Algoritmo kNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia

Carlos Andres Ferrero

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:

Algoritmo kNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia*

Carlos Andres Ferrero

Orientador: ***Profa. Dra. Maria Carolina Monard***

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional.

USP – São Carlos
Janeiro/2009

* Trabalho desenvolvido com auxílio do Centro de Estudos Avançados em Segurança de Barragens (CEASB) da Fundação Parque Tecnológico Itaipu (FPTI-BR) e do Programa de Desenvolvimento Tecnológico Avançado (PDTA).

Dedicatória

Aos meus Pais,
Mónica e Carlos.

Aos meus Orientadores e Amigos,
Carolina, Huei, Paulo e Renato.

Agradecimentos

À professora Maria Carolina Monard, minha orientadora, pela amizade, carinho, que me brinda nas atividades do dia-a-dia, pelo incentivo e motivação, brindados nesses anos de mestrado e pela paciência para ensinar. Para mim, um exemplo de educadora a ser seguido.

Ao professor Gustavo Batista, pelo apoio, incentivo e ensinamentos durante as atividades de mestrado e pela amizade. Também à sua esposa Cláudia por todo o apoio e incentivo.

Aos professores, orientadores e amigos Huei Diana Lee e Feng Chung Wu (Paulo), pela amizade, carinho e fraternidade que me brindam e pela inigualável companhia que representam em todos os dias e todas estas etapas maravilhosas da minha vida.

Ao professor Renato Bobsin Machado, pela amizade, fraternidade e apoio, que me brinda para crescer em cada passo deste nobre e interessante caminho.

Aos professores Fabiano e Leticia, Fernando Nogueira, Fabiana Frata e Eliane Pereira, pelo incentivo e motivação para a educação.

À professora Ana Carolina Lorena e ao professor Dimas Betioli Ribeiro, pelo incentivo e motivação.

Ao professor Homero de Cuffa, pela amizade e os momentos compartilhados. Também pelo apoio e incentivo para o desenvolvimento do trabalho.

A Simone Frederigi Benassi, pelo apoio e incentivo, e pela notável colaboração para o desenvolvimento deste trabalho. A Matheus Rometo Neto, Anderson Braga Mendes, Nelton Miguel Friedrich da Superintendência de Meio Ambiente, Divisão de Reservatórios da Itaipu Binacional, pela colaboração para o desenvolvimento do trabalho. A Jocylaine Nunes Maciel pela amizade e incentivo, e pela colaboração, e a Alexandre Jung pelo importante auxilio no trabalho.

A André Gustavo Maletzke, pela grande amizade construída nesses seis anos, e pela notável colaboração para a elaboração do trabalho de mestrado. A Bárbara, Irno e Marly, pelo carinho e pela companhia.

A Rinaldo Antonio Ribeiro Filho, pelo incentivo e pela colaboração em diferentes etapas do trabalho.

Aos meus grandes amigos e colegas do LABI, pelos momentos compartilhados como irmãos, Richardson Floriani Voltolini, Maksoel Agustín Krauspenhar Niz, Eduardo Lucas Konrad Burin, Daniel de Faveri Honorato, Joylan Nunes Maciel, Willian Zalewski, Everton Alvares Cherman, Newton Spolaôr, Sidney

Bruce Shiki, Neimar Neitzel, Adeildo Fernandes e Luiz Henrique Dutra da Costa. Às “meninas” do LABI, Bianca Espindola, Carla Dávila, Adrieli Cristina da Silva, Rafaella Aline Lopes da Silva, Dabna Hellen Tomim, Giselle Colpani, Chris Mayara dos Santos Tibes e Cecilia Noro Pfeifer.

À toda minha família, pelo apoio no meu projeto de vida.

Em especial, à minha mãe, Mónica e ao meu pai, Carlos, pelo apoio incondicional em cada etapa da minha vida, pela educação que me foi dada com carinho e dedicação e pela coragem para tomar decisões, que fizeram com que pudesse chegar até esse momento e direcionar a minha vida.

À minha avó, Celia, pelo carinho, amor e apoio incondicional em todas as etapa da minha vida.

Ao meu irmão, Damián e à minha cunhada Marisol, pelo apoio, carinho e incentivo que me brindam em cada etapa da minha vida.

A Cristina e Mário, pelo carinho e, principalmente, pelo incentivo e motivação para seguir o caminho da educação.

Ao meu tio Juan Carlos, à minha tia Yussy e aos meus primos Rafaela, Julián y Luana, pelo apoio que me brindaram desde que estou neste país e pelo carinho que sempre me brindam.

A todos meus amigos e colegas do LABIC, pelas bons momentos compartilhados nesses últimos dois anos, Bruno Magalhães Nogueira, Robson Carlos da Motta, Leonardo Jesus Almeida, Ronaldo Prati e Silvia Fini, Edson Takashi Matsubara, Márcio Basgalupp, André Rossi, Renato Ramos da Silva, Rafael Giusti, Merley da Silva Conrado, Victor Antonio Laguna Gutiérrez, Ana Trindade Winck e Ricardo Marcacini.

A Mauro Miazaki e Sidney Seiji Sato, por todos os bons momentos compartilhados e pelo apoio e incentivo.

A Beth, Laura, Ana Paula, Livia (Seção de Pós-graduação do ICMC-USP) e Marília (Seção de Eventos do ICMC-USP), pela disposição e eficiência no tratamento dos assuntos relacionado à Pós-graduação.

À professora Rosaly Mara Senapeschi Garita (Zazá), pela amizade e pelos ensinamentos pedagógicos e educacionais.

Agradeço aos professores Wu Feng Chung e Renato Bobsin Machado e a Ana Carolina Cainelli, Kamila Severo Amaral e Diogo Rafael Dammann, do PDTA, por sempre serem prestativos(as), pela colaboração e incentivo, e pelo exemplo de excelente desempenho nas atividades.

Aos meus grandes amigos de Necochea, Lucas Andrés Rabioglio, Leonardo Ganga, Pedro Ganga, Cristian Dominguez, Pablo Constantino, Federico Dominguez, Facundo Multini, Andrés Yañez, Arturo Iglesias, Walter Vogel, Marlene Constantin e Maria Lourdes Carlini, pelo apoio, incentivo e amizade.

A Gastón Guarracino, pelo apoio, incentivo e consideração.

Ao Centro de Estudos Avançados em Segurança de Barragens — CEASB — e ao Programa de Desenvolvimento Tecnológico Avançado — PDTA — da Fundação Parque Tecnológico Itaipu — FPTI-BR —, pelo auxílio por meio da linha de financiamento de bolsas de mestrado.

Resumo

A análise de dados contendo informações sequenciais é um problema de crescente interesse devido à grande quantidade de informação que é gerada, entre outros, em processos de monitoramento. As séries temporais são um dos tipos mais comuns de dados sequenciais e consistem em observações ao longo do tempo. O algoritmo *k-Nearest Neighbor - Time Series Prediction* — *kNN-TSP* — é um método de previsão de dados temporais. A principal vantagem do algoritmo é a sua simplicidade, e a sua aplicabilidade na análise de séries temporais não-lineares e na previsão de comportamentos sazonais. Entretanto, ainda que ele frequentemente encontre as melhores previsões para séries temporais parcialmente periódicas, várias questões relacionadas com a determinação de seus parâmetros continuam em aberto. Este trabalho, foca-se em dois desses parâmetros, relacionados com a seleção de vizinhos mais próximos e a função de previsão. Para isso, é proposta uma abordagem simples para selecionar vizinhos mais próximos que considera a similaridade e a distância temporal de modo a selecionar os padrões mais similares e mais recentes. Também é proposta uma função de previsão que tem a propriedade de manter bom desempenho na presença de padrões em níveis diferentes da série temporal. Esses parâmetros foram avaliados empiricamente utilizando várias séries temporais, inclusive caóticas, bem como séries temporais reais referentes a variáveis ambientais do reservatório de Itaipu, disponibilizadas pela Itaipu Binacional. Três variáveis limnológicas fortemente correlacionadas são consideradas nos experimentos de previsão: temperatura da água, temperatura do ar e oxigênio dissolvido. Uma análise de correlação é realizada para verificar se os dados previstos mantem a correlação das variáveis. Os resultados mostram que, o critério de seleção de vizinhos próximos e a função de previsão, propostos neste trabalho, são promissores.

Abstract

Treating data that contains sequential information is an important problem that arises during the data mining process. Time series constitute a popular class of sequential data, where records are indexed by time. The *k-Nearest Neighbor - Time Series Prediction — kNN-TSP* — method is an approximator for time series prediction problems. The main advantage of this approximator is its simplicity, and is often used in nonlinear time series analysis for prediction of seasonal time series. Although *kNN-TSP* often finds the best fit for nearly periodic time series forecasting, some problems related to how to determine its parameters still remain. In this work, we focus in two of these parameters: the determination of the nearest neighbours and the prediction function. To this end, we propose a simple approach to select the nearest neighbours, where time is indirectly taken into account by the similarity measure, and a prediction function which is not disturbed in the presence of patterns at different levels of the time series. Both parameters were empirically evaluated on several artificial time series, including chaotic time series, as well as on a real time series related to several environmental variables from the Itaipu reservoir, made available by Itaipu Binacional. Three of the most correlated limnological variables were considered in the experiments carried out on the real time series: water temperature, air temperature and dissolved oxygen. Analyses of correlation were also accomplished to verify if the predicted variables values maintain similar correlation as the original ones. Results show that both proposals, the one related to the determination of the nearest neighbours as well as the one related to the prediction function, are promising.

Esta dissertação foi preparada com o formatador de textos LaTeX. Foi utilizado um estilo (*style*) desenvolvido por Ronaldo Cristiano Prati. O sistema de citações de referências bibliográficas utiliza o padrão *Apalike* do sistema BibTeX.

