 5$ e 15,3280% para $k = 1$. Já as demais medidas L_p Fracionárias apresentaram valores intermediários de erro médio.

A medida *DTW* apresentou valores intermediários de *MAPE* para os dados quando utili-

zados todos os valores de k avaliados agrupados. Para o valor do erro médio quando utilizado os valores de $k = 1$ e $k = 10$, essa medida continuou a apresentar valores intermediários de erro médio. Já para $k = 5$, essa medida apresentou o menor erro médio, 08,8574%.

A medida Canberra, assim como a medida *DTW*, apresentou valores intermediários de erro médio para os dados que utilizaram todos os valores de k avaliados agrupados. Essa medida também alcançou valores intermediários de erro médio para os valores de k avaliados individualmente.

A medida Geodésica, assim como a *DTW* e a Canberra, apresentou valores intermediários de *MAPE* considerando todos os valores de k avaliados agrupados. Para $k = 10$, essa medida foi a que atingiu o menor valor de erro médio, com 08,3195%. Já para $k = 5$ e $k = 1$, a medida Geodésica alcançou valores intermediários de erro médio.

A medida Composta, proposta neste trabalho, apresentou o segundo maior valor de *MAPE* para todos os valores de k agrupados. Essa medida continuou apresentando o segundo maior erro médio para $k = 10$ e $k = 5$. Quando $k = 1$, essa medida alcançou o mesmo valor que a medida $L_{p0.1}$, 15,3280%, tendo sido esse o maior valor de erro médio.

Na Figura 6.2 são apresentadas a variação da média, em forma de pontos, e do desvio padrão, representado pela área sombreada, do *MAPE* das medidas de similaridades avaliadas para a série de fluxo diário de pacientes. As medidas Manhattan, Euclidiana e Métrica L_3 foram agrupadas em L_p Inteiras e as medidas $L_{p0.1}$, $L_{p0.3}$, $L_{p0.5}$ e $L_{p0.7}$ foram agrupadas em L_p Fracionárias.

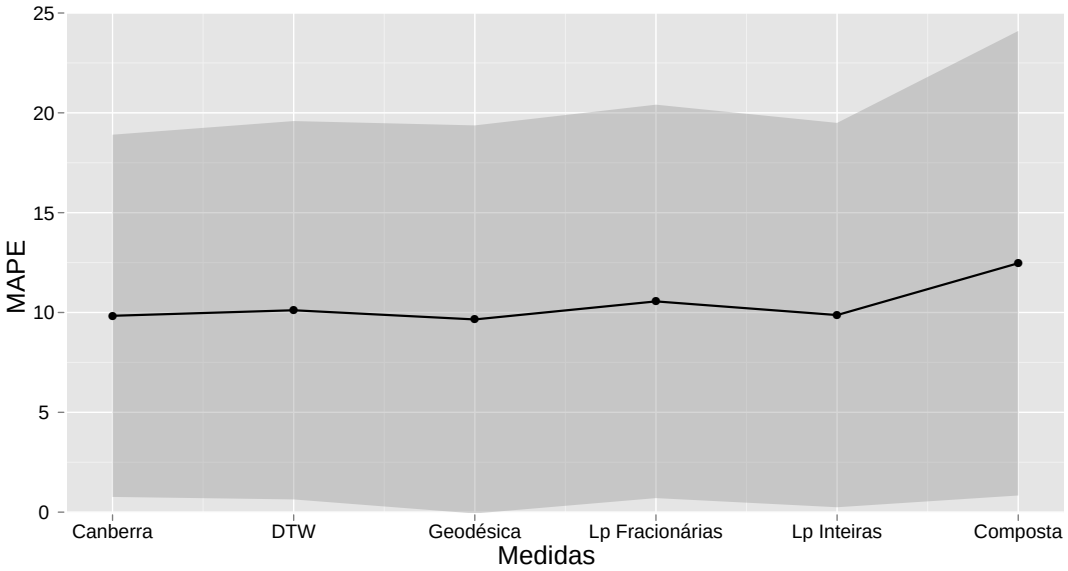


Figura 6.2: Média e desvio padrão de *MAPE* da série de fluxo diário de pacientes.

Nessa figura, é possível observar que a medida Composta apresentou o maior valor de erro médio, bem como o maior desvio padrão entre todas as medidas. As medidas L_p Fracionárias

alcançaram tanto o segundo maior erro médio quanto o segundo maior desvio padrão. As medidas Canberra, Geodésica e L_p Inteiras atingiram os menores valores de *MAPE* e de desvio padrão, encontrando-se visualmente muito próximas. A medida *DTW* apresentou um valor de erro médio maior do que as medidas Geodésica e L_p Inteiras, entretanto um desvio padrão menor. Observa-se ainda que, exceto para a medida Composta, a variação dos valores de média e desvio padrão entre as medidas foi pequena, encontrando-se próxima à 10%.

Do mesmo modo que foi realizado para as séries da avaliação experimental, utilizou-se o teste estatístico de Friedman para verificar a existência de **d.e.s** entre os valores de *MAPE* das medidas avaliadas para a série de fluxo diário de pacientes. O resultado da aplicação do teste estatístico para essa série é apresentado na Tabela 6.7. Nessa tabela:

*****:** Representa a existência de **d.e.s** com $p\text{-}valor < 0,001$;

****:** Representa a existência de **d.e.s** com $p\text{-}valor < 0,01$;

***:** Representa a existência de **d.e.s** com $p\text{-}valor < 0,05$;

Célula em Branco: Representa que não foi possível constatar a existência de **d.e.s**;

t.d.f: Representa o total de **d.e.s** favorável (t.d.f) para a medida da respectiva linha;

t.d.d: Representa o total de **d.e.s** desfavorável (t.d.d) para a medida da respectiva coluna.

Assim, a notação *** na célula de encontro entre a linha da medida Manhattan e a coluna da medida $L_{p0.1}$, representa a existência de **d.e.s** com $p\text{-}valor < 0,001$ entre essas medidas, sendo que a distância Manhattan apresentou um valor de *MAPE* menor que a $L_{p0.1}$.

Tabela 6.7: Comparativo sobre a existência de **d.e.s** entre as medidas de similaridade para a série de fluxo diário de pacientes.

	Manhattan	**Euclidiana**	**Métrica L_3**	$L_{p0.1}$	$L_{p0.3}$	$L_{p0.5}$	$L_{p0.7}$	***DTW***	**Canberra**	**Geodésica**	**Composta**	**t.d.f**
Manhattan	——			***							***	**2**
Euclidiana		——		***							***	**2**
Métrica L_3			——	***							***	**2**
$L_{p0.1}$				——		***						**1**
$L_{p0.3}$				***	——						***	**2**
$L_{p0.5}$						——						**0**
$L_{p0.7}$				**			——				**	**2**
DTW				*				——			*	**2**
Canberra				***					——		***	**2**
Geodésica				***						——	***	**2**
Composta						***					——	**1**
t.d.d	**0**	**0**	**0**	**8**	**0**	**2**	**0**	**0**	**0**	**0**	**8**	——

Nessa tabela, é possível observar que não foi constatada **d.e.s** das medidas L_p Inteiras entre si. Todas as medidas L_p Inteiras apresentaram **d.e.s** favorável para com as medidas $L_{p0.1}$

e Composta, e não foi evidenciada **d.e.s** desfavorável para com nenhuma das demais medidas avaliadas.

A medida $L_{p0.1}$ apresentou **d.e.s** favorável apenas para com a medida $L_{p0.5}$ e desfavorável para com todas as demais medidas, exceto a $L_{p0.5}$ e a Composta. Já para a medida $L_{p0.5}$ foi evidenciada **d.e.s** desfavorável com a $L_{p0.1}$ e com a Composta, e as medidas $L_{p0.3}$ e $L_{p0.7}$ apresentaram **d.e.s** favorável para com a medida Composta. Para os demais casos, não foi constatada **d.e.s** para as medidas L_p Fracionárias.

As medidas *DTW*, Canberra e Geodésica apresentaram **d.e.s** favorável para com as medidas $L_{p0.1}$ e Composta. Para essas medidas não foi constatada **d.e.s** desfavorável para com nenhuma das medidas avaliadas na série de fluxo diário de pacientes.

A medida Composta apresentou **d.e.s** favorável apenas para com a medida $L_{p0.5}$ e desfavorável para com todas as medidas, exceto a $L_{p0.1}$ e a $L_{p0.5}$.

Ainda nessa tabela, pode ser observado que todas as medidas L_p Inteiras, *DTW*, Canberra e Geodésica apresentaram o mesmo t.d.f, dois, sendo essas **d.e.s** favoráveis para com as mesmas medidas, $L_{p0.1}$ e Composta. Já a medida $L_{p0.5}$ foi a única em que não foi possível evidenciar nenhuma **d.e.s** favorável.

As medidas $L_{p0.1}$ e Composta apresentaram o mesmo t.d.d, oito, sendo que a única medida além dessas duas que apresentou **d.e.s** desfavorável foi a $L_{p0.5}$, com um t.d.d igual a dois. Para as demais medidas não foi constatada nenhuma **d.e.s** desfavorável para a série de fluxo diário de pacientes.

Na Tabela 6.8 são apresentadas as diferenças entre o t.d.f e o t.d.d das medidas de similaridade avaliadas para a série de fluxo diário de pacientes.

Tabela 6.8: Diferença entre t.d.f e t.d.d para a série de fluxo diário de pacientes.

	Manhattan	**Euclidiana**	**Métrica** L_3	$L_{p0.1}$	$L_{p0.3}$	$L_{p0.5}$	$L_{p0.7}$	***DTW***	**Canberra**	**Geodésica**	**Composta**
(t.d.f-t.d.d)	2	2	2	-7	2	-2	2	2	2	2	-7

Nessa tabela, é possível observar que a maioria das medidas apresentou a mesma diferença entre t.d.f e t.d.d, 2 (dois), para a série de fluxo diário de pacientes. Exceções ocorrem apenas para diferenças negativas, sendo que as medidas $L_{p0.1}$ e Composta apresentaram os menores valores de índices, -7 (sete negativo), e a medida $L_{p0.5}$ alcançou uma diferença de -2 (dois negativo). Desse modo, as medidas $L_{p0.1}$ e Composta se apresentaram desfavoráveis para a maioria das medidas avaliadas, seguidas pela medida $L_{p0.5}$, enquanto as demais encontraram-se equilibradas para a série de fluxo diário de pacientes.