Algumas palavras utilizadas neste trabalho não foram traduzidas da língua inglesa para a portuguesa por serem amplamente conhecidas e difundidas na comunidade acadêmica.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AR	Auto-regressivos
ARIMA	Auto-regressivos Integrados de Médias Móveis
ARMA	Auto-regressivos de Médias Móveis
CEASB	Centro de Estudos Avançados em Segurança de Barragens
d.e.s	Diferença estatisticamente significativa
DBO	Demanda Bioquímica de Oxigênio
DQO	Demanda Química de Oxigênio
EDRS	*Edit Distance on Real Sequence*
EMA	Erro Médio Absoluto
FNN	*False Nearest Neighbor*
FPTI	Fundação Parque Tecnológico Itaipu
IAP	Instituto Ambiental do Paraná
k***NN***	k*-Nearest Neighbor*
k***NN-TSP***	k*-Nearest Neighbor - Time Series Prediction*
LABI	Laboratório de Bioinformática
LABIC	Laboratório de Inteligência Computacional
LCSS	*Longest Common Subsequence*
MA	Médias Móveis

MV Média de Valores

MVR Média de Valores Relativos

RNA Rede Neural Artificial

SIA Sistema de Informações Ambientais

TimeSSys *Time Series System*

CAPÍTULO

1

Introdução

A avaliação de fenômenos temporais é uma tarefa de crescente interesse em diversas áreas do conhecimento. A contínua coleta de informações ao longo do tempo, tais como, em casos de monitoramento, tem contribuído para o surgimento de bases de dados com grandes volumes de informação sequencial, o que torna difícil a sua interpretação por seres humanos, com o objetivo de identificar padrões relevantes que permitam descobrir novos conhecimentos.

Nesse cenário, a representação desses dados como séries temporais torna possível a aplicação de uma ampla variedade de métodos que podem auxiliar na extração de informações de acordo com determinadas tarefas de interesse, como recuperação de conteúdo, agrupamento, classificação, extração de regras de associação, identificação de padrões, detecção de anomalias e previsão, entre outras. Geralmente, essas tarefas são realizadas por algoritmos que combinam ideias de diversas áreas relacionadas à matemática e à ciência da computação, como a estatística e a inteligência computacional.

A previsão de dados temporais consiste em uma das tarefas de maior interesse para muitas áreas do conhecimento, pois permite prever dados desconhecidos, a partir de um conjunto de informações conhecidas. Para isso, têm sido propostos métodos para a previsão de comportamentos lineares e não-lineares. Os métodos que permitem modelar comportamentos lineares, normalmente, assumem que os dados respeitam alguma distribuição estatística e, com base nessa informação, são definidos parâmetros de funções lineares para ajustar um modelo aos dados (Chan, 2002, p. 26–35). Porém, uma grande parte de séries temporais envolve fenômenos naturais, os quais

são não-lineares. As abordagens para modelagem não-linear, também denominadas de regressões não-paramétricas, são comumente classificadas como globais e locais (Karunasinghe e Liong, 2006). As primeiras utilizam a série temporal inteira para a construção de um modelo que represente toda a série, enquanto que as segundas utilizam somente um subconjunto de sequências, consideradas mais importantes, para estimar o valor futuro.

Uma das estratégias para contornar o problema de previsão local para comportamentos não-lineares trata da adaptação do algoritmo de aprendizado de máquina *k-Nearest Neighbor — kNN*. O algoritmo *kNN* foi proposto por Aha et al. (1991) e consiste em predizer a classe, ou rótulo, de um novo exemplo com base em exemplos similares já rotulados. A adaptação desse algoritmo para séries temporais consiste em encontrar k sequências similares dentro da série a partir de uma sequência de referência e, com base nos valores futuros das sequências similares, é realizado o cálculo do valor futuro da sequência de referência (McNames, 1998; Illa et al., 2004). Neste trabalho, o algoritmo resultante dessa adaptação é denominado *k-Nearest Neighbor - Time Series Prediction — kNN-TSP*.

Este trabalho utiliza o algoritmo *kNN-TSP* e está inserido no projeto Análise Inteligente de Dados de Séries Temporais, que está sendo desenvolvido em uma parceria entre o Laboratório de Inteligência Computacional — LABIC — da Universidade de São Paulo — USP / São Carlos —, o Laboratório de Bioinformática — LABI — da Universidade Estadual do Oeste do Paraná — UNIOESTE / Foz do Iguaçu —, a Superintendência de Meio Ambiente, Divisão de Reservatórios da Itaipu Binacional e o Centro de Estudos Avançados em Segurança de Barragens — CEASB — da Fundação Parque Tecnológico Itaipu — FPTI-BR. Pesquisadores das áreas de ciência da computação, biologia e saúde colaboram para o desenvolvimento de métodos e ferramentas que permitam auxiliar na extração de informações e conhecimentos relevantes para o monitoramento ambiental e a segurança de barragens. Na Figura 1.1 é ilustrado o projeto Análise Inteligente de Dados de Séries Temporais aplicado para as áreas de monitoramento ambiental e segurança de barragens.

O projeto contempla três etapas, ilustradas na Figura 1.1. A primeira consiste no pré-processamento dos dados temporais. A ideia é utilizar diversos métodos de limpeza e transformação de dados, bem como métodos de extração e seleção de características, com o objetivo de obter uma descrição estruturada dos dados a serem analisados, além de construir modelos matemáticos que representem o comportamento das séries temporais. Posteriormente, na Etapa 2, com base nessa descrição dos dados, pode ser aplicado o processo de mineração de dados. Os padrões extraídos nesse processo, conjuntamente

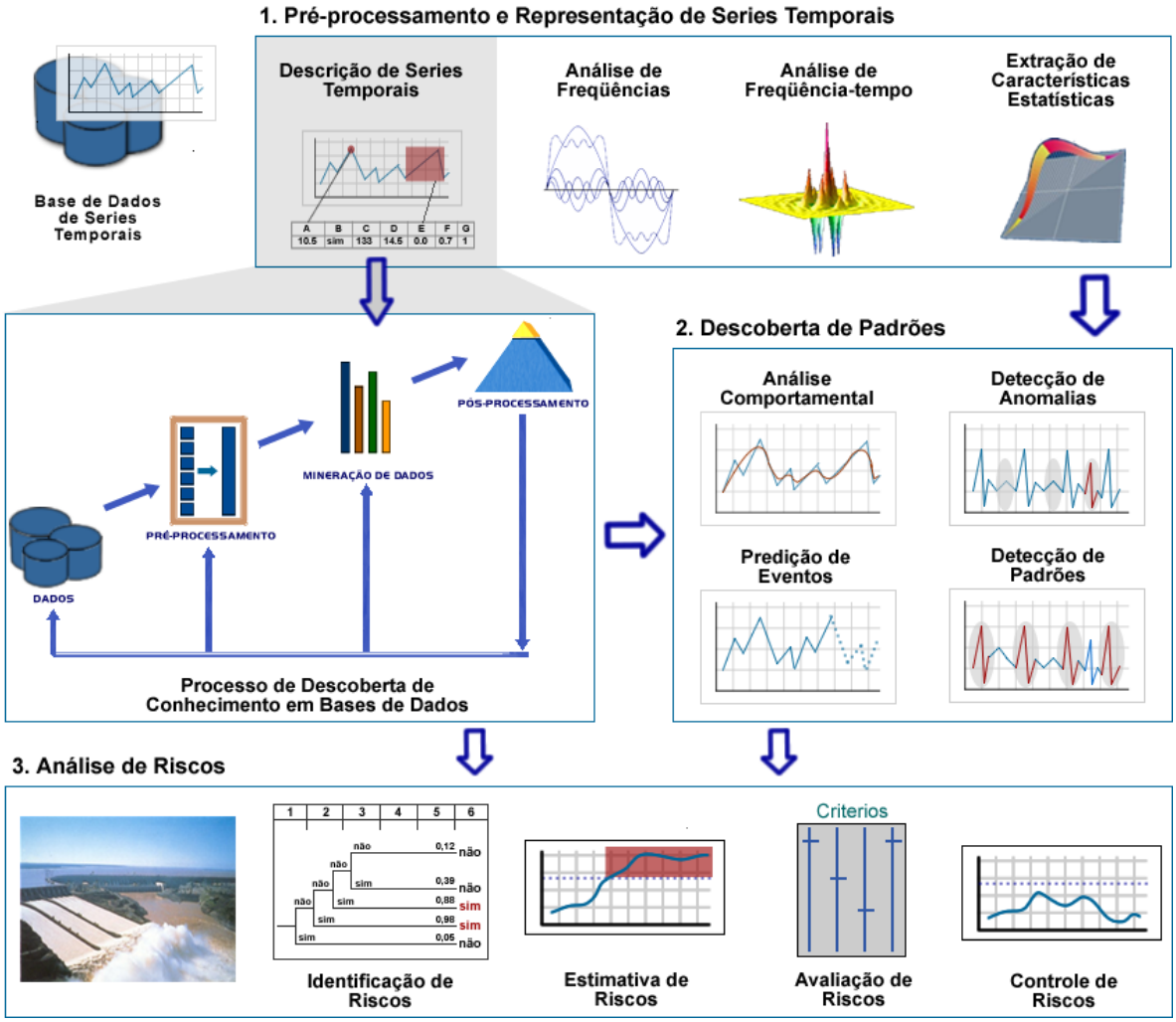


Figura 1.1: Projeto e Análise Inteligente de Dados de Séries Temporais.

com os modelos matemáticos construídos, podem ser considerados para a realização de diversas tarefas de interesse, tais como a análise comportamental de fenômenos, a detecção de anomalias, a previsão de eventos e a detecção de padrões, entre outras. Na terceira etapa, avaliação de riscos, o conhecimento extraído e os padrões encontrados na etapa anterior, podem ser utilizados para realizar as atividades de identificação, estimação, avaliação e controle de riscos.

Neste trabalho, situado na Etapa 2 do projeto, é estudado o método *kNN-TSP* na previsão de dados temporais.

1.1 Objetivos

Uma das principais vantagens do algoritmo *kNN-TSP* é a sua simplicidade. Entretanto, ainda que ele frequentemente encontre as melhores previsões para séries temporais parcialmente periódicas, várias questões relacionadas com a

determinação de seus parâmetros continuam em aberto. O objetivo deste trabalho consiste na proposta de outras abordagens para dois desses parâmetros, relacionadas com a seleção de vizinhos mais próximos e a função de previsão.

Com esse fim é proposto um critério simples para selecionar os vizinhos mais próximos que leva em consideração a similaridade e a distância temporal de modo a selecionar os padrões mais similares e mais recentes da série. Quanto à função de previsão proposta, ela tem a propriedade de manter um bom desempenho na presença de padrões em níveis diferentes da série temporal. Essas propostas estão integradas em um *framework* denominado *TimeSSys* que tem como finalidade auxiliar nas tarefas de visualização, pré-processamento, previsão, recuperação de conteúdo, entre outras.