Assim como o ocorrido com as séries utilizadas na avaliação experimental, as medidas L_p Inteiras apresentaram para a série de fluxo diário de pacientes, em geral, os menores valores de *MAPE* entre as medidas avaliadas, tendo demonstrado novamente a viabilidade de sua utilização

para a previsão de ST empregando o algoritmo *kNN-TSP*. Novamente não foi constatada **d.e.s** entre as medidas L_p Inteiras, reafirmando a viabilidade da utilização da medida de menor custo computacional entre elas, isto é, a medida Manhattan.

De modo semelhante ao ocorrido com as medidas L_p Inteiras, para a maioria dos casos repetiu-se, para as medidas L_p Fracionárias, o comportamento encontrado com as séries da avaliação experimental, onde essas medidas apresentaram, para a maioria dos casos, valores de erro médio intermediários (séries artificiais) ou os maiores valores (séries da *NN GC1*). Exceção ocorreu no caso da medida $L_{p0.3}$ que apresentou-se como a melhor medida quando agrupados todos os valores de k bem como quando $k = 1$. Apesar da medida $L_{p0.1}$ ter alcançado os menores valores de *MAPE* em alguns casos, os testes estatísticos demonstraram que ela, assim como as demais medidas L_p Fracionárias, não apresentou **d.e.s** favorável para com nenhuma medida, exceto para com a medida Composta.

A *DTW* atingiu valores intermediários de erro médio, assim como ocorrido nas séries da *NN GC1*, o que representa uma melhora se comparado com as séries artificiais, onde na média geral, essa medida foi a que alcançou o maior *MAPE*. Entretanto, essa medida se apresentou estatisticamente vantajosa apenas para a medida Composta e a $L_{p0.1}$. Dessa maneira, seu alto custo computacional a torna desfavorável para utilização de previsão de ST também para a série de fluxo diário de pacientes.

A medida Canberra, da mesma maneira que para as séries da avaliação experimental, apresentou para a série de fluxo diário de pacientes valores de erro médio intermediários. Foi constatada, para essa medida, **d.e.s** favorável apenas para com as medidas $L_{p0.1}$ e Composta. Apesar de não haver **d.e.s** desfavorável, essa medida alcançou valores de erro médio intermediários e maior custo computacional, quando comparada com as L_p Inteiras, o que desfavorece sua utilização para a previsão de ST com o algoritmo *kNN-TSP*.

A medida Geodésica apresentou para série de fluxo diário de pacientes, comportamento semelhante ao encontrado nas séries da avaliação experimental, tendo alcançado valores intermediários de *MAPE* para a maioria dos casos, e em alguns, o menor valor de *MAPE*. Apesar de na maioria das vezes ter atingido valores intermediários de erro médio, esses, em geral, estão próximos aos menores valores. Entretanto, para essa medida foi constatada **d.e.s** favorável apenas para com as medidas $L_{p0.5}$ e para com a Composta, o que, somado ao seu maior custo computacional, acaba por desfavorecer a escolha dessa medida como medida de similaridade a ser adotada.

A medida Composta, proposta neste trabalho, apresentou para a maioria dos casos valores intermediários de *MAPE*, porém, muito próximos aos maiores valores de *MAPE* para a série de fluxo diário de pacientes. Somando esse comportamento encontrado à constatação de **d.e.s** desfavorável para com a maioria das medidas e favorável apenas para com a $L_{p0.5}$, e o seu alto custo computacional, essa medida acabar por ser inadequada para a previsão de ST na série de

fluxo diário de pacientes utilizando o algoritmo *kNN-TSP*.

Isso indica que a medida composta por meio da ponderação inversa do *MAPE*, como proposta inicialmente neste trabalho, utilizando seleção de atributos para reduzir em 50% o conjunto de medidas, pode não ser adequada para a previsão da série de fluxo diário de pacientes empregando o algoritmo *kNN-TSP*.

Cabe ressaltar que a medida Composta foi desenvolvida realizando a SA em um conjunto de dados que, apesar de possuir quantidade considerável de ST, pode não ter apresentado todas as características e combinações de características existentes. Assim, é possível que a escolha das medidas para a composição seja diferente em um conjunto de dados mais abrangente. Além disso, a SA, utilizada para indicar as medidas para composição da medida proposta, seleciona as medidas individuais que melhor representam a variabilidade do conjunto, sendo que isso não necessariamente implica em escolher as medidas com melhor desempenho para a maioria das situações. Devido a isso, medidas como a *DTW* e a $L_{p0.1}$, que apresentaram, em geral, valores de *MAPE* altos, foram selecionadas e, apesar de devido a ponderação dos pesos exercerem uma influência menor do que as medidas de melhor desempenho, sua influência no resultado final da medida Composta deve ser considerada. Além disso, a diminuição do conjunto de dados em 50% não implica, necessariamente, na melhor redução possível, sendo que a redução de melhor desempenho pode estar abaixo ou acima dessa porcentagem de corte.

Desse modo, verifica-se que ainda existem muitos estudos a serem realizados para o desenvolvimento da medida Composta, sendo a abordagem empregada neste trabalho, um estudo ainda em fase inicial.

6.6 Comparativo com o Estudo de Kam

Na Tabela 6.9 são apresentados as três melhores configurações e seus respectivos valores de *MAPE* para os dados do estudo inicial de Kam et al. (2010) e para este trabalho. Nessa tabela, a notação k*NN-TSP*$_{(k,M_s)}$ indica as configurações do algoritmo *kNN-TSP* utilizadas, onde k representa a quantidade de vizinhos mais próximos e M_s representa a medida de similaridade utilizada.

Nessa tabela, é possível observar que as melhores configurações do método SARIMA apresentaram menores valores de *MAPE* do que o algoritmo *kNN-TSP* quando utilizadas as configurações avaliadas (Seção 6.4). O SARIMA multivariado, que fez uso das variáveis explanatórias, alcançou o melhor desempenho entre todas as configurações avaliadas. O algoritmo *kNN-TSP*, na versão utilizada neste trabalho, não está preparado para dados multivariados, o que pode ter influência negativa na exatidão do algoritmo.

Quando utilizados dados univariados, ou seja, apenas a série de fluxo diário de pacientes

Tabela 6.9: Três melhores configurações e valores de *MAPE* do estudo inicial de Kam et al. (2010) e deste trabalho para a previsão da série de fluxo de pacientes.

Origem	Configuração	MAPE
Estudo Inicial de Kam et al. (2010)	MA(2)	12,909
	SARIMA $(1,0,1)(0,1,1)_7$ Univariado	07,788
	SARIMA $(1,0,2)(0,1,1)_7$ Multivariado	07,372
Este Trabalho	*kNN-TSP* $_{(10,\text{ Geodésica})}$	08,319
	kNN-TSP $_{(10,\text{ Euclidiana})}$	08,596
	kNN-TSP $_{(10,\ Lp\ 0.7)}$	08,637

sem variáveis explanatórias, o método SARIMA alcançou o segundo menor valor de *MAPE* (07,788%). Já o algoritmo *kNN-TSP* com $k = 10$ e a medida Geodésica atingiu seu menor valor de *MAPE* (08,319%), tendo apresentado uma diferença de 0,531% para com a melhor configuração SARIMA univariado.

Os autores do estudo inicial não disponibilizaram informações relativas a variabilidade do *MAPE* para as suas previsões. Assim, apesar dos métodos SARIMA terem apresentado os menores valores de *MAPE* para a série de fluxo diário de pacientes, quando comparado ao algoritmo *kNN-TSP*, não foi possível a realização de comparações sobre a variabilidade de ambas as abordagens.

Cabe ressaltar que apenas um parâmetro do algoritmo *kNN-TSP* foi avaliado neste trabalho, a medida de similaridade. Desse modo, é possível que a otimização de outros parâmetros, como a quantidade de vizinhos mais próximos (k) ou o tamanho da janela (w), ajude a reduzir ainda mais o erro do algoritmo. É possível perceber, por exemplo, uma diminuição do valor de *MAPE* a medida que se aumenta a quantidade de vizinhos mais próximos, assim outros valores de k além dos usados no escopo deste trabalho devem ser avaliados, pois podem diminuir ainda mais o erro do algoritmo.

Um outro fator a ser considerado é a quantidade de parâmetros explícitos a serem estimados para ambos os métodos, a qual é menor no algoritmo *kNN-TSP* quando comparado ao método SARIMA. Dessa maneira, a aplicação do primeiro apresenta uma maior facilidade do que a do segundo, sendo um fator de influência positiva para a escolha da utilização do algoritmo *kNN-TSP*.

6.7 Considerações Finais

Neste capítulo foram apresentados os problemas de estimação do fluxo diário de pacientes em AE de hospitais, bem como um estudo de caso realizado por Kam et al. (2010), em que foram utilizadas as abordagens de Médias Móveis, SARIMA univariado e SARIMA multivariado. Foi apresentado também o estudo de caso relativo a estimação do mesmo fluxo diário de paci-

entes utilizando o algoritmo *kNN-TSP*, além de um comparativo entre os resultados alcançados pelas duas abordagens. Foi introduzido também um estudo inicial sobre o desenvolvimento de uma medida Composta, empregando seleção de atributos para a escolha das medidas a serem utilizadas para a composição e o inverso do valor de *MAPE* como critério para atribuição de pesos.

No próximo capítulo são apresentadas as conclusões deste trabalho, bem como as principais contribuições, limitações e trabalhos futuros.

Capítulo 7

Conclusão

A previsão de séries temporais constitui um tema de grande interesse para várias áreas do conhecimento. Ao longo do tempo foram desenvolvidas diversas abordagens para a realização dessas estimações, tais como os métodos paramétricos de Médias Móveis (MA) e os Auto-regressivos de Médias Móveis Integrados (ARIMA). Atualmente, com a proximidade da computação em várias áreas do conhecimento, observa-se uma maior utilização da adaptação de métodos computacionais, como métodos de aprendizagem de máquina, relacionados à Mineração de Dados, para a solução de diversos problemas, tais como a previsão de dados temporais. Um exemplo dessas adaptações consiste no algoritmo *k-Nearest Neighbor - Time Series Prediction* (*kNN-TSP*), que é uma variação do algoritmo *k-Nearest Neighbor*, para previsão de séries temporais.

A utilização do algoritmo *kNN-TSP* é dependente da definição de alguns parâmetros, sendo um deles a medida de similaridade. Esse parâmetro é responsável por determinar a maneira que o algoritmo considera duas subsequências semelhantes para a seleção de vizinhos próximos.