O comportamento de k*NN-TSP* utilizando as duas abordagens propostas foi avaliado experimentalmente e foi comparado com o comportamento utilizando outras abordagens propostas na literatura. Nessas avaliações foram usadas várias séries temporais artificiais, inclusive séries caóticas. Foi também realizado um estudo de caso utilizando séries temporais reais de variáveis ambientais do reservatório de Itaipu, no qual as variáveis limnológicas fortemente correlacionadas foram utilizadas para avaliar os parâmetros.

As previsões obtidas por k*NN-TSP* foram analisadas individualmente para cada variável e uma análise de correlação foi realizada com o objetivo de verificar se os valores previstos mantêm a alta correlação dos valores medidos.

Os resultados mostram que o critério de seleção de vizinhos próximos e a função de previsão propostos neste trabalho são promissores.

1.2 Organização

O restante deste trabalho está organizado da seguinte maneira:

No Capítulo 2 é realizada uma breve introdução ao tema de séries temporais. São apresentadas a notação e as definições mais importantes consideradas no restante do trabalho, bem como aplicações de utilização da representação de séries temporais em problemas reais.

No Capítulo 3 é abordado o tema de previsão de valores em séries temporais. Primeiramente são descritos métodos lineares de previsão considerando processos estacionários e não-estacionários e, após, são apresentados métodos não-lineares de previsão. Também são apresentadas métricas para avaliação de métodos de previsão e são comentadas algumas aplicações desses métodos em problemas reais.

No Capítulo 4, são apresentados conceitos básicos de aprendizado de má-

quina e introduzido o algoritmo de aprendizado supervisionado k*NN*. O algoritmo k*NN-TSP* para previsão de dados temporais, as suas características e as fases para a execução do algoritmo também são apresentadas nesse capítulo.

No Capítulo 5 é descrita a metodologia utilizada para avaliar o desempenho do algoritmo k*NN-TSP*. Essa metodologia é constituída de três fases, as quais são descritas detalhadamente. Nesse capítulo também é descrita a ferramenta computacional desenvolvida que implementa a metodologia e o *framework* do qual faz parte.

No Capítulo 6 é apresentada a avaliação experimental do algoritmo k*NN-TSP* utilizando bases de dados temporais artificiais. São também descritos os parâmetros utilizados para a execução dos experimentos considerando diferentes configurações. A discussão dos resultados experimentais é focada na comparação entre o desempenho dos critérios de seleção de vizinhos próximos e das funções de previsão.

No Capítulo 7 é realizado um estudo de caso por meio da aplicação do algoritmo k*NN-TSP* a séries temporais de variáveis ambientais em limnologia. A avaliação dos resultados é realizada considerando a precisão de previsão para cada variável individualmente e pela correlação dos valores previstos entre pares de variáveis correlacionadas.

No Capítulo 8 são apresentadas as conclusões do trabalho, as principais contribuições, as limitações e os trabalhos futuros.

CAPÍTULO

2

Séries Temporais

2.1 Considerações Iniciais

A representação e a análise de eventos e comportamentos no tempo é uma tarefa complexa e dependente do domínio de aplicação. A utilização de métodos e técnicas de análise de séries temporais para representar informações intrínsecas dessas séries, possibilita agrupar, classificar, compreender e prever eventos futuros. Essas tarefas não são triviais e requerem que as séries temporais sejam compreendidas em termos das componentes que as constituem de modo a utilizar esse entendimento na realização dessas tarefas.

Neste capítulo, são apresentados conceitos e definições referentes ao tema de séries temporais. Também são apresentadas as componentes que constituem as séries temporais e são descritas algumas aplicações de séries temporais em problemas reais.

2.2 Notação e Definições

Uma série temporal é descrita como uma série de observações de interesse ordenadas cronologicamente e pode ser denotada como:

$$Z(t) = (z_1, z_2, \ldots, z_n) \tag{2.1}$$

onde z_t representa a observação no instante t, n o número de observações coletadas e $Z(t)$ a função que descreve a série temporal em termos de t.

Essa representação no domínio do tempo é de notável interesse devido a que a relação entre observações adjacentes no tempo torna-se uma informação importante, considerando a dependência de uma determinada observação com as observações anteriores. Outras maneiras de representação incluem o domínio das frequências em que a série é descrita em termos das frequências presentes na série temporal, e a junção de ambas, que permite evidenciar quais frequências estão presentes em determinados intervalos de tempo (Shumway e Stoffer, 2006, p. 232–245).

De acordo com Morettin e Toloi (2006, p. 49–62), as séries temporais podem ser consideradas, basicamente, como compostas por três componentes básicas:

- Tendência;
- Sazonalidade;
- Resíduo.

O entendimento de problemas temporais em termos dessas componentes proporciona informações importantes para a realização de tarefas como modelagem, previsão e simulação. Nesse sentido, cada observação que constitui uma dada série temporal está sujeita à influência dessas componentes. Assim, é possível definir $Z(t)$ em termos dessas componentes pela Equação 2.2:

$$Z(t) = T(t) + S(t) + R(t) \tag{2.2}$$

onde $Z(t)$ representa a série de observações, $T(t)$ e $S(t)$ representam, respectivamente, a tendência e a sazonalidade e $R(t)$ representa valores de resíduo, os quais podem ser considerados ruídos que seguem alguma distribuição estatística. A seguir, essas três componentes são abordadas em maior profundidade.

2.3 Componentes

A trajetória apresentada por uma série temporal é influenciada pelo conjunto de variáveis, T, S e R, que atuam com uma determinada força em cada instante de tempo. Nesse sentido, a tendência, a sazonalidade e os resíduos contribuem na formação dessa série temporal. Deve ser observado que o isolamento dessas componentes a partir de uma dada série temporal é uma tarefa complexa.

2.3.1 Tendência

A tendência corresponde à trajetória geral dos valores observados em uma série temporal. A característica principal dessa componente é o quase constante e suave movimento ao longo da série, que pode ser influenciado por vários fatores (Cortés e Zimmermann, 2006, p. 6). Para ilustrar, na Figura 2.1 é mostrado um exemplo de série de dados apresentado em (Shumway e Stoffer, 2006, p. 55), referente à mortalidade vascular semanal na cidade de Los Angeles — CA, USA, no período de 10 anos compreendido entre os anos 1970 e 1979. Nessa figura são ilustradas a série temporal e a respectiva tendência, a qual é indicada pela linha contínua preta.

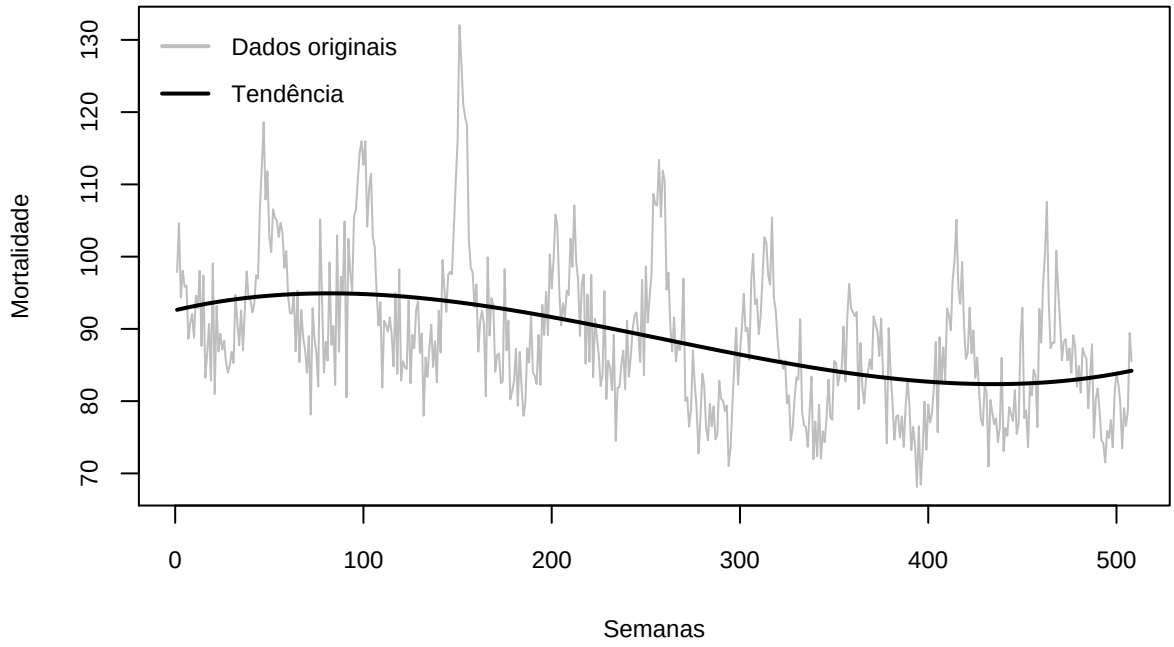


Figura 2.1: Série de dados referentes a mortalidade vascular. A linha contínua preta representa a tendência da série de dados

Existem vários tipos de tendência nos quais se baseiam os métodos para extração de tendências. Na Figura 2.2 são apresentados exemplos dos tipos de tendência linear, quadrática e cúbica.