Neste trabalho foi apresentado um estudo sobre a influência do parâmetro medida de similaridade na exatidão da previsão de séries temporais utilizando o algoritmo *kNN-TSP*.

De maneira a estudar essa influência, foram selecionadas várias medidas de similaridade, conforme descrito no Capítulo 4. O algoritmo *kNN-TSP*, em conjunto com essas medidas, foi submetido a uma avaliação experimental contendo séries temporais artificiais, de características sazonais e caóticas, e várias séries temporais reais relacionadas a transporte, com características diversas, conforme descrito no Capítulo 5. Os resultados foram avaliados por meio do *Mean Absolute Percentage Error* (*MAPE*), e foram realizadas análises por meio de estatística descritiva e analítica, permitindo dessa maneira verificar a existência de diferença estatisticamente significativa (**d.e.s**) que possa apoiar a escolha de medidas de similaridade que apresentem algum benefício.

Além da avaliação experimental utilizando ST artificiais e reais, foi realizado um estudo

de caso com dados reais relacionados à saúde, onde os métodos MA, SARIMA univariado e SARIMA multivariado foram comparados ao *kNN-TSP*, conforme descrito no Capítulo 6.

Neste trabalho, foi também introduzida uma proposta inicial de desenvolvimento de uma medida de similaridade composta, utilizando a seleção de atributos como método de escolha das medidas a serem empregadas na composição, as quais foram ponderadas pelo inverso de seu erro *MAPE*. Essa medida proposta foi então utilizada na previsão da série do estudo de caso.

As medidas L_p Inteiras apresentaram os menores valores de *MAPE* para a maioria das situações, ou, nos piores casos, valores de erro médio intermediários. Em geral, para essas medidas, foram constatadas **d.e.s** favoráveis para com a maioria das medidas avaliadas, sem terem sido evidenciadas **d.e.s** desfavoráveis. Esse resultado demonstra a capacidade de adaptação do algoritmo *kNN-TSP* utilizando essas medidas para as séries temporais avaliadas, favorecendo a sua utilização.

As medidas L_p Fracionárias, por sua vez, alcançaram valores de erro médio elevados para a maioria dos casos, tendo sido encontradas como as medidas de maior erro médio, ou, quando apresentaram erro médio intermediário, em geral, esse estava próximo do maior. Seu elevado custo computacional, quando comparadas às medidas L_p Inteiras, somado ao fato de, em grande parte das situações ter sido possível comprovar **d.e.s** desfavorável para com as demais medidas, acabam por tornar a utilização das medidas L_p Fracionárias desfavorável para a maioria dos casos estudados neste trabalho.

O desempenho da medida *Dynamic Time Warping* (*DTW*) para as tarefas de classificação, incluindo classificação de dados temporais é, em geral, melhor que os das medidas da Norma L_p. Esse melhor desempenho é, em parte, devido à sua característica de buscar o melhor alinhamento das séries temporais antes da comparação ponto-a-ponto, favorecendo que pequenas diferenciações possam ter impacto mínimo na comparação geral das séries. Entretanto, para a previsão de dados temporais utilizando o algoritmo *kNN-TSP*, essa mesma característica pode estar influenciando negativamente na seleção dos melhores vizinhos próximos, já que nesses casos, pequenas diferenças locais nas séries podem ser fatores determinantes para a escolha dos melhores vizinhos. Assim, essa medida acabou apresentando valores de erro médio elevados, em geral acrescidos de **d.e.s** desfavorável. Desse modo, o alto custo computacional dessa medida, somado ao fato de ter alcançado, em alguns casos, os maiores valores de erros, desfavorecem a sua utilização para a previsão de séries temporais com o algoritmo *kNN-TSP*.

A medida Canberra apresentou, para a maioria das situações verificadas neste trabalho, valores de erro médio intermediários. Apesar de seu custo computacional ser inferior aos das medidas Euclidiana e Métrica L_3, devido a sua característica de normalização, é ainda superior ao da medida Manhattan, não tendo evidenciado **d.e.s** favorável para com essas medidas para a maioria dos casos. Esses resultados desfavorecem a utilização dessa medida, em comparação

com as medidas L_p Inteiras, para a maioria das situações. Entretanto, seus valores de erro médio geralmente baixos e seu custo computacional superior apenas ao da medida Manhattan, favorecem a utilização dessa medida em comparação com as medidas fracionárias e *DTW*.

Em relação à medida Geodésica, foram encontrados valores de erro médio intermediários, porém, próximos aos menores valores. Entretanto, seu elevado custo computacional, somado ao fato de não ter sido constatada, para a maioria dos casos, **d.e.s** favorável quando comparada à medidas de custo computacional inferior, como as L_p Inteiras, acabam por desfavorecer a utilização dessa medida neste trabalho.

Assim, por meio da avaliação experimental, verificou-se que as medidas pertencentes à Norma L_p, utilizando valores de p inteiros entre um e três, apresentaram os menores valores de erros para a maioria dos casos, tanto para as séries artificiais quanto reais. A não constatação de **d.e.s** entre essas medidas indica que não há vantagem estatística de alguma dessas medidas sobre as demais, dessa forma, medidas que apresentam um menor custo computacional são preferenciais. Adicionalmente, verifica-se que a medida Manhattan pode ser candidata interessante como medida de similaridade a ser adotada para a maioria dos casos em que se desejam realizar previsões utilizando o algoritmo *kNN-TSP*.

A medida Canberra também pode ser considerada interessante, por apresentar um custo computacional baixo, acima somente ao da medida Manhattan, e ter alcançado resultados de previsão próximos aos dessa medida.

A medida Composta, proposta inicialmente neste trabalho, atingiu para alguns casos valores intermediários de *MAPE* próximos aos maiores valores, e em outro, o maior valor de *MAPE* para a série do estudo de caso. Desse modo, a utilização das medidas Métrica L_3, $L_{p0.1}$, *DTW*, Canberra e Geodésica, selecionadas pelos algoritmos de SA, e o inverso do valor de *MAPE* para ponderação de sua influência, não demonstraram vantagens consideráveis para a previsão de séries temporais utilizando o algoritmo *kNN-TSP*. Cabe ressaltar que essa é uma proposta inicial, sendo que várias melhorias podem ser realizadas (algumas delas serão Seção 7.2).

Os resultados do estudo de caso se assemelharam, no que diz respeito ao desempenho das medidas, aos resultados encontrados na avaliação experimental, reafirmando assim que medidas de menor custo computacional, como a Manhattan, podem ter preferencia para a maioria dos casos.

Como mencionado foi realizado um estudo comparativo entre os métodos MA, SARIMA univariado, SARIMA multivariado e o algoritmo *kNN-TSP*, tendo como foco de previsão uma série de fluxo diário de pacientes na Área de Emergência de um hospital coreano. O algoritmo *kNN-TSP* alcança, em sua melhor configuração, 08,319% de *MAPE*, tendo apresentado melhor exatidão do que a melhor configuração do método MA (12,909%). Já as duas melhores configurações do SARIMA alcançaram os menores valores de erro médio, 07,732% para o multivariado

e 07,788% para o univariado.

Cabe ressaltar que o algoritmo *kNN-TSP* é univariado, utilizando apenas a série temporal como entrada, e que apenas um parâmetro do algoritmo foi otimizado neste trabalho: a medida de similaridade. Mesmo com essas limitações, o algoritmo *kNN-TSP* apresentou apenas 00,531% de erro médio superior ao método SARIMA univariado e 00,947% de erro médio superior ao método SARIMA multivariado. Assim, foi possível verificar que o algoritmo *kNN-TSP* é competitivo com métodos tradicionais da literatura, mesmo tendo apenas um de seus parâmetros otimizados. A otimização de outros parâmetros ou a utilização de entradas multivariadas podem reduzir ainda mais o *MAPE* alcançado por esse algoritmo.

Outro ponto a ser destacado é a menor quantidade e a maior simplicidade de estimação dos parâmetros do algoritmo *kNN-TSP*, em relação à alguns dos métodos paramétricos, tais como ARIMA e SARIMA. Essa característica, somada à competitividade dos resultados desse algoritmo, torna vantajosa a sua aplicação.

Considerando os resultados da avaliação experimental empregando o algoritmo *kNN-TSP* com séries temporais artificiais e reais, e o estudo de caso comparativo, podemos concluir que:

1. As medidas da Norma L_p com valores de p inteiros variando entre um e três apresentaram, em geral, os menores valores de *MAPE* para a maioria das situações utilizando o algoritmo *kNN-TSP* para previsão de séries temporais;

2. A não existência de **d.e.s** entre as medidas L_p Inteiras, indica que a medida de menor custo computacional pode ser preferencial. Assim, a medida Manhattan pode ser de grande interesse como medida de similaridade para a previsão de dados temporais utilizando o algoritmo *kNN-TSP*. Essa medida pode ser interessante também quando não há a possibilidade de desenvolver estudo empírico para determinar a melhor medida a ser usada, ou ainda se possível uma avaliação experimental, pode ser interessante considerar apenas as medidas de menores valores de *MAPE* encontradas neste trabalho, buscando aquela que melhor se enquadre em um domínio específico;

3. A medida Canberra pode ser escolhida como segunda medida preferencial, já que possui baixo custo computacional e alcançou valores de erro médio intermediários próximos aos das medidas L_p Inteiras;

4. A medida Composta, proposta inicialmente neste trabalho, não apresentou bom desempenho, podendo mediante pesquisas mais profundas, se tornar uma medida competitiva;

5. O algoritmo *kNN-TSP* atingiu, mesmo com apenas um parâmetro otimizado, exatidão competitiva com os métodos tradicionais na literatura para a previsão de séries de fluxo diário de pacientes.

7.1 Principais Contribuições

As principais contribuições deste trabalho podem ser organizadas da seguinte maneira:

- O auxílio para a escolha otimizada do parâmetro de medida de similaridade para a previsão de séries temporais utilizando o algoritmo *kNN-TSP*;
- O maior conhecimento do comportamento do algoritmo *kNN-TSP* frente a uma grande quantidade de séries com características distintas;
- A proposta inicial de desenvolvimento de uma medida de similaridade composta;
- O desenvolvimento de um conjunto de ferramentas computacionais para a aplicação do algoritmo.

Como mencionado, o algoritmo *kNN-TSP* depende da escolha de um conjunto de parâmetros. Apesar de alguns desses parâmetros já terem sido explorados, como a função de previsão (Ferrero et al., 2009), outros ainda permanecem em aberto para pesquisa. Um desses parâmetros em aberto era a medida de similaridade, a qual foi estudada neste trabalho. Assim, os resultados deste trabalho podem auxiliar futuras pesquisas com o algoritmo *kNN-TSP*, facilitando a otimização da escolha do parâmetro de medida de similaridade, o que pode levar a previsões mais exatas.

A avaliação experimental foi realizada utilizando cinco séries artificiais de características sazonais conhecidas e características caóticas, além de uma grande quantidade de séries reais com características diversas. A utilização desse grande conjunto de séries, em especial das séries reais, permitiu conhecer melhor o comportamento do algoritmo *kNN-TSP* frente a uma grande quantidade de comportamentos de séries temporais.

Foi proposto também, um método para o desenvolvimento de medidas de similaridade compostas por outras medidas, utilizando a seleção de atributos e a ponderação da influência das medidas por meio do inverso do valor de seu erro individual.

Adicionalmente, foram desenvolvidos um conjunto de ferramentas computacionais, que incluem, além do algoritmo *kNN-TSP* e suas configurações, a análise de resultados e a geração de gráficos. O algoritmo adaptado para permitir a seleção da medida de similaridade, bem como as demais contribuições computacionais, está implementado de maneira a ser facilmente agregado ao *framework TimeSys*. Esse *framework* realiza diversas tarefas de interesse de séries temporais, como pré-processamento, previsão, agrupamento, recuperação por conteúdo entre outros.

7.2 Limitações

Ao longo do desenvolvimento do trabalho foram detectadas as seguintes limitações:

- A possibilidade da seleção de vizinhos próximos que apresentam sobreposição;
- A identificação de parâmetros para o algoritmo *kNN-TSP* não é automatizada;
- O método de seleção de medidas de similaridade para a composição da medida composta é limitado;
- A implementação do *kNN-TSP* se restringe a dados univariados;

O critério de seleção de vizinhos adotado atualmente na implementação do algoritmo *kNN-TSP* permite a seleção de vizinhos próximos que apresentem sobreposição, ou seja, permite *trivial matches*. A possibilidade de ocorrência de *trivial matches* aumenta à medida que se aumenta o tamanho da janela (w), pois a possibilidade de duas sequências adjacentes terem valores de similaridade próximos aumenta. A seleção contendo sequências adjacentes pode estar influenciando negativamente o desempenho da função de previsão utilizada com o algoritmo.

Ainda que a identificação de parâmetros no algoritmo *kNN-TSP* seja mais simples que a de métodos como ARIMA e SARIMA, essa identificação é uma tarefa delicada e que merece atenção especial. Embora tenham sido usados parâmetros sugeridos na literatura para a maioria dos casos, parâmetros como a determinação do tamanho da janela poderiam ser identificados automaticamente utilizando critérios específicos, como a função de autocorrelação. Outro parâmetro que poderia ter identificação automatizada é a quantidade de vizinhos próximos, através de critérios como a utilização de falsos vizinhos próximos.

O método empregado para a seleção das medidas que compõem a medida proposta foi restringido a utilizar apenas dados relativos a $k = 5$. Esses dados poderiam ser ampliados a outros valores de k, inclusive a combinação de todos, ou ainda restritos apenas ao valor de k que apresenta o melhor resultado, isto é, o menor valor de *MAPE*. Poderiam ainda ser utilizados outros métodos de SA, em busca de um método que possa ser mais adequado ao formato de dados empregados.

O algoritmo *kNN-TSP* implementado neste trabalho está preparado para lidar com dados univariados. Assim, esse algoritmo considera como entrada apenas os dados das séries temporais, sem considerar possíveis variáveis explanatórias.

7.3 Trabalhos Futuros

Durante o desenvolvimento deste trabalho foram encontradas diversas questões de interesse que permanecem abertas. Algumas dessas questões, que podem dar seguimento a este trabalho, são:

- O estudo da influência da quantidade de vizinhos próximos na exatidão da previsão de série temporais utilizando o algoritmo *kNN-TSP*. Dependendo das características da série, o acréscimo da quantidade de vizinhos próximos exerce influência positiva ou negativa na exatidão da previsão, sendo necessário então um estudo aprofundado dessa influência;
- A seleção de vizinhos próximos desconsiderando os *trivial matches*;
- O estudo sobre a utilização de critérios objetivos para a definição do parâmetro de tamanho da janela. Considerando que esse é um dos principais parâmetros a serem definidos para a utilização do algoritmo *kNN-TSP*, estudos sobre critérios objetivos para sua definição, como a função de autocorrelação, são de grande interesse;
- A verificação da influência das medidas de similaridades utilizada no algoritmo *kNN-TSP* em séries temporais reais de outros domínios;
- A utilização de medidas de similaridade de famílias e grupos de medidas diferentes, por exemplo, a utilização de medidas baseadas em características da série ou medidas de similaridade utilizando dimensão fractal;
- A paralelização do algoritmo *kNN-TSP*, de maneira que seja possível tomar proveito de computadores de vários núcleos, ou mesmo, de *clusters* de computadores;
- A utilização de técnicas de otimização, como algoritmos genéticos, para a busca de parâmetros do algoritmo, como a quantidade de vizinhos próximos, bem como uma possível combinação de medidas;
- A ampliação das ferramentas computacionais empregadas no desenvolvimento deste trabalho, de maneira a prover um sistema mais completo, dessa forma tornando-se mais amigável para o usuário;
- A implementação do algoritmo *kNN-TSP* utilizada neste trabalho está limitada a dados univariados. Como foi demonstrado no Capítulo 6, a adição de variáveis explanatórias pode auxiliar na exatidão de previsão de dados temporais. Desse modo, realizar modificações no algoritmo de maneira a utilizar dados multivariados pode ser de grande interesse;
- A melhora do método para o desenvolvimento da medida Composta, utilizando, por exemplo, outros métodos para a seleção das medidas a serem utilizadas para a composição, combinando o tratamento de redundância e relevância;

- A avaliação da medida de similaridade Composta, proposta inicialmente neste trabalho, considerando os conjuntos de dados artificiais e outros reais.

Referências Bibliográficas

Aggarwal, C., Hinneburg, A. and Keim, D. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space, *Database Theory* pp. 420–434.

Aguirre, L. A. (2007). *Introdução à Identificação de Sistemas: Técnicas Lineares e Não-Lineares Aplicadas a Sistemas Reais*, UFMG, Belo Horizonte.

Aikes Junior, J., Lee, H. D., Ferrero, C. A., Zalewski, W. and Wu, F. C. (2011). Estudo da Influência de Medidas de Similaridade da Norma Lp no Algoritmo kNN-TSP para Previsão de Dados Temporais, *X Conferência Brasileira de Dinâmica, Controle e Aplicações*, Águas de Lindóia.

Aitkenhead, M. J. and Cooper, R. J. (2008). Neural Network Time Series Prediction of Environmental Variables in a Small Upland Headwater in NE Scotland, *Hydrological Processes* **22**(16): 3091–3101.

Alpaydin, E. (2004). *Introduction to Machine Learning*, MIT Press, Cambridge.

Antunes, C. M. and Oliveira, A. L. (2001). Temporal Data Mining: An Overview, *KDD Workshop on Temporal Data Mining*, ACM Press, pp. 1–13.

Berndt, D. and Clifford, J. (1994). Using Dynamic Time Warping to Find Patterns in Time Series, *Workshop on Knowledge Discovery in Databases*, pp. 359–370.

Berthold, M. and Hand, D. J. (2003). *Intelligent Data Analysis*, 2 edn, Springer-Verlag, Berlin.

Brocklebank, J. C. and Dickey, D. A. (2003). *SAS for Forecasting Time Series*, 2 edn, SAS Institute, Cary.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series Forecasting*, 2 edn, Springer, New York.

Cammarota, C. and Curione, M. (2008). Analysis of Extrema of Heartbeat Time Series in Exercise Test., *Mathematical Medicine and Biology : A Journal of the IMA* **25**(1).

Chatfield, C. (2004). *The Analysis of Time Series: An Introduction*, 6 edn, Chapman and Hall/CRC, New York.

Chen, L. and Ozsu, M. T. (2003). Similarity-based Retrieval of Time Series Data Using Multi-scale Histograms, *Technical report*, School of Computer Science - University of Waterloo, Waterloo.

Chiu, B., Keogh, E. and Lonardi, S. (2003). Probabilistic Discovery of Time Series Motifs, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 493–498.

Chu, S., Keogh, E. and Hart, D.and Pazzani, M. (2002). Iterative Deepening Dynamic Time Warping for Time Series, *Proceedings SIAM International Conference on Data Mining*, Citeseer, pp. 195–212.

Cryer, J. D. and Chan, K. (2008). *Time Series Analysis: With Applications in R*, 2 edn, Springer, New York.

Derlet, R. W. and Richards, J. R. (2000). Overcrowding in the Nation's Emergency Departments: Complex Causes and Disturbing Effects, *Annals of Emergency Medicine* **35**(1): 63–68.

Deza, M. and Deza, E. (2006). *Dictionary of Distances*, Elsevier, Amsterdam.

Ding, Y., Yang, X., Li, J. and Kavs, A. J. (2010). A New Representation and Distance Measure for Financial Time Series, *2nd IEEE International Conference on Information and Financial Engineering*, pp. 220–224.

Ducklow, H. W., Doney, S. C. and Steinberg, D. K. (2009). Contributions of Long-Term Research and Time-Series Observations to Marine Ecology and Biogeochemistry, *Annual Review of Marine Science* **1**(1): 279–302.

Ehlers, R. S. (2009). Análise de Séries Temporais, *Technical report*, Departamento de Matematica Aplicada e Estatistica - Instituto de Ciências Matematicas e da Computação - Universidade de São Paulo, São Carlos.

Fabris, F., Drago, I. and Varejão, F. (2008). A Multi-measure Nearest Neighbor Algorithm for Time Series Classification, *Proceedings of the 11th Ibero-American conference on AI: Advances in Artificial Intelligence*, Springer-Verlag, Lisboa, pp. 153–162.

Felipe, J., Traina, A. and Traina, C. (2006). Perceptual Distance Functions for Similarity Retrieval of Medical Images, *Image and Video Retrieval*, Vol. 4071 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 432–442.

Ferrero, C. A. (2009). *Algoritmo kNN para Previsão de Dados Temporais: Funções de Previsão e Critérios de Seleção de Vizinhos Próximos Aplicados a Variáveis Ambientais em Limnologia*, Master's thesis, Instituto de Ciências Matemáticas e Computação (ICMC) - Universidade de São Paulo (USP/São Carlos).