2.3.2 Sazonalidade

A componente de sazonalidade de uma série temporal representa as flutuações de acordo com alguma característica ao longo da linha de tendência. A identificação dessa sazonalidade em uma série temporal é importante por dois motivos. Em primeiro lugar, variações sazonais podem ser informações relevantes em um determinado domínio; em segundo lugar, eliminar de

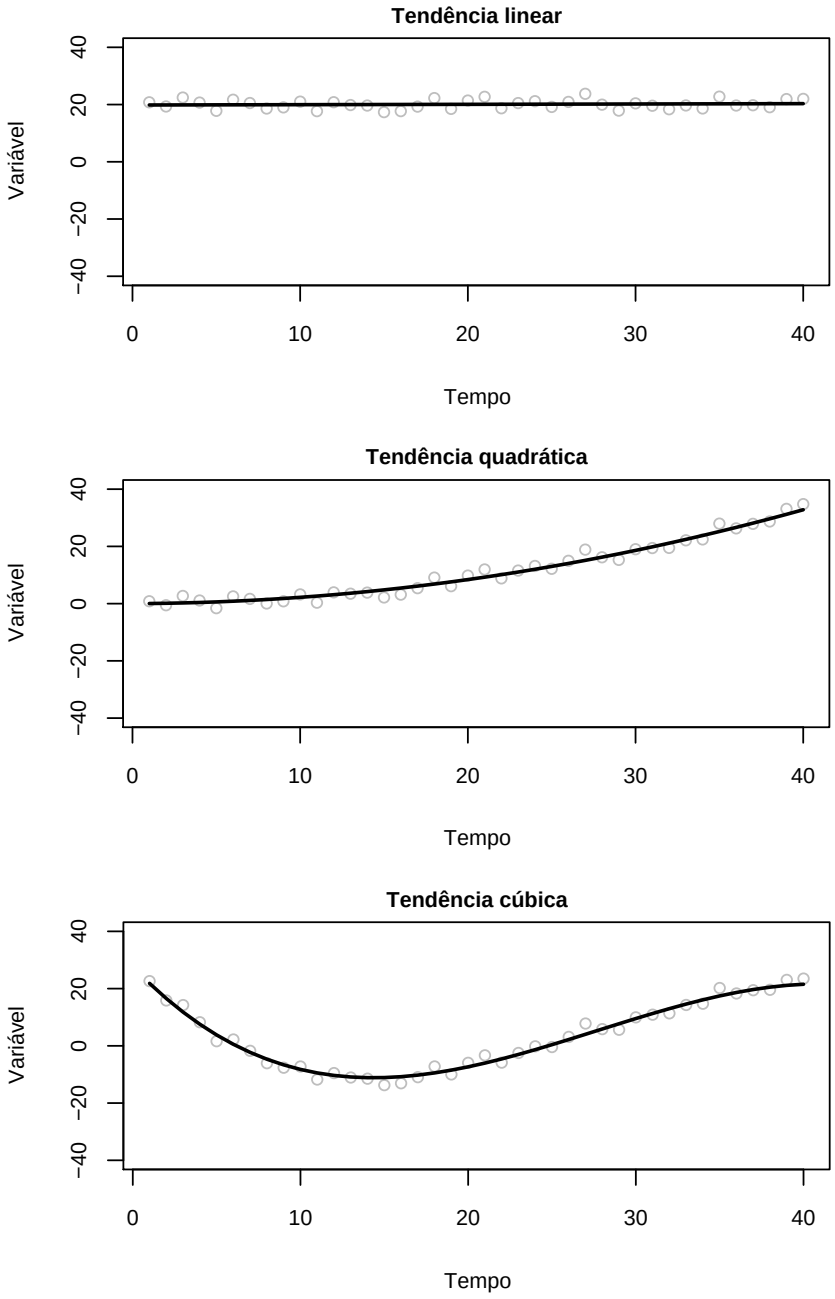


Figura 2.2: Exemplos de tendência linear, quadrática e cúbica

uma série temporal a componente de sazonalidade permite mais facilmente a identificação de fenômenos que, no caso de não serem eliminadas, implicam em uma perturbação no reconhecimento visual de eventos não-sazonais. De modo empírico, pode-se citar como exemplos de variações sazonais eventos que ocorrem de ano em ano (ou outro ciclo temporal), como o aumento de fluxo de carros na estrada perto de praias durante o verão e a procura por apartamentos para estudantes em cidades com grandes universidades nos primeiros meses do ano. Segundo Morettin e Toloi (2006, p. 66–68), existem dois tipos

de relações que podem ser observadas[1]:

- Observações entre meses sucessivos de um ano particular;
- Observações em um mesmo mês durante anos sucessivos.

Na Figura 2.3 é ressaltada a componente sazonal da série temporal da Figura 2.1.

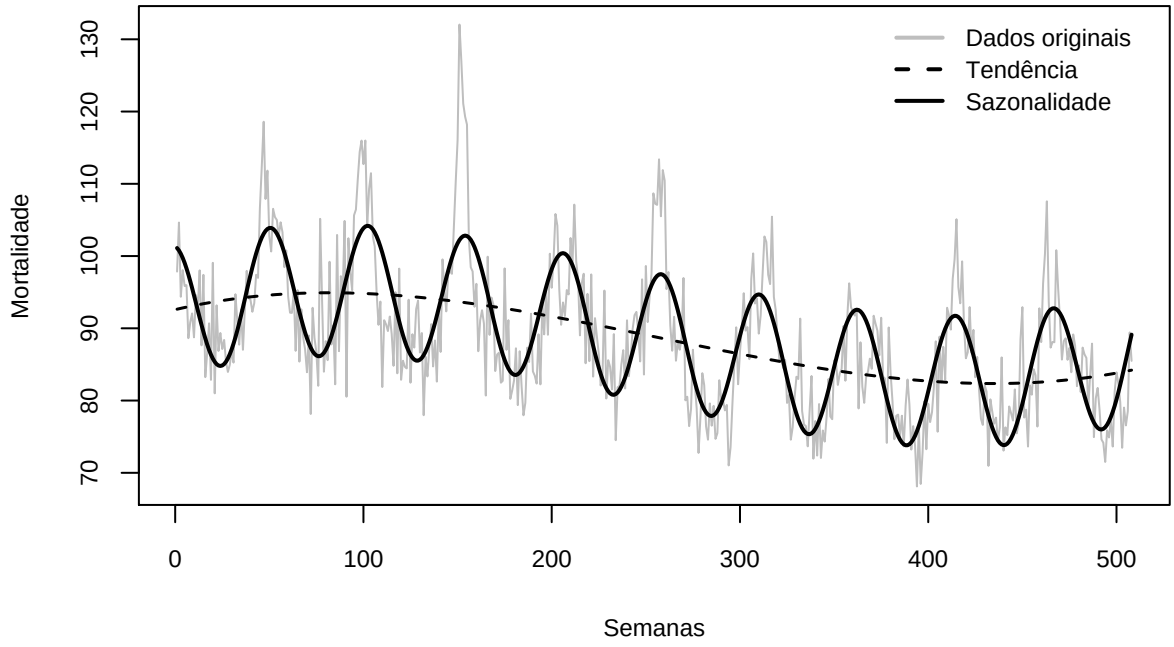


Figura 2.3: Série de dados referentes a mortalidade vascular. A linha contínua preta representa a sazonalidade da série de dados

2.3.3 Resíduo

O resíduo é a terceira das componentes que conformam as séries temporais e, muitas vezes, é considerado como o ruído da série temporal. Essa variável é de grande importância, pois a sua presença pode influenciar na identificação da tendência e da sazonalidade.

A série que representa o resíduo em cada instante t da série temporal é dada pela Equação 2.3:

$$R(t) = Z(t) - (T(t) + S(t)) \tag{2.3}$$

[1]A definição dessas relações também é válida para qualquer outro ciclo temporal, como diário, semanal, mensal, etc.

Para ilustrar o cálculo de resíduo, na Figura 2.4 é apresentado o gráfico referente ao resíduo da série temporal de mortalidade cardiovascular, eliminando as componentes de tendência e sazonalidade, apresentadas nas Seções 2.3.1 e 2.3.2. Deve ser observado que a ordenada desse gráfico encontra-se em uma escala diferente à do gráfico da Figura 2.3, a fim de ilustrar que não pode ser facilmente identificado um padrão que relacione visualmente os dados dispostos no gráfico (resíduos). Assim, é importante realizar a modelagem desses dados utilizando métodos estatísticos e/ou matemáticos para verificar a existência de padrões.

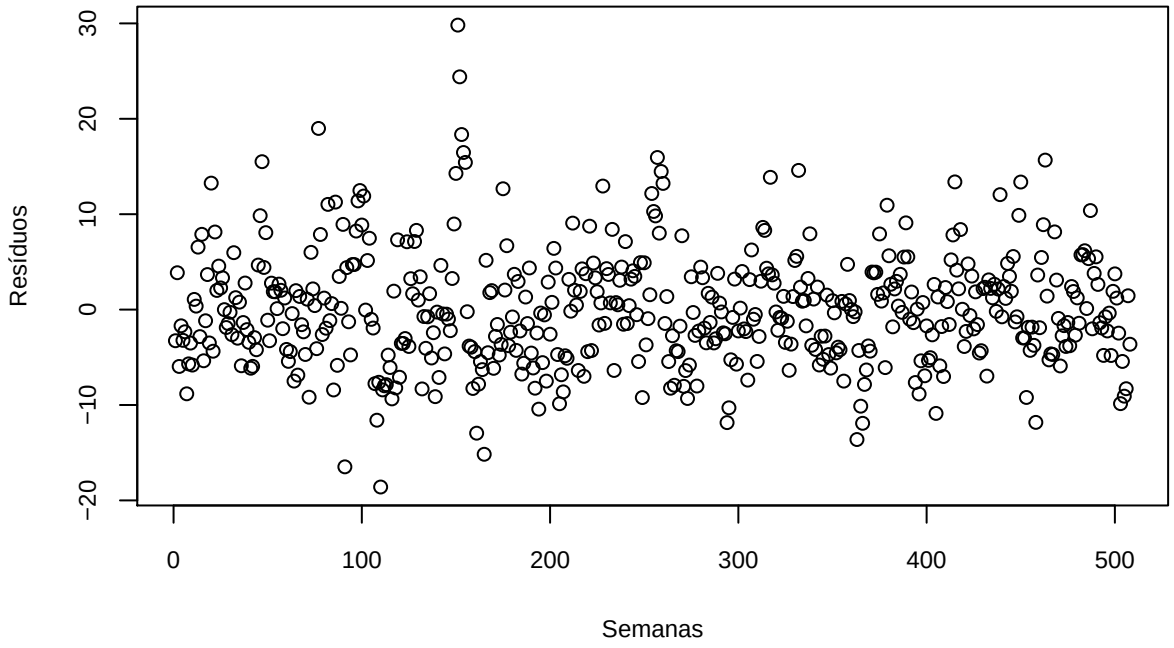


Figura 2.4: Série de dados referente ao resíduo de série temporal de mortalidade vascular

2.4 Aplicações

Séries temporais são utilizadas para descrever eventos nas mais diversas áreas. Por exemplo, um dos temas de maior relevância dentro da teoria econômica diz respeito aos ciclos econômicos. Nesse sentido, um dos objetivos dos especialistas dessa área consiste na possibilidade de prever pontos de revisão e controle das causas e efeitos de ciclos econômicos, tentando diminuir ao máximo a amplitude das oscilações. Para esse fim, os economistas devem conhecer e entender as variáveis que influenciam nas séries temporais que antecedem, coincidem e sucedem os ciclos econômicos (Cortés e Zimmermann,

2006, p. 25). Cientistas sociais também observam séries temporais de população com diversos fins, como análise das taxas de natalidade por períodos ou o número de crianças matriculadas em escolas ano a ano. Nesse modelo de tempo, epidemiologistas também analisam a distribuição e o comportamento das enfermidades que se propagam nas sociedades (Nygård e Glattre, 2003).

Em medicina, a variação de pressão sanguínea no tempo pode ser de grande interesse para avaliar drogas utilizadas no tratamento de doenças como a hipertensão (Shumway e Stoffer, 2006). Também na área médica destacam-se as aplicações de análise de séries temporais para a diferenciação entre eventos epileptogênicos (relacionados à epilepsia) e eventos normais em encefalogramas. Hadad et al. (2007) mostraram a representação de séries temporais em diversos níveis de abstração para o monitoramento de pacientes com o objetivo de identificar arritmias. A representação de séries temporais de dados está também presente em exames de manometria anorretal, que constitui um exame importante para o diagnóstico de doenças como a incontinência fecal ou a constipação intestinal. O interesse pela identificação de informações relevantes nos exames, assim como pela busca de padrões de doenças, tem incentivado pesquisadores das áreas de medicina e de ciência da computação ao desenvolvimento de trabalhos conjuntos que permitam uma análise mais completa desses exames (Cherman et al., 2008; Shiki et al., 2008).

Na área de segurança de barragens, a análise de séries temporais tem proporcionado avanços na modelagem, compreensão e previsão de fenômenos. Sáfadi (2004) utilizou a análise de séries temporais para modelar o comportamento de dados de vazão de água na represa de Furnas — PR, Brasil. Nesse trabalho foi estudado o efeito da sazonalidade, da tendência e da intervenção[2], concluindo que a modelagem da série de dados de vazão de água representa satisfatoriamente o comportamento desses dados.

2.5 Considerações Finais

A representação e a análise de problemas como séries temporais está em grande parte relacionada ao auxílio em processos de tomada de decisão. Neste capítulo, foram apresentados conceitos, definições e as componentes que constituem as séries temporais, bem como aplicações de abordagens baseadas no estudo e na análise de séries temporais presentes na literatura. A construção de modelos, com base na representação de fenômenos como séries tem-

[2]De acordo com Morettin e Toloi (1989) esse fenômeno consiste na mudança de nível ou inclinação que pode ocorrer com a série de dados durante um determinado instante de tempo, por motivos que podem ser, ou não, conhecidos — interferência. Em outros domínios do conhecimento, como na hidrologia, esse fenômeno é também denominado de efeito catástrofe.

porais, utilizando métodos matemáticos e/ou de inteligência computacional é de grande importância na busca por padrões que permitam a construção de modelos a partir desses dados. No próximo capítulo é abordado o tema de previsão de séries temporais, o qual é de interesse em diversas áreas do conhecimento, pois possibilita utilizar informação coletada no passado para construir modelos que permitam prever eventos futuros de interesse.

CAPÍTULO

3

Previsão de Dados em Séries Temporais

3.1 Considerações Iniciais

Uma das tarefas de maior interesse para qualquer área de conhecimento que esteja interessada em analisar algum fenômeno do ponto de vista temporal, consiste na previsão de valores futuros a partir do histórico da série em questão. As previsões podem ter diversos direcionamentos de acordo com o objetivo da tarefa. Por exemplo, em problemas de monitoramento, é desejável que as previsões sejam realizadas a curto prazo. Por outro lado, em problemas de perspectivas futuras populacionais ou de produtividade, as previsões poderiam ser mais úteis em períodos de longo prazo. No contexto de processos de tomada de decisões, a tarefa de previsão é uma das mais comuns e de maior interesse, sendo que o apoio computacional permite auxiliar nesse processo.

O problema de previsão em séries temporais consiste em prever valores futuros com base em valores prévios. Em outras palavras, para prever o valor de z_{n+1} de uma série $Z(t) = (z_1, z_2, \ldots, z_n)$ podem ser o utilizados os valores $z_n, z_{n-1}, z_{n-2}, \ldots, z_{n-m+1}$, onde m corresponde ao número de valores prévios da série $Z(t)$ utilizados para realizar a previsão de z_{n+1} (Sorjamaa et al., 2007). O valor futuro z_{n+1} de uma série temporal utilizando a função de previsão f_1 pode ser definido conforme a Equação 3.1.

$$z_{n+1} = f_1(z_n, z_{n-1}, z_{n-2}, \ldots, z_{n-m+1}) \tag{3.1}$$

As técnicas de previsão utilizam, em geral, duas abordagens. A primeira corresponde à utilização de métodos lineares, basicamente os modelos Auto-regressivos — AR, de Médias Móveis — MA (*moving average*), Auto-regressivos de Médias Móveis — ARMA (*Auto-regressive Moving Average*) e Auto-regressivos de Médias Móveis Integrados — ARIMA (*Auto-regressive Integrated Moving Average*). A segunda abordagem corresponde à utilização de modelos não lineares. Um exemplo desse tipo de modelos são as Redes Neurais Artificiais — RNA. A seguir é apresentada a ideia conceitual da abordagem de previsão linear e não linear, englobando os modelos que pertencem a cada uma dessas abordagens.

3.2 Modelos Lineares

Os modelos lineares para a modelagem de séries temporais podem ser divididos em modelos para processos estacionários e para processos não-estacionários, os quais são descritos a seguir.

3.2.1 Processos Estacionários

Os processos estacionários assumem que os valores das séries temporais são gerados a partir de um processo estocástico, isto é, os valores da série oscilam ao redor de uma média com variância constante ao longo do tempo. Considerando $Z(t) = (z_1, z_2, \ldots, z_n)$, um processo estocástico consiste em uma distribuição $D(z_t)$, tal que, para cada $z_t \in Z(t)$, z_t é o valor de uma variável aleatória de acordo com $D(z_t)$. Na Figura 3.1 pode ser visibilizada a série $Z(t)$, considerando cada valor da série como uma variável aleatória. Nessa figura, t_1, t_2 e t_3 descrevem três instantes de tempo, nos quais cada valor de um instante é gerado a partir de uma variável aleatória distribuida normalmente. Portanto, a série de dados é gerada a partir de uma série de variáveis aleatórias (uma para cada instante t) que têm a particularidade de serem independentes e identicamente distribuídas.

Alguns métodos podem auxiliar na modelagem desse tipo de séries temporais. A seguir são descritos o método de previsão exponencial simples e os modelos AR, MA e ARMA.

Método de previsão exponencial simples: dada uma série $Z(t)$ sem tendência e sem sazonalidade, é possível estimar o valor futuro, z_{t+1}, por meio da soma ponderada das observações do passado. Essa ponderação diz respeito à intuição de que, nesse tipo de séries temporais, as observações mais recentes têm maior influência no valor futuro do que as menos

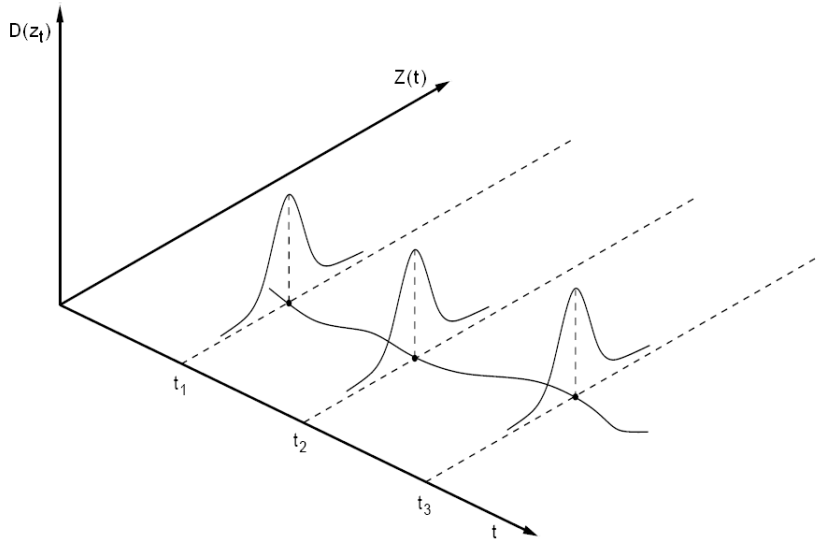


Figura 3.1: Processo estocástico como um grupo de variáveis aleatórias (Morettin, 2008, p. 26)

recentes (Ereira Souto et al., 1999). Nesse sentido, a estimativa de um valor futuro de acordo com esse método é definido pela Equação 3.2.

$$z_{t+1} = a_0 z_t + a_1 z_{t-1} + a_2 z_{t-2} + \ldots + a_{m-1} z_{t-(m-1)} \tag{3.2}$$

em que z_{t+1} representa o valor futuro calculado, m o número de observações prévias e $a_0, a_1, \ldots, a_{m-1}$ o peso de cada observação. Esse peso para qualquer instante pode ser calculado pela Equação 3.3.

$$a_j = \alpha(1-\alpha)^j \tag{3.3}$$

onde $0 < \alpha < 1$, tal que valores de α próximos de 0 produzem previsões que dependem de muitas observações passadas e valores de α próximos de 1 produzem previsões que dependem das observações mais recentes. O valor de α pode ser estimado utilizando o critério de minimização da soma dos quadrados dos erros de previsão (Ehlers, 2005, p. 37–39).

Em (Ehlers, 2005, p. 39–40) é apresentado um exemplo da aplicação do método de previsão exponencial, juntamente com a utilização do critério de minimização da soma dos quadrados dos erros de previsão, para estimar o melhor valor de α para a previsão de valores futuros. Nesse exemplo, é utilizado um conjunto de dados que contém uma série temporal da quantidade de um tipo de hormônio em amostras de sangue coletadas a cada 10 minutos de uma pessoa de sexo feminino. Na Figura 3.2 são apresentados dois gráficos. O gráfico superior representa os

valores de soma dos erros médios quadráticos para cada valor de α. O menor valor de erro foi para $\alpha = 0,945$, o qual foi utilizado para realizar a previsão dos valores da série temporal.