Ferrero, C. A., Monard, M. C., Lee, H. D., Benassi, S. F. and Wu, F. C. (2008). Previsão da Temperatura da Água no Reservatório de Itaipu Utilizando o Método Não-Linear k-Nearest Neighbor, *III Congresso da Acadêmia Trinacional de Ciências*, Foz do Iguaçu.

Ferrero, C. A., Monard, M. C., Lee, H. D. and Wu, F. C. (2009). Proposta de uma Função de Previsão de Dados Temporais para o Algoritmo dos Vizinhos mais Próximos, *Anais do XXXV Confêrencia Latinoamericana de Informática*, Pelotas, pp. 1–10.

Freitas, A. A. (2002). *Data Mining and Knowledge Discovery with Evolutionary Algorithms*, Springer-Verlag, Berlin.

Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*, 2 edn, Morgan Kaufmann, San Francisco.

Hanias, M. P. and Curtis, P. G. (2008). Time Series Prediction of Dollar \ Euro Exchange Rate Index, *International Research Journal of Finance and Economics* (15): 232–239.

Hanslmeier, A., Kucera, A., Rybák, J. and Wöhl, H. (2004). Two-dimensional Spectroscopic Time Series of Solar Granulation, *Solar Physics* **223**(1-2): 13–26.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*, 2 edn, Prentice Hall, New Jersey.

He, X., Cai, D. and Niyogi, P. (2006). Laplacian Score for Feature Selection, *Advances in Neural Information Processing Systems* **18**: 507–514.

Hetland, M. L. (2004). A Survey of Recent Methods for Efficient Retrieval of Similar Time Sequences, *Data Mining in Time Series Databases*, World Scientific Publishing, Danvers, pp. 23–42.

Hyndman, R. and Koehler, A. (2006). Another Look at Measures of Forecast Accuracy, *International Journal of Forecasting* **22**(4): 679–688.

Jurman, G., Riccadonna, S., Visintainer, R. and Furlanello, C. (2009). Canberra Distance on Ranked lists, *Advances in Ranking*, Neural Information Processing Systems Foundation, Whistler, pp. 22–29.

Kam, H. J., Sung, J. O. and Park, R. W. (2010). Prediction of Daily Patient Numbers for a Regional Emergency Medical Center using Time Series Analysis, *Healthcare Informatics Research* **16**(3): 158.

Karunasinghe, D. S. and Liong, S.-Y. (2006). Chaotic Time Series Prediction With a Global Model: Artificial Neural Network, *Journal of Hydrology* **323**(1-4): 92–105.

Keogh, E. J. and Pazzani, M. J. (1999). Relevance Feedback Retrieval of Time Series Data, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* pp. 183–190.

Keogh, E. and Kasetty, S. (2003). On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration, *Data Mining and Knowledge Discovery* **7**(4): 349–371.

Kirchgässner, G. and Wolters, J. (2007). *Introduction to Modern Time Series Analysis*, Springer-Verlag, Berlin.

Kulesh, M., Holschneider, M. and Kurennaya, K. (2008). Adaptive Metrics in the Nearest Neighbours Method, *Physica D: Nonlinear Phenomena* **237**(3): 283–291.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley & Sons, Hoboken.

Last, M., Kandel, A. and Bunke, H. (2004). *Data Mining in Time Series Databases*, World Scientific Publishing, Danvers.

Layte, R., O'Hara, S. and Bennett, K. (2010). Explaining Structural Change in Cardiovascular Mortality in Ireland 1995-2005: A Time Series Analysis., *European Journal of Public Health* pp. 1–6.

Lee, H. D. (2005). *Seleção de Atributos Importantes para a Extração de Conhecimento de Bases de Dados*, PhD thesis, Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo.

Lee, H. D. and Monard, M. C. (2006). Seleção de Atributos Importantes para a Extração de Conhecimento de Bases de Dados, *Proceedings of the International Joint Conference X Ibero-American Artificial Intelligence Conference - Brazilian Artificial Intelligence Symposium*, Ribeirão Preto.

Lemke, C. and Gabrys, B. (2010). Meta-learning for Time Series Forecasting in the NN GC1 Competition, *World Congress on Computational Intelligence*, IEEE, Barcelona, pp. 1–5.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, *Technical Report 8*.

Lhermitte, S., Verbesselt, J., Verstraeten, W. W. and Coppin, P. (2011). A Comparison of Time Series Similarity Measures for Classification and Change Detection of Ecosystem Dynamics, *Remote Sensing of Environment* . Aceito para publicação.

Lin, J., Keogh, E., Lonardi, S., Lankford, J. P. and Nystrom, D. M. (2004). Visually Mining and Monitoring Massive Time Series, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, pp. 460–469.

Lin, J., Keogh, E., Lonardi, S. and Patel, P. (2002). Finding Motifs in Time Series, *Workshop on Temporal Data Mining, 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Citeseer, pp. 23–26.

Liu, H., Motoda, H., Setiono, R. and Zhao, Z. (2010). Feature Selection: An Ever Evolving Frontier in Data Mining, *Proc. The Fourth Workshop on Feature Selection in Data Mining*, Vol. 4, Hyderabad, pp. 4–13.

Lo, J. (2011). A Study of Applying ARIMA and SVM Model to Software Reliability Prediction, *Uncertainty Reasoning and Knowledge Engineering (URKE), 2011 International Conference on*, Vol. 1, IEEE, pp. 141–144.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*, Springer, Berlin.

Ma, J. and Perkins, S. (2003). Online Novelty Detection on Temporal Sequences, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 613–618.

Maletzke, A. G. (2009). *Uma Metodologia para a Extração de Conhecimento em Séries Temporais por meio da Identificação de Motifs e da Extração de Características*, PhD thesis, Instituto de Ciências Matemáticas e Computação (ICMC) - Universidade de São Paulo (USP/São Carlos).

McNames, J. (1999). *Innovations in Local Modeling for Time Series Prediction*, PhD thesis, Stanford.

Meyer, D. and Bucht, C. (2011). *Package Proxy: Distance and Similarity Measures*.

Mitchell, T. M. (1997). *Machine Learning*, McGraw-Hill, Boston.

Morchen, F. (2006). *Time Series Knowledge Mining*, PhD thesis, Philipps-University, Marburg.

Morettin, P. A. and Toloi, C. M. C. (2006). *Análise de Séries Temporais*, 2 edn, Edgard Blücher LTDA, São Paulo.

Motulsky, H. (1995). GraphPad InStat 3.0 User's Guide. http://www.graphpad.com. Disponível em: *www.graphpad.com*

Mueller, A. (1996). *Uma Aplicação de Redes Neurais Artificiais na Previsão do Mercado Acionária*, Master's thesis, Departamento de Pós-Graduação em Engenharia de Produção - Universidade Federal de Santa Catarina.

Nye, J. A., Bundy, A., Shackell, N., Friedland, K. D. and Link, J. S. (2009). Coherent Trends in Contiguous Survey Time Series of Major Ecological and Commercial Fish Species in the Gulf of Maine Ecosystem, *ICES Journal of Marine Science* **67**(1): 26–40.

Odan, F. K., Ferrero, C. A., Reis, L. F. R. and Monard, M. C. (2009). Análise Comparativa dos Modelos kNN-TSP e Série de Fourier para Previsão de Demanda Horária para Abastecimento de Água, *Anais do XVIII Simpósio Brasileiro de Recursos Hídricos*, Campo Grande, pp. 1–30.

Petitjean, F., Ketterlin, A. and Gançarski, P. (2010). A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering, *Pattern Recognition* **44**: 678–693.

Pyle, D. (1999). *Data Preparation for Data Mining*, Morgan Kaufmann, Califórnia.

Ratanamahatana, C. and Keogh, E. (2005). Three Myths About Dynamic Time Warping Data Mining, *Proceedings of SIAM International Conference on Data Mining*, Citeseer, pp. 506–510.

Rebbapragada, U., Protopapas, P., Brodley, C. E. and Alcock, C. (2008). Finding Anomalous Periodic Time Series, *Machine Learning* **74**(3): 281–313.

Rezende, S. O. (2005). *Sistemas Inteligentes: Fundamentos e Aplicações*, Manole, Barueri.

Roddick, J. F. and Spiliopoulou, M. (2002). A Survey of Temporal Knowledge Discovery Paradigms and Methods, *IEEE Transactions on Knowledge and Data Engineering* **14**(4): 750–767.

Rodrigues, L. C., Silva, P. P. C. D. and Linden, R. (2007). Séries Temporais no Consumo de Energia Elétrica no Estado do Rio de Janeiro, *Revista Visões* **1**.

Rodrigues, P. P. (2008). *Hierarchical Clustering of Time Series Data Streams*, PhD thesis, Faculdade de Ciências do Porto.

Russel, S. J. and Norvig, P. (2004). *Inteligência Artificial: Tradução da Segunda Edição*, Elsevier, Rio de Janeiro.

Salvador, S. and Chan, P. (2007). Toward Accurate Dynamic Time Warping in Linear Time and Space, *Intelligent Data Analysis* **11**(5): 561–580.

Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y. and Lendasse, A. (2007). Methodology for Long-term Prediction of Time Series, *Neurocomputing* **70**: 2861–2869.

Spolaôr, N. (2010). *Aplicação de Algoritmos Genéticos Multiobjetivo ao Problema de Seleção de Atributos*, Master's thesis, Centro de Matemática, Computação e Cognição – Universidade Federal do ABC.

Theodoridis, S. and Koutroumbas, K. (2009). *Pattern Recognition*, 4 edn, Academic Press, Burlington.

Traina, C., Traina, A. and Faloutsos, C. (2003). *MDE - Measure Distance Exponent Manual (Documento Interno)*.

Verplancke, T., Van Looy, S., Steurbaut, K., Benoit, D., De Turck, F., De Moor, G. and Decruyenaere, J. (2010). A Novel Time Series Analysis Approach for Prediction of Dialysis in Critically Patients Using Echo-state Networks., *BMC Medical Informatics and Decision Making* **10**: 4.

Vlachos, M., Gunopulos, D. and Das, G. (2004). Indexing Time-Series Under Condition of Noise, *Data Mining in Time Series Databases*, World Scientific Publishing, Danvers.

Wang, C. and Huang, Y. (2009). Evolutionary-based Feature Selection Approaches with New Criteria for Data Mining: A Case Study of Credit Approval Data, *Expert Systems with Applications* **36**(3): 5900–5908.

Wei, L. and Keogh, E. (2006). Semi-supervised Time Series Classification, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, New York, p. 7.

Weiss, G. M. (2004). Mining with Rarity: a Unifying Framework, *ACM SIGKDD Explorations Newsletter* **6**(1): 7–19.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2 edn, Elsevier, San Francisco.

Xie, M. and Ho, S. (1999). Analysis of Repairable System Failure Data Using Time Series Models, *Journal of Quality in Maintenance Engineering* **5**(1): 50–61.

Yan, Q., Xia, S. and Shi, Y. (2010). An Anomaly Detection Approach Based on Symbolic Similarity, *Control and Decision Conference*, IEEE, pp. 3003–3008.

Yan, W. (2007). Fusion in Multi-criterion Feature Ranking, *International Conference on Information Fusion*, IEEE, pp. 1–6.

Yankov, D., Keogh, E., Medina, J., Chiu, B. and Zordan, V. (2007). Detecting Time Series Motifs Under Uniform Scaling, *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 844–853.

Apêndice A

Característica das ST da *NN GC1*

Tabela A.1: Características das ST disponíveis pela *NN GC1*.

Id	Base de Dados	Aquisição	Tamanho	Início	Término
1.B-001	1.B	Quaternal	40	jan-1993	abr-2002
1.B-002	1.B	Quaternal	31	jan-1990	mar-1997
1.B-003	1.B	Quaternal	148	jan-1967	abr-2003
1.B-004	1.B	Quaternal	148	jan-1967	abr-2003
1.B-005	1.B	Quaternal	148	jan-1967	abr-2003
1.B-006	1.B	Quaternal	108	jan-1977	abr-2003
1.B-007	1.B	Quaternal	108	jan-1977	abr-2003
1.B-008	1.B	Quaternal	148	jan-1967	abr-2003
1.B-009	1.B	Quaternal	148	jan-1967	abr-2003
1.B-010	1.B	Quaternal	148	jan-1967	abr-2003
1.B-011	1.B	Quaternal	148	jan-1967	abr-2003
1.C-001	1.C	Mensal	48	jan-1999	dez-2002
1.C-002	1.C	Mensal	48	jan-1999	dez-2002
1.C-003	1.C	Mensal	198	set-1987	fev-2004
1.C-004	1.C	Mensal	172	jan-1990	abr-2004
1.C-005	1.C	Mensal	118	out-1993	jul-2003
1.C-006	1.C	Mensal	118	out-1993	jul-2003
1.C-007	1.C	Mensal	118	out-1993	jul-2003
1.C-008	1.C	Mensal	57	abr-1998	dez-2002
1.C-009	1.C	Mensal	227	jan-1983	nov-2001
1.C-010	1.C	Mensal	132	abr-1993	mar-2004
1.C-011	1.C	Mensal	228	mar-1986	fev-2005
1.D-001	1.D	Semanal	527	02-jan-1995	31-jan-2005

Continua na página seguinte.

Tabela A.1 – Características das ST disponíveis pela *NN GC1*.

Continuação da página anterior.

Id	Base de Dados	Aquisição	Tamanho	Início	Término
1.D-003	1.D	Semanal	437	03-jan-1997	13-mai-2005
1.D-004	1.D	Semanal	549	11-nov-1994	13-mai-2005
1.D-005	1.D	Semanal	437	03-jan-1997	13-mai-2005
1.D-006	1.D	Semanal	618	16-jul-1993	13-mai-2005
1.D-007	1.D	Semanal	618	16-jul-1993	13-mai-2005
1.D-008	1.D	Semanal	548	18-nov-1994	13-mai-2005
1.D-009	1.D	Semanal	548	18-nov-1994	13-mai-2005
1.D-010	1.D	Semanal	593	07-jan-1994	13-mai-2005
1.D-011	1.D	Semanal	594	10-jan-1994	23-mai-2005
1.E-001	1.E	Diária	377	01-jan-2005	12-jan-2006
1.E-002	1.E	Diária	377	01-jan-2005	12-jan-2006
1.E-004	1.E	Diária	466	07-fev-2003	17-mai-2004
1.E-005	1.E	Diária	716	01-jan-2002	17-dez-2003
1.E-006	1.E	Diária	502	01-jan-2002	17-mai-2003
1.E-007	1.E	Diária	502	01-jan-2002	17-mai-2003
1.E-008	1.E	Diária	747	01-nov-2003	16-nov-2005
1.E-009	1.E	Diária	747	01-nov-2003	16-nov-2005
1.E-010	1.E	Diária	654	01-jul-2003	14-abr-2005
1.E-011	1.E	Diária	654	01-jul-2003	14-abr-2005
1.F-003	1.F	Horária (5:00-24:00)	1742	02-jan-2005	29-mar-2005
1.F-004	1.F	Horária (5:00-24:00)	902	05-set-2005	19-out-2005
1.F-005	1.F	Horária (5:00-24:00)	902	05-set-2005	19-out-2005
1.F-006	1.F	Horária (5:00-24:00)	1742	02-jan-2005	29-mar-2005
1.F-007	1.F	Horária (5:00-24:00)	1742	02-jan-2005	29-mar-2005
1.F-008	1.F	Horária (5:00-24:00)	1742	02-jan-2005	29-mar-2005
1.F-009	1.F	Horária (5:00-24:00)	902	05-set-2005	19-out-2005
1.F-010	1.F	Horária (5:00-24:00)	902	05-set-2005	19-out-2005
1.F-011	1.F	Horária (5:00-24:00)	902	05-set-2005	19-out-2005

Apêndice B

Valores de *MAPE* para as Séries da Avaliação Experimental

Tabela B.1: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries artificiais para um vizinho próximo.

Artificiais											
Série	**Estatística**	**Manhattan**	**Euclidiana**	**Métrica L_3**	**$Lp_{0.1}$**	**$Lp_{0.3}$**	**$Lp_{0.5}$**	**$Lp_{0.7}$**	***DTW***	**Canberra**	**Geodésica**
STS1	**Média**	00,0015	00,0015	00,0015	00,0015	00,0015	00,0015	00,0015	00,0037	00,0017	00,0015
	Desv. Padrão	00,0023	00,0023	00,0023	00,0023	00,0023	00,0023	00,0023	00,0123	00,0028	00,0023
	Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000
	Máximo	00,0187	00,0187	00,0187	00,0187	00,0187	00,0187	00,0187	00,0811	00,0209	00,0187
STS2	**Média**	09,6758	09,6758	09,6758	09,6758	09,6758	09,6758	09,6758	38,0265	09,6758	09,6758
	Desv. Padrão	06,5685	06,5685	06,5685	06,5685	06,5685	06,5685	06,5685	33,5702	06,5685	06,5685
	Mínimo	00,6885	00,6885	00,6885	00,6885	00,6885	00,6885	00,6885	02,5614	00,6885	00,6885
	Máximo	26,9426	26,9426	26,9426	26,9426	26,9426	26,9426	26,9426	120,0507	26,9426	26,9426
STS3	**Média**	15,1964	15,3418	15,3418	15,0673	15,0673	14,9954	14,9643	54,8981	16,5991	16,0609
	Desv. Padrão	37,1086	27,8379	27,8379	37,0290	37,0290	37,0407	37,0427	58,1728	28,4241	27,3731
	Mínimo	00,0415	00,0561	00,0561	00,0415	00,0415	00,0415	00,0415	13,5229	00,3334	00,2338
	Máximo	261,0275	188,5809	188,5809	261,0275	261,0275	261,0275	261,0275	350,3710	188,5809	188,5809
STC1	**Média**	09,6648	09,6524	09,6920	09,6438	09,6280	09,7403	09,6648	20,5628	11,3984	09,6933
	Desv. Padrão	19,8997	19,9041	19,8904	19,8957	19,9013	19,8796	19,8997	57,2849	21,1495	19,8896
	Mínimo	00,0770	00,0770	00,0770	00,0770	00,0770	00,0770	00,0770	00,1848	00,0770	00,0770
	Máximo	153,9658	153,9658	153,9658	153,9658	153,9658	153,9658	153,9658	419,8998	153,9658	153,9658
STC2	**Média**	00,4838	00,4726	00,4337	00,6022	00,5347	00,5014	00,4935	00,4838	00,5093	00,3966
	Desv. Padrão	00,6251	00,6211	00,5531	00,8027	00,6603	00,6335	00,6320	00,6251	00,5619	00,3871
	Mínimo	00,0032	00,0032	00,0032	00,0032	00,0032	00,0032	00,0032	00,0032	00,0032	00,0032
	Máximo	04,0303	04,0303	04,0303	05,0844	04,0303	04,0303	04,0303	04,0303	04,0303	02,0002

Tabela B.2: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries artificiais para cinco vizinhos próximos.

Artificiais											
Série	**Estatística**	**Manhattan**	**Euclidiana**	**Métrica L_3**	**$Lp_{0.1}$**	**$Lp_{0.3}$**	**$Lp_{0.5}$**	**$Lp_{0.7}$**	***DTW***	**Canberra**	**Geodésica**
STS1	**Média**	00,0010	00,0010	00,0010	00,0067	00,0051	00,0026	00,0010	00,0140	00,0297	00,0010
	Desv. Padrão	00,0015	00,0015	00,0015	00,0172	00,0147	00,0095	00,0015	00,0217	00,0472	00,0015
	Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0002	00,0000	00,0000
	Máximo	00,0127	00,0127	00,0127	00,0827	00,0773	00,0682	00,0127	00,1038	00,1898	00,0127
STS2	**Média**	21,8642	21,8931	22,6029	22,9898	22,1302	22,2035	21,9355	35,3455	21,9617	21,8931
	Desv. Padrão	14,8771	14,9258	15,3247	15,6597	15,2902	15,0075	14,5405	20,9340	14,9394	14,9258
	Mínimo	01,5573	01,5573	01,5573	01,8220	01,5573	01,5573	01,5573	07,3126	01,5573	01,5573
	Máximo	60,8743	60,8743	60,8743	61,7673	61,3213	59,3067	59,3067	103,7948	60,8743	60,8743
STS3	**Média**	14,7772	12,4930	12,8972	15,5665	14,6272	14,6946	14,7628	67,6721	15,6741	18,7865
	Desv. Padrão	46,0933	35,5229	40,3670	48,1918	46,1323	46,1063	46,0881	151,6853	39,1072	48,6403
	Mínimo	00,1593	00,1593	00,0241	00,1593	00,1593	00,1593	00,1593	01,0607	00,0547	00,3654
	Máximo	342,4242	261,8539	299,0466	355,7659	342,4242	342,4242	342,4242	1133,0019	290,8451	361,3063
STC1	**Média**	12,4864	12,2115	11,9253	11,8670	13,0869	13,5858	13,3551	11,9303	13,3144	12,1461
	Desv. Padrão	23,2682	23,3362	21,9512	22,3074	25,7411	25,6156	25,3007	27,6317	16,4882	21,8943
	Mínimo	00,0919	00,0919	00,0643	00,0668	00,0668	00,1769	00,1769	00,0671	00,0719	00,0919
	Máximo	145,4835	145,4835	145,3459	145,4835	145,4835	145,4835	145,4835	208,2192	87,5692	145,4835
STC2	**Média**	00,5559	00,5224	00,5568	00,8540	00,6628	00,6184	00,5880	00,5660	00,8305	00,6556
	Desv. Padrão	00,8011	00,7800	00,8957	01,0774	00,8299	00,7957	00,7847	00,7561	01,0646	00,6024
	Mínimo	00,0073	00,0074	00,0074	00,0073	00,0073	00,0073	00,0073	00,0073	00,0038	00,0033
	Máximo	06,6778	06,6768	06,8501	06,0314	05,7290	06,2060	06,4398	06,0265	08,5149	03,2869

Tabela B.3: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries artificiais para dez vizinhos próximos.