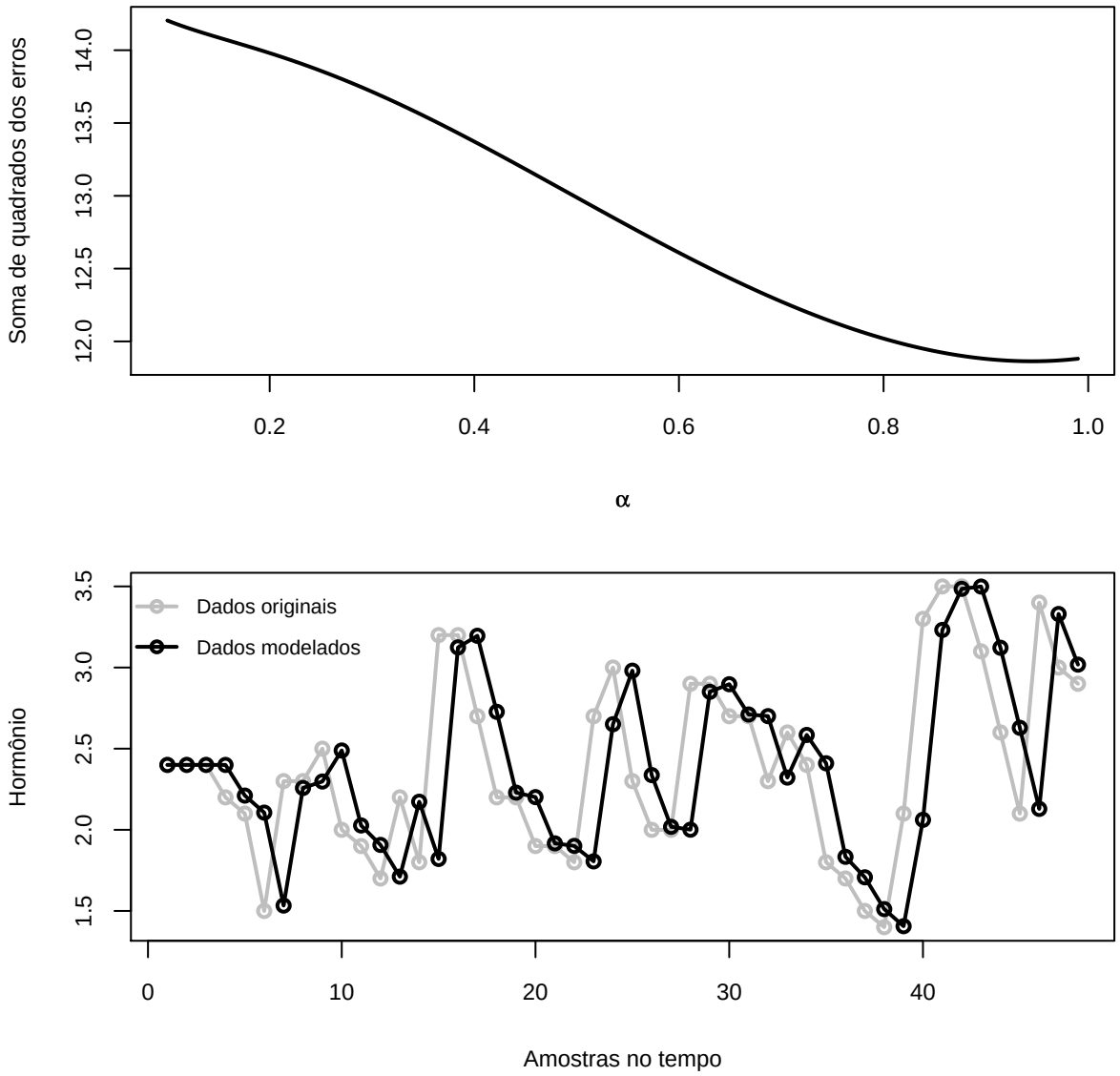


Figura 3.2: Aplicação do método exponencial para a previsão de valores futuros. O gráfico superior representa a busca pelo melhor valor do parâmetro de previsão e o gráfico inferior representa a previsão dos dados

Modelo auto-regressivo — AR: consiste em estimar o valor z_{t+1} usando uma combinação linear de p valores já observados, incluindo também um valor e_t, considerado como um ruído branco (erros distribuídos normalmente, com média zero, variância constante e não-correlacionados), como mostra a Equação 3.4.

$$z_{t+1} = e_t + \sum_{j=0}^{p-1} \alpha_j z_{t-j} \tag{3.4}$$

onde p é o número de observações prévias consideradas e os coeficientes $\alpha_0, \alpha_1, \ldots, \alpha_{p-1}$, ponderam esses valores observados. Esses coeficientes são determinados utilizando técnicas de minimização de erro.

Os modelos AR são limitados por considerar a existência de uma relação linear entre os elementos da sequência, baseando-se na hipótese de que a série em questão é uma série estacionária ou que a média e a variância das observações não variam ao longo do tempo.

Modelo de médias móveis — MA: consiste em um modelo em que cada valor futuro pode ser calculado pela combinação linear dos sinais de ruídos $e_t, e_{t-1}, \ldots, e_{t-q}$, aleatórios e independentes entre si. O cálculo de um valor futuro de acordo com o modelo de MA é definido pela Equação 3.5.

$$z_{t+1} = e_t - \sum_{j=0}^{q-1} \beta_j e_{t-j} \tag{3.5}$$

onde q representa o número de observações consideradas. Os valores dos coeficientes $\beta_0, \beta_1, \ldots, \beta_{q-1}$ também devem ser ajustados de acordo com algum método de minimização de resíduos. Tanto o modelo de MA como o modelo de ARMA, o qual é apresentado a seguir, apresentam as mesmas limitações do modelo de previsão auto-regressivo.

Modelo auto-regressivo de médias móveis — ARMA: consiste em uma combinação entre o modelo auto-regressivo e o de médias móveis. A Equação 3.6 define a previsão de um valor futuro de acordo o modelo ARMA.

$$z_{t+1} = e_t + \sum_{j=0}^{p-1} \alpha_j z_{t-j} - \sum_{j=0}^{q-1} \beta_j e_{t-j} \tag{3.6}$$

onde p corresponde ao parâmetro do modelo auto-regressivo e q corresponde ao parâmetro do modelo de médias móveis.

3.2.2 Processos Não-estacionários

No mundo real, muitas séries temporais não são estacionárias. A maioria das séries, na prática, contém algum tipo de tendência ou sazonalidade como parte do próprio comportamento. Entretanto, técnicas podem ser aplicadas

para tornar séries não-estacionárias em séries estacionárias, como o método de diferenciação (Ehlers, 2005, p. 6).

O modelo auto-regressivo integrado de médias móveis permite uma modelagem mais adequada para séries não-estacionárias que não tenham comportamento explosivo e que apresentem alguma homogeneidade no comportamento não-estacionário. O modelo ARIMA é aplicado em função dos parâmetros p, q, r. Os dois primeiros foram apresentados na Seção 3.2.1 e o parâmetro r consiste na ordem do método de diferenciação a ser aplicado antes do ajuste do modelo ARMA. Em outras palavras, supõe-se que a r-ésima diferença da série temporal em questão pode ser representada por um modelo ARMA, isto é, como um processo estacionário (Ehlers, 2005, p. 24).

Uma variação do modelo ARIMA para a modelagem de séries temporais com sazonalidade é o modelo Sazonal Auto-regressivo de Médias Móveis Integrado — SARIMA (*Sazonal Auto-regressive Integrated Moving Average*). Um exemplo de aplicação desse modelo é apresentado em (Ehlers, 2005, p. 42–45), utilizando uma série temporal do total de mortes por acidentes nos Estados Unidos, desde janeiro de 1973 até dezembro de 1978. A partir de 1979, os valores para os primeiros seis meses daquele ano foram previstos. Na Figura 3.3 são apresentados os valores da série temporal e os valores previstos. É importante notar que os valores observados no período previsto encontram-se dentro de um intervalo de confiança, e que os valores previstos, indicados na imagem por círculos preenchidos, estão dentro desse intervalo. Isso indica que, dentro desse intervalo de confiança, o modelo teve um bom desempenho de previsão.

3.3 Modelos Não-lineares

Algumas séries temporais apresentam comportamentos que dificilmente podem ser modelados com precisão utilizando os modelos apresentados anteriormente (Camilleri, 2004). Essas séries, normalmente, apresentam possibilidades de previsão a curto prazo. Métodos para prever valores futuros dessas séries são denominados métodos por espaço de estados. A ideia básica desses métodos é uma relação de funções entre o estado atual, denominado S_t, e o estado futuro, definido por S_{t+h}, onde h representa o número de valores a serem previstos em relação ao valor atual e S_t e S_{t+h} correspondem a estados w-dimensionais do sistema no instante t, contidos no espaço de estados S. A Equação 3.7 define o processo de previsão de valores futuros de um método não-linear.

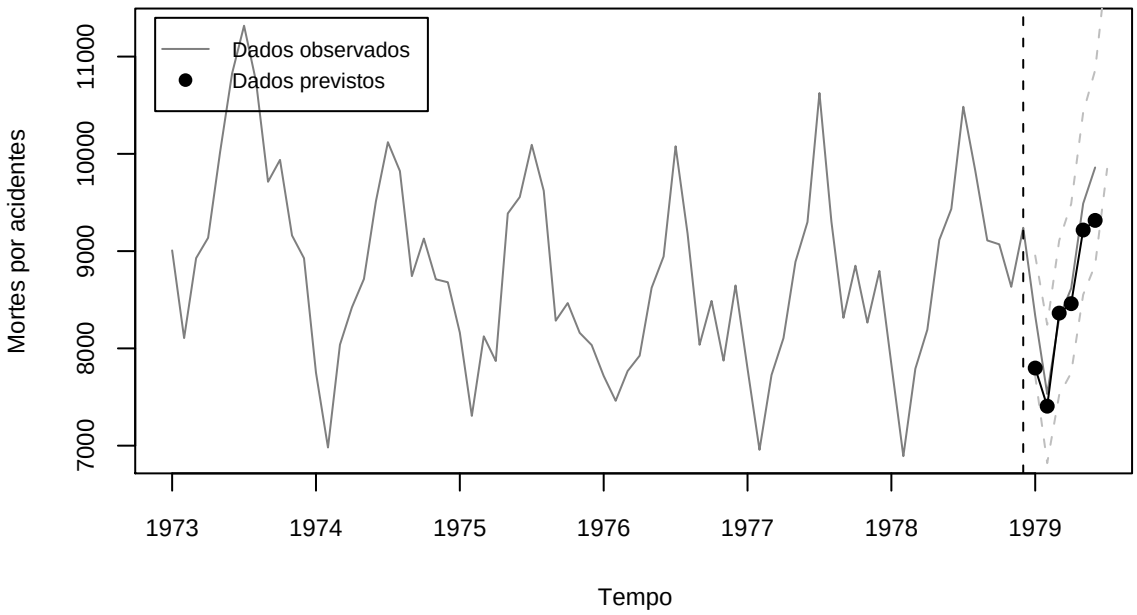


Figura 3.3: Previsões da série temporal de mortalidade por acidentes nos Estados Unidos aplicando o modelo de previsão SARIMA

$$S_{t+h} = f(S_t) \tag{3.7}$$

onde $f(S_t)$ é a função de previsão a partir do estado S_t.

De acordo com Karunasinghe e Liong (2006), existem duas estratégias para obter uma aproximação $f(S_t)$ da função ideal de previsão. A primeira consiste em realizar uma aproximação global, *i.e.*, são utilizados todos os dados observados para a previsão do valor futuro e, a segunda, consiste em realizar uma aproximação local, que utiliza somente os dados observados mais importantes para a previsão do valor futuro. As aproximações globais podem ser realizadas por meio de redes neurais artificiais, funções polinomiais e funções racionais. Por outro lado, na estratégia de aproximações locais, a série temporal é subdividida em subsequências menores. Desse modo, o conjunto de subsequências mais próximas do estado atual, considerando alguma medida de similaridade, é utilizada para o cálculo do valor futuro. Comumente, o cálculo do valor futuro é realizado pelo cálculo de médias locais ponderadas e não-ponderadas.

A seguir são descritos alguns métodos para realizar aproximação local e global de uma série temporal.

Aproximação local — método de previsão não-linear: nesse método apenas os estados mais próximos ao estado atual são utilizados para a previsão do valor futuro. Para prever o estado futuro S_{t+h} é utilizada uma me-

dida de similaridade para encontrar, no espaço de estados, os vizinhos mais próximos do estado S_t. Uma lista S', de cardinalidade $|S'|$ é criada contendo os estados mais próximos de S_t. Assim, usando S' podem ser projetados um ou mais valores à frente de S_t baseado nos valores à frente dos estados similares. Uma função $f(S')$ é utilizada para estimar o valor de S_{t+h}. Dentre as diversas funções, o valor da média dos valores dos estados é uma das mais aplicadas. A Equação 3.8 define o cálculo do valor futuro utilizando o método de aproximação local aplicando o critério da média.

$$S_{t+h} = \frac{\sum_{j=1}^{|S'|} S_{i+h}}{|S'|}, S_i = S'_j \tag{3.8}$$

Outro critério que pode ser utilizado na previsão consiste em utilizar uma função polinomial, na qual o mapeamento é realizado por cada estado da série temporal. Em (Karunasinghe e Liong, 2006) esse critério é descrito com maiores detalhes.

Aproximação global — redes neurais artificiais: esses modelos, que não necessitam de conhecimento prévio do domínio, resultam na criação de uma relação entre os valores de entrada do modelo, que nesse caso seriam valores observados de uma série temporal, e a saída, isto é, o valor futuro estimado. Essa relação é encontrada por meio de uma etapa de treino para aprendizado. As RNA são caracterizadas pela sua arquitetura, a qual consiste em um conjunto de nós organizados com uma disposição particular, em que a informação flui na direção das entradas para a saída. As redes denominadas *Multi Layer Perceptrons* e comumente chamadas de *Sigmoidal Networks*, *Feed foward Neural Networks* e *Back-propagation Networks*, têm proporcionado bons resultados para funções de aproximação e para reconhecimento de padrões, sendo utilizada em 90% das aplicações de RNA (Karunasinghe e Liong, 2006).

De acordo com a Equação 3.7, em que S_t e S_{t+h} são estados w-dimensionais que descrevem o estado do sistema no tempo t e $t+h$ respectivamente, o problema consiste em encontrar uma boa aproximação da função $f(S_t)$ que estime precisamente o valor futuro. Nesse caso, o interesse consiste em obter apenas a última componente do vetor do estado S_{t+h}, que é z_{t+h}, portanto essa procura tem como objetivo encontrar uma função de mapeamento $f : \Re^w \Rightarrow \Re$, ao invés de $f : \Re^m \Rightarrow \Re^w$. As RNA permitem receber como entrada um vetor w-dimensional retornando como saída um valor escalar. Desse modo, é utilizado $S(t)$ como entrada da

rede e retornada a estimativa do valor de $z_{t+h} \in S_{t+h}$.

3.4 Aplicações

Como mencionado, existem diversas maneiras de modelar séries temporais. Alguns métodos permitem modelar algumas séries com maior exatidão do que outras, dependendo da natureza dos comportamentos nas diversas áreas. A seguir são apresentadas aplicações de métodos lineares e não-lineares em problemas reais.

3.4.1 Modelos Lineares

A área de análise de séries temporais tem proporcionado avanços na modelagem, compreensão e previsão de comportamentos que apresentam característica linear.

Em (Slini et al., 2002) uma análise estatística foi utilizada para desenvolver um aplicativo de previsão de variáveis ambientais. Foi utilizado o modelo ARIMA com o intuito de prever a concentração máxima de ozônio na cidade de Atenas, na Grécia. Nesse trabalho, a metodologia Box-Jenkins (Box et al., 1994) para a escolha do modelo ARIMA foi aplicada a uma série temporal com observações a respeito da qualidade do ar durante um período de nove anos. O resultado da modelagem foi satisfatório, porém os autores apontam alta sensibilidade na previsão de alarmes. Os autores sugerem utilizar uma análise mais completa de avaliação a partir da análise de dados de qualidade do tempo e do ar, para prever a concentração de ozônio.

Em (Guerra et al., 1997) foi aplicado um modelo linear ARIMA para prever a inflação na Venezuela. Nesse trabalho, foram realizadas projeções a partir de modelos construídos com o objetivo de prever valores a curto e a longo prazo. Inicialmente, foi realizada uma modelagem orientada ao Índice de Preços do Consumidor. Posteriormente, foi incluído mais um parâmetro referente à influência do setor externo na inflação e foi ajustado um modelo ARIMA aos dados, o qual apresentou uma modelagem conhecida pelos especialistas. O modelo foi validado com um alto grau de significância para projetar o comportamento futuro da inflação na Venezuela.

Em (Sáfadi, 2004) foi analisado o comportamento de uma série de vazão de água na represa de Furnas — PR, Brasil, e estudado o efeito da sazonalidade, da tendência e da intervenção. Para a análise foram considerados modelos com e sem presença de intervenção. A partir de uma série de dados coletados diariamente no período de janeiro de 1963 até dezembro de 1994, foi

aplicado o modelo SARIMA, que é apropriado para a modelagem de séries com comportamento sazonal. Para diminuir a dimensionalidade do problema foi utilizada a média mensal de dados. O autor do trabalho conclui que o ajuste do modelo, juntamente com a incorporação de um parâmetro de intervenção, fornece informações relevantes para uma análise satisfatória.

3.4.2 Modelos Não-lineares

O interesse pela modelagem de comportamentos temporais da natureza é uma questão emergente. Os comportamentos da natureza têm, comumente, um comportamento não-linear, em muitos casos considerado caótico.

Em (Karunasinghe e Liong, 2006) é apresentado um estudo comparativo das três técnicas de modelos não-lineares de previsão mencionadas anteriormente: redes neurais artificiais e as técnicas de aproximação local por média e por polinômios. Inicialmente, foi aplicada a tarefa de previsão em uma série temporal artificial amplamente utilizada para medir o desempenho de novos métodos, denominada série de Lorenz (McNames, 1998; Kulesh et al., 2008). Essa série foi utilizada com e sem inserção de ruído. Posteriormente, os modelos de previsão foram aplicados em duas séries temporais reais de fluxo de água em rios. Nesse estudo foi concluído que as RNA apresentaram melhor precisão em relação aos modelos de aproximação local, na previsão de 1, 3 e 5 dados futuros, para as quatro séries temporais utilizadas.

Tang e Fishwick (1993) estudaram métodos de redes neurais como modelos para previsão em séries temporais. Nesse estudo foi comparado o método ARIMA com o método de RNA para séries de memória curta e memória longa[1]. O trabalho indicou que para séries de memória longa ambos métodos produzem resultados similares. No caso de séries temporais de memória curta, os métodos de RNA tiveram melhor desempenho do que o método ARIMA, permitindo concluir que as RNA são modelos adequados para a previsão de dados temporais.

Daliakopoulosa et al. (2004) estudaram a construção das arquiteturas de RNA para prover ferramentas robustas na área de águas subterrâneas, com o objetivo de auxiliar tanto na modelagem de comportamentos como na previsão. Sete tipos diferentes de RNA para a previsão de níveis de águas subterrâneas foram utilizadas para identificar uma arquitetura de RNA que permitisse simular o comportamento de tendência decrescente do nível das águas e possibilitar previsões aceitáveis para 18 meses. Os experimentos foram realizados a partir de dados do vale de Messara da ilha de Creta, na Grécia,

[1]Séries temporais de memória curta e longa, referem-se à quantidade de observações que são consideradas na previsão de um valor futuro.

onde, pela exploração de recursos nessa região nos últimos 15 anos, o nível de águas subterrâneas tem decrescido paulatinamente. Os resultados do trabalho mostram que a melhor previsão foi realizada pela RNA *Feed-foward Neural Network* treinada com o algoritmo Levenberg-Marquardt.

Outro trabalho importante relacionado à aplicação de modelos não-lineares para previsão de valores futuros de séries temporais foi realizado por Aitkenhead e Cooper (2008). Nesse trabalho são realizadas previsões a curto prazo de séries temporais de variáveis ambientais, como taxa de fluxo de águas, coletadas no nordeste da Escócia. A previsão do comportamento futuro dessas variáveis ambientais permite prever a ocorrência de desastres ambientais ou o descobrimento de fenômenos e eventos interessantes. Foi utilizado um método baseado em RNA, treinada com valores observados da série temporal, para dar suporte a avisos sobre eventos futuros nessa região. De acordo com os autores, obteve-se um sistema rápido e efetivo, com o qual poderia ser implementado um pacote *WEB* de previsão e monitoramento.

3.5 Considerações Finais

Diferentes abordagens podem ser utilizadas para realizar a previsão, as quais dependem das características das séries temporais que representam o fenômeno em avaliação. As duas abordagens principais, métodos lineares e não-lineares, bem como alguns modelos utilizados nessas abordagens, foram apresentados neste capítulo. Foram também descritas várias aplicações em problemas reais que utilizam modelos dessas duas abordagens. No próximo capítulo é apresentado o algoritmo de aproximação local para previsão de dados temporais utilizado, bem como as as abordagens, propostas neste trabalho, para a aplicação do algoritmo.

CAPÍTULO

4

Algoritmo k-*Nearest Neighbor* para Previsão de Séries Temporais

4.1 Considerações Iniciais

A adaptação de métodos provenientes da área de aprendizado de máquina no contexto de análise de séries temporais tem possibilitado a aproximação dessas áreas. Especificamente em problemas de previsão têm sido aplicados, por exemplo, métodos de aprendizado baseados em redes neurais artificiais (Frank et al., 2001; Karunasinghe e Liong, 2006; Aitkenhead e Cooper, 2008), *support vector machines* (Bray e Han, 2004), entre outros. Neste capítulo são apresentados inicialmente alguns conceitos básicos da área de aprendizado de máquina, descrito o algoritmo de aprendizado k*NN* convencional e a adaptação desse algoritmo para a previsão de dados temporais. São também apresentadas as propostas deste trabalho, referentes a dois parâmetros desse algoritmo. Mais especificamente, os parâmetros relacionados com a seleção dos vizinhos mais próximos e a função de previsão.

4.2 Conceitos Básicos de Aprendizado de Máquina

A área de aprendizado de máquina tem como objetivos desenvolver técnicas computacionais que permitam simular o processo de aprendizado e a construção de sistemas capazes de adquirir conhecimento automaticamente (Mitchell, 1997). Usualmente, algoritmos de aprendizado utilizam experiências anterio-

res, denominadas casos ou exemplos, para auxiliar o processo de tomada de decisão e melhorar seu desempenho.

De acordo com a característica desses casos ou exemplos têm-se três diferentes modos de aprendizado: supervisionado, não-supervisionado e semi-supervisionado. O que distingue esses modos de aprendizado é a presença ou não do atributo classe, que rotula os exemplos do conjunto de dados fornecido ao algoritmo, denominado conjunto de treinamento. No aprendizado supervisionado, esse rótulo é conhecido, enquanto que no aprendizado não-supervisionado os exemplos não estão previamente rotulados. Já no aprendizado semi-supervisionado, o conjunto de treinamento consiste de uns poucos exemplos rotulados e muitos não rotulados (Chapelle et al., 2006).

Em geral, o conjunto de treinamento é representado por uma estrutura denominada tabela atributo-valor. Na Tabela 4.1 é mostrada essa estrutura para aprendizado supervisionado, utilizado no desenvolvimento deste trabalho.

	A_1	A_2	$\dots$	A_M	Classe (Y)
E_1	x_{11}	x_{12}	$\vdots$	x_{1M}	y_1
E_2	x_{21}	x_{22}	$\vdots$	x_{2M}	y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
E_N	x_{N1}	x_{N2}	$\vdots$	x_{NM}	y_N

Tabela 4.1: Representação do conjunto de dados por meio da tabela atributo-valor

O conjunto de treinamento para um algoritmo de aprendizado supervisionado consiste, usualmente, de um conjunto E de N exemplos (ou casos) de treinamento $E = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N)\}$ rotulados com os valores y de uma função f desconhecida, $y = f(\boldsymbol{x})$, onde os valores $\boldsymbol{x}_i$ são vetores da forma $(x_{i1}, x_{i2}, \dots, x_{iM})$ cujos componentes são valores discretos ou contínuos relacionados ao conjunto de atributos $X = \{A_1, A_2, \dots, A_M\}$. Ou seja, x_{il} denota o valor do atributo A_l do exemplo i. Dado esse conjunto de exemplos de treinamento, o algoritmo constrói uma hipótese hip que deve aproximar a verdadeira função f, tal que, dado um novo exemplo $\boldsymbol{x}$, $hip(\boldsymbol{x})$ prediz o valor y correspondente. Para valores nominais dos rótulos $y_1, y_2, \dots, y_N$ o processo é denominado classificação, enquanto que para valores numéricos o processo é denominado regressão.

A qualidade de previsão de algoritmos supervisionados é avaliada utilizando um conjunto de exemplos rotulados disjunto do conjunto de treinamento, o qual é denominado de conjunto de teste.

Em geral, algoritmos supervisionados podem ser do tipo *eager* ou *lazy*.

Algoritmos *eager* usam o conjunto de exemplos de treinamento para construir a hipótese hip. Após construída a hipótese, os exemplos de treinamento são descartados já que somente hip é necessária para predizer o valor y de um novo exemplo $\boldsymbol{x}$. Por outro lado, algoritmos *lazy* não constroem explicitamente uma hipótese e necessitam lembrar os exemplos de treinamento, pois eles são necessários para predizer o valor y de um novo exemplo. O algoritmo k*NN*, descrito a seguir, é um algoritmo do tipo *lazy*.

4.3 Algoritmo kNN — k-Nearest Neighbor

Uma maneira de predizer o valor y de um novo exemplo consiste em comparar esse exemplo com outros cuja classe é conhecida e atribuir a classe do caso mais próximo. Por exemplo, em um consultório médico, o especialista, a partir do conjunto de sintomas que descrevem o estado de saúde do paciente pode procurar, em fichas médicas de pacientes já diagnosticados, conjuntos de sintomas similares a este, no intuito de auxiliar no diagnóstico de determinada doença.

O algoritmo k-*Nearest Neighbor* — k*NN* — é um algoritmo de aprendizado supervisionado do tipo *lazy*, introduzido por Aha et al. (1991). A ideia geral desse algoritmo consiste em encontrar os k exemplos rotulados mais próximos do exemplo não classificado e, com base no rótulo desses exemplos mais próximos, é tomada a decisão relativa à classe do exemplo não rotulado. Os algoritmos da família k*NN* requerem pouco esforço durante a etapa de treinamento. Em contrapartida, o custo computacional para rotular um novo exemplo é relativamente alto, pois, no pior dos casos, esse exemplo deverá ser comparado com todos os exemplos contidos no conjunto de exemplos de treinamento.

Na Figura 4.1 é ilustrada essa ideia para um problema de classificação, com um conjunto de exemplos de treinamento descrito por dois atributos, no qual, exemplos com rótulo positivo (+) referem-se a pacientes doentes e exemplos com rótulo negativo (−) a não doentes. Considerando o algoritmo k*NN* para classificação, com $k = 1$, o novo exemplo E_i seria classificado de acordo com o único vizinho mais próximo, que é da classe positiva (+).

Três parâmetros importantes devem ser determinados para a execução de k*NN*:

1. quais exemplos rotulados, *i.e.*, exemplos de treinamento, devem ser lembrados;

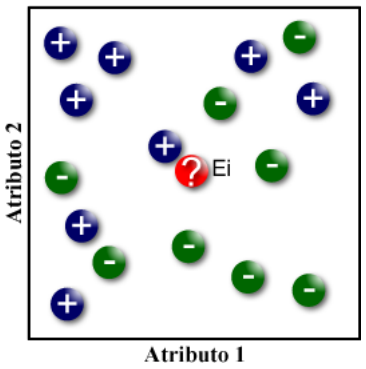


Figura 4.1: Exemplo de classificação do método k-*Nearest Neighbor*

2. qual a medida que quantifica a similaridade entre o exemplo não classificado e os exemplos de treinamento; e

3. quantos/quais vizinhos mais próximos devem ser considerados.

A seguir são descritos esses três parâmetros.

4.3.1 Conjunto de Exemplos de Treinamento

A quantidade de exemplos de treinamento a serem lembrados tem influência direta no tempo de busca pelos exemplos mais próximos do exemplo a ser classificado, pois é necessário comparar esse exemplo com todos os armazenados (Alpaydin, 2004). Dependendo do domínio, essa quantidade de exemplos pode ser muito grande e tornar o processo de classificação lento, até o ponto de não atender ao requisito de tempo máximo de resposta para determinado problema.

Ao invés de utilizar muitos exemplos de treinamento, o ideal é armazenar somente os exemplos mais representativos de cada classe, resumindo a informação mais importante em um conjunto menor de exemplos. Em (Aha et al., 1991) são descritas algumas estratégias para selecionar os exemplos mais representativos de cada classe, a partir do conjunto de exemplos rotulados disponíveis, contribuindo para a redução do custo para classificar novos exemplos e do espaço ocupado em memória pelos exemplos de treinamento.

4.3.2 Medida de Similaridade entre Exemplos

Outra questão relevante está relacionada com a definição da medida utilizada para determinar o grau de similaridade entre o exemplo a ser rotulado e os exemplos no conjunto de treinamento. Diversas medidas têm sido propostas, entre as quais estão as medidas de distância e de correlação. Uma revisão

de vários tipos de medidas de similaridade pode ser encontrada em (Jain e Dubes, 1988) e (Everitt, 1993).

Quando o conjunto de dados é descrito por atributos numéricos, as medidas de distância podem ser devidamente aplicadas para o cálculo da similaridade entre os exemplos, tal que menor distância corresponde a maior similaridade. Diversos índices de proximidade têm sido propostos para o cálculo da similaridade entre pares de exemplos (E_i, E_j), os quais devem satisfazer as seguintes condições:

i. $dist(E_i, E_j) \geq 0, \forall(i, j)$ (positividade)

ii. $dist(E_i, E_j) = 0$ se e somente se $E_i = E_j$ (identidade)

iii. $dist(E_i, E_j) = dist(E_j, E_i)$ (simetria)

Para que um índice de proximidade seja considerado