Artificiais											
Série	**Estatística**	**Manhattan**	**Euclidiana**	**Métrica L_3**	**$Lp_{0.1}$**	**$Lp_{0.3}$**	**$Lp_{0.5}$**	**$Lp_{0.7}$**	***DTW***	**Canberra**	**Geodésica**
STS1	**Média**	00,0181	00,0134	00,0107	00,0290	00,0244	00,0221	00,0206	00,0177	00,0754	00,0131
	Desv. Padrão	00,0318	00,0255	00,0190	00,0406	00,0382	00,0366	00,0354	00,0276	00,0766	00,0255
	Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0001	00,0000	00,0000
	Máximo	00,1349	00,1284	00,0910	00,1476	00,1476	00,1476	00,1412	00,1148	00,4102	00,1284
STS2	**Média**	30,5229	31,1815	29,1318	30,5683	29,8606	30,3826	30,1353	39,5033	30,1032	29,8945
	Desv. Padrão	18,0229	16,1861	18,6996	19,8079	17,4654	16,7715	18,2770	26,3900	19,8372	20,4586
	Mínimo	00,6641	07,7743	01,5143	00,9185	02,6502	04,4371	00,6641	01,7333	04,7205	02,1157
	Máximo	75,2234	68,4527	70,2030	73,1782	71,5481	68,7250	75,2234	104,8800	79,2077	83,0236
STS3	**Média**	15,2911	13,5914	13,4997	15,7982	15,3734	15,2289	15,2493	57,6690	19,7962	21,5606
	Desv. Padrão	48,3509	42,9276	42,9343	48,4875	48,1461	48,1546	48,1327	146,3030	57,2744	57,4784
	Mínimo	00,0224	00,0471	00,1813	00,2634	00,2634	00,0224	00,0224	00,0621	00,5027	00,2554
	Máximo	356,5351	317,6741	317,6741	356,5351	356,5351	356,5351	356,5351	1091,7397	426,2583	426,0089
STC1	**Média**	12,5469	11,3784	11,4549	10,9093	13,0873	12,3719	12,5873	12,0029	13,7677	12,0753
	Desv. Padrão	28,1519	20,6899	21,7469	22,1898	31,0530	27,7885	28,1788	21,6573	16,4311	21,8914
	Mínimo	00,0835	00,0015	00,1813	00,0864	00,0835	00,0835	00,0835	00,0332	00,1118	00,0015
	Máximo	256,9434	171,9324	187,0545	172,5227	256,9434	256,9434	256,9434	140,9618	102,2488	187,0545
STC2	**Média**	00,7770	00,7617	00,7780	01,2290	00,9936	00,8783	00,8038	00,8037	01,2811	00,9292
	Desv. Padrão	00,9918	01,1446	01,2050	01,5614	01,2228	01,0319	00,9733	01,0625	01,5798	00,8879
	Mínimo	00,0007	00,0007	00,0202	00,0061	00,0064	00,0064	00,0064	00,0110	00,0191	00,0053
	Máximo	06,9189	08,9251	08,9251	08,1749	07,1601	06,4179	06,7833	06,3156	09,5243	05,4541

Tabela B.4: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries da *NN GC1* para um vizinho próximo.

Reais											
Base	**Cálculo**	**Manhattan**	**Euclidiana**	**Métrica L_3**	**$Lp_{0.1}$**	**$Lp_{0.3}$**	**$Lp_{0.5}$**	**$Lp_{0.7}$**	***DTW***	**Canberra**	**Geodésica**
1.B	**Média**	04,6302	04,8986	05,0524	05,3490	04,8566	04,5038	04,5484	04,9185	04,6916	04,6923
	Desv. Padrão	03,6399	04,0529	04,2057	04,2562	04,3738	03,7597	03,6879	04,1134	03,8198	03,8751
	Mínimo	00,0000	00,0000	00,0000	00,0762	00,0762	00,0000	00,0000	00,0000	00,0000	00,0000
	Máximo	17,6832	18,2254	18,2254	21,5812	21,5812	17,6832	17,6832	18,4245	16,2323	20,0571
1.C	**Média**	09,6166	09,2810	09,0316	13,0388	11,6335	11,6064	11,0187	14,8719	13,0271	09,1855
	Desv. Padrão	11,0629	10,5494	09,5959	17,9600	14,1802	14,1629	13,7043	16,3682	15,8616	10,2427
	Mínimo	00,0265	00,0272	00,0287	00,0243	00,0243	00,0243	00,0243	00,0201	00,0511	00,0118
	Máximo	57,0833	57,0833	49,0141	127,3973	82,4205	82,4205	82,4205	82,4205	125,9717	55,1495
1.D	**Média**	12,3070	12,5101	12,2886	14,1818	12,7472	11,7749	11,5568	15,8322	14,7972	12,4406
	Desv. Padrão	23,6603	25,3179	24,6425	30,2978	24,2815	21,2396	21,0988	31,0709	30,0148	23,3709
	Mínimo	00,0535	00,0000	00,0000	00,0000	00,0535	00,0535	00,0535	00,0535	00,0000	00,0000
	Máximo	181,9672	168,1034	168,1034	273,2558	181,9672	181,9672	181,9672	234,4398	182,1429	168,1034
1.E	**Média**	38,0041	27,4373	26,7985	58,5661	35,0830	38,5498	37,7416	32,9881	48,2330	59,7926
	Desv. Padrão	142,7694	51,5270	51,0977	219,4583	108,6842	139,9510	142,6062	73,6589	154,9711	146,7638
	Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,2338	00,0000
	Máximo	1600,0000	350,0000	350,0000	2000,0000	1200,0000	1600,0000	1600,0000	550,0000	1600,0000	850,0000
1.F	**Média**	19,2140	16,5344	15,8744	22,2511	22,1087	21,9252	20,1126	21,6531	16,9682	12,4269
	Desv. Padrão	60,5405	50,6060	45,3273	78,1411	79,9546	79,9284	63,1475	34,6334	50,2891	27,1412
	Mínimo	00,0000	00,0000	00,0071	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000
	Máximo	655,0980	655,0980	602,2274	1070,7317	1070,7317	1070,7317	655,0980	230,6874	613,6577	246,2036

Tabela B.5: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries da *NN GC1* para cinco vizinhos próximos.

Reais											
Base	**Cálculo**	**Manhattan**	**Euclidiana**	**Métrica L_3**	**$Lp_{0.1}$**	**$Lp_{0.3}$**	**$Lp_{0.5}$**	**$Lp_{0.7}$**	***DTW***	**Canberra**	**Geodésica**
1.B	**Média**	04,2624	04,1821	03,9896	04,5679	04,3076	04,2957	04,3358	04,2036	04,3629	04,2600
	Desv. Padrão	03,3661	03,3133	03,0388	03,4231	03,2406	03,2880	03,3621	03,2737	03,4994	03,3732
	Mínimo	00,0942	00,1466	00,1466	00,0330	00,0358	00,3681	00,0942	00,0942	00,0000	00,0312
	Máximo	16,7934	18,0109	18,0109	16,8286	16,7934	16,7934	16,7934	16,7934	17,2219	17,8343
1.C	**Média**	08,9164	09,2470	08,7133	10,1851	09,7198	09,2645	08,7069	09,7626	10,6589	09,2400
	Desv. Padrão	09,3386	09,5330	09,1912	11,4973	11,0557	09,8500	09,4871	10,5556	11,3378	10,7461
	Mínimo	00,0027	00,0557	00,0471	00,0127	00,0027	00,0027	00,0027	00,0021	00,0000	00,0038
	Máximo	43,8343	42,6158	39,3803	57,7739	57,7739	44,8253	43,8343	49,1268	49,9106	61,9282
1.D	**Média**	10,6838	10,0233	10,6493	10,3517	10,3619	10,1547	10,5198	11,2532	10,5582	09,1940
	Desv. Padrão	23,0189	22,2424	22,8784	19,7119	21,0623	22,5360	22,8387	23,2016	22,2410	18,9121
	Mínimo	00,0263	00,0191	00,0394	00,0000	00,0054	00,0000	00,0263	00,0096	00,0090	00,0000
	Máximo	152,3556	164,6552	174,4643	128,0328	145,8091	160,3279	152,3556	182,5000	167,7586	148,4483
1.E	**Média**	33,7424	29,8862	28,0565	30,1889	38,5174	37,4565	36,2955	29,8106	40,9019	51,0474
	Desv. Padrão	120,4311	98,1960	80,3676	113,2307	148,4708	141,0741	130,9136	94,7436	156,1880	126,5669
	Mínimo	00,0000	00,0570	00,0028	00,0273	00,0000	00,0510	00,0510	00,0570	00,0152	00,0473
	Máximo	1090,0000	950,0000	680,0000	1280,0000	1420,0000	1370,0000	1270,0000	960,0000	1520,0000	846,6667
1.F	**Média**	22,0724	18,8412	18,1167	23,4491	23,3005	22,8535	22,8453	20,7652	16,3929	11,0063
	Desv. Padrão	66,7014	52,9404	50,1878	72,5637	73,3975	70,2367	70,4735	31,6978	49,2091	21,8898
	Mínimo	00,0153	00,0398	00,0398	00,0039	00,0170	00,0577	00,0320	00,0896	00,0000	00,0122
	Máximo	565,0293	556,3423	552,0985	565,0293	739,0244	573,1707	580,9756	206,5104	601,0317	205,3297

Tabela B.6: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para as séries da *NN GC1* para dez vizinhos próximos.

		Reais									
Base	**Cálculo**	**Manhattan**	**Euclidiana**	**Métrica L_3**	**$Lp_{0.1}$**	**$Lp_{0.3}$**	**$Lp_{0.5}$**	**$Lp_{0.7}$**	***DTW***	**Canberra**	**Geodésica**
1.B	**Média**	04,0987	04,0456	03,9998	04,1325	04,2273	04,2049	04,1631	04,1169	04,2339	04,1137
	Desv. Padrão	03,1375	03,2924	03,3530	03,1650	03,1533	03,1039	03,1101	03,1348	03,4369	03,2095
	Mínimo	00,1031	00,0787	00,0255	00,0291	00,0964	00,1171	00,1171	00,0819	00,0190	00,1441
	Máximo	17,2058	18,2049	18,2049	16,6469	16,6469	17,2058	17,2058	17,3372	18,1825	18,3941
1.C	**Média**	09,4691	09,3172	09,4028	09,8513	09,8674	09,4455	09,6825	10,2220	10,8287	08,9714
	Desv. Padrão	10,7532	10,2949	10,4273	11,2040	11,1507	10,7135	10,6726	10,7891	11,6703	09,5026
	Mínimo	00,0338	00,0513	00,0392	00,0385	00,0385	00,0385	00,0146	00,0400	00,0248	00,0284
	Máximo	45,9717	45,3371	43,3387	51,5123	44,8587	44,8587	44,8587	50,3315	49,5936	42,9488
1.D	**Média**	10,2889	09,9691	10,0362	10,1049	10,4959	10,1452	10,2154	10,6857	10,6324	09,9163
	Desv. Padrão	21,3677	22,1168	21,2143	20,1449	21,9182	20,7137	20,8776	21,8542	22,3813	20,6518
	Mínimo	00,0213	00,0070	00,0590	00,0139	00,0438	00,0110	00,0175	00,0040	00,0052	00,0000
	Máximo	150,7377	172,5000	135,3571	133,9556	140,8929	134,5082	134,5082	175,1786	156,6222	130,4310
1.E	**Média**	33,4394	30,1056	26,8900	19,7046	35,8950	35,0471	34,0661	27,0752	41,8554	53,3438
	Desv. Padrão	117,6177	97,6436	86,7500	39,7230	136,1961	135,5137	125,3976	89,2625	150,5762	129,4903
	Mínimo	00,0652	00,0061	00,0603	00,0000	00,0547	00,1512	00,1512	00,0116	00,0622	00,0559
	Máximo	1075,0000	865,0000	810,0000	275,0000	1275,0000	1290,0000	1205,0000	885,0000	1370,0000	800,0000
1.F	**Média**	23,2458	20,7147	19,7472	24,0973	24,1319	23,9942	24,0192	25,1576	16,1956	11,0491
	Desv. Padrão	68,5159	58,4745	53,2036	69,3949	69,9203	70,3590	71,7590	43,0821	42,5864	20,3721
	Mínimo	00,0189	00,0885	00,0118	00,0378	00,0375	00,0163	00,0606	00,0130	00,0049	00,0000
	Máximo	584,4021	584,4021	472,5205	513,7246	517,3171	516,5854	604,1463	279,1304	541,4830	178,6179

Tabela B.7: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para o agrupamento das séries artificiais e das séries da *NN GC1* para um vizinho próximo.

Série	**Estatística**	**Manhattan**	**Euclidiana**	**Métrica L_3**	**$Lp_{0.1}$**	**$Lp_{0.3}$**	**$Lp_{0.5}$**	**$Lp_{0.7}$**	***DTW***	**Canberra**	**Geodésica**
Artificiais	**Média**	04,8001	04,8101	04,8102	04,8048	04,7900	04,7970	04,7792	15,0466	05,2497	04,8741
	Desv. Padrão	15,4810	13,4956	13,4962	15,4477	15,4502	15,4511	15,4502	38,0473	14,1542	13,4625
	Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000
	Máximo	261,0275	188,5809	188,5809	261,0275	261,0275	261,0275	261,0275	419,8998	188,5809	188,5809
Reais	**Média**	17,5834	15,1073	14,6781	22,5198	18,7641	18,8769	17,9011	19,4722	19,0709	17,6800
	Desv. Padrão	66,6232	40,1285	37,1992	96,7877	66,5383	73,7520	67,4121	39,1422	67,8928	59,8104
	Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000
	Máximo	1600,0000	655,0980	602,2274	2000,0000	1200,0000	1600,0000	1600,0000	550,0000	1600,0000	850,0000

Tabela B.8: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para o agrupamento das séries artificiais e das séries da *NN GC1* para cinco vizinhos próximos.

Série	**Estatística**	**Manhattan**	**Euclidiana**	**Métrica L_3**	**$Lp_{0.1}$**	**$Lp_{0.3}$**	**$Lp_{0.5}$**	**$Lp_{0.7}$**	***DTW***	**Canberra**	**Geodésica**
Artificiais	**Média**	07,1781	06,9046	07,0103	07,3763	07,3322	07,4300	07,3478	14,3607	07,4880	07,5315
	Desv. Padrão	20,1568	17,9143	18,7177	20,6430	20,7809	20,7546	20,5932	53,6264	17,4543	20,6256
	Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0002	00,0000	00,0000
	Máximo	342,4242	261,8539	299,0466	355,7659	342,4242	342,4242	342,4242	1133,0019	290,8451	361,3063
Reais	**Média**	17,6702	15,7016	15,2323	17,8653	18,8350	18,4009	18,2667	16,8521	16,4867	15,1011
	Desv. Padrão	62,9619	51,1266	45,5399	63,5405	73,1825	69,9556	67,3685	42,4589	66,8055	51,1898
	Mínimo	00,0000	00,0191	00,0028	00,0000	00,0000	00,0000	00,0027	00,0021	00,0000	00,0000
	Máximo	1090,0000	950,0000	680,0000	1280,0000	1420,0000	1370,0000	1270,0000	960,0000	1520,0000	846,6667

Tabela B.9: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para o agrupamento das séries artificiais e das séries da *NN GC1* para dez vizinhos próximos.

Série	Estatística	Manhattan	Euclidiana	Métrica L_3	$Lp_{0.1}$	$Lp_{0.3}$	$Lp_{0.5}$	$Lp_{0.7}$	*DTW*	Canberra	Geodésica
Artificiais	**Média**	08,6383	08,3632	08,0493	08,4887	08,6798	08,5988	08,5866	14,0900	09,3416	09,0939
	Desv. Padrão	23,1695	20,4122	20,5934	22,2217	23,6222	22,8716	23,1034	51,4198	23,4726	24,4024
	Mínimo	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0000	00,0001	00,0000	00,0000
	Máximo	356,5351	317,6741	317,6741	356,5351	356,5351	356,5351	356,5351	1091,7397	426,2583	426,0089
Reais	**Média**	18,0698	16,4842	15,6825	16,5969	18,8724	18,5615	18,4848	18,2020	16,5615	15,5568
	Desv. Padrão	62,9727	53,4769	48,2625	48,5135	68,4710	68,3608	66,3513	44,9100	63,2028	52,2289
	Mínimo	00,0189	00,0061	00,0118	00,0000	00,0375	00,0110	00,0146	00,0040	00,0049	00,0000
	Máximo	1075,0000	865,0000	810,0000	513,7246	1275,0000	1290,0000	1205,0000	885,0000	1370,0000	800,0000

Apêndice C

Valores de *MAPE* para a Série de Fluxo Diário de Pacientes

Tabela C.1: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para a série de quantidade de fluxo diário de pacientes para um vizinhos próximo.

Estatística	Manhattan	Euclidiana	Métrica L_3	$Lp_{0.1}$	$Lp_{0.3}$	$Lp_{0.5}$	$Lp_{0.7}$	*DTW*	Canberra	Geodésica	Composta
Média	12,7512	11,6024	10,8145	15,3280	11,6046	11,4695	12,3002	12,5262	11,1490	11,3773	15,3280
Desv. Padrão	11,7845	11,3011	10,6305	14,0232	09,9566	09,0381	10,3402	10,8975	09,8204	10,3592	14,0232
Mínimo	00,0000	00,0000	00,0000	00,0000	00,4367	00,3509	00,0000	00,0000	00,0000	00,0000	00,0000
Máximo	57,8125	57,8125	48,7685	76,5625	60,5263	35,5263	48,7685	57,8125	45,6432	48,4536	76,5625

Tabela C.2: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para a série de quantidade de fluxo diário de pacientes para cinco vizinhos próximo.

Estatística	Manhattan	Euclidiana	Métrica L_3	$Lp_{0.1}$	$Lp_{0.3}$	$Lp_{0.5}$	$Lp_{0.7}$	*DTW*	Canberra	Geodésica	Composta
Média	09,0756	09,2812	09,1852	11,6069	09,3931	09,2491	09,2348	08,8574	08,8865	09,2594	11,5988
Desv. Padrão	09,4209	09,0117	08,2110	10,5088	08,6005	08,5767	08,5767	08,1940	08,6029	09,7303	10,5124
Mínimo	00,0980	00,0851	00,1923	00,0893	00,0962	00,0000	00,0980	00,0976	00,0000	00,1770	00,0893
Máximo	51,9171	48,8083	40,0000	52,6042	45,5263	40,0000	36,4035	34,8688	42,4155	48,3938	52,6042

Tabela C.3: Valores de média, desvio padrão, máximo e mínimo de *MAPE* para a série de quantidade de fluxo diário de pacientes para dez vizinhos próximo.

Estatística	Manhattan	Euclidiana	Métrica L_3	$Lp_{0.1}$	$Lp_{0.3}$	$Lp_{0.5}$	$Lp_{0.7}$	*DTW*	Canberra	Geodésica	Composta
Média	08,6963	08,5957	08,8368	10,4977	08,6529	08,6977	08,6366	08,9547	09,4678	08,3195	10,4763
Desv. Padrão	08,5760	07,9814	08,3726	09,3870	08,9195	08,8412	08,3817	08,7557	08,6878	08,8634	09,4250
Mínimo	00,2410	00,1347	00,1347	00,0483	00,0000	00,0830	00,2542	00,0948	00,3346	00,0775	00,0483
Máximo	37,8238	34,0816	40,0518	50,4167	38,2895	39,5313	37,6042	42,3834	46,5700	41,1979	50,4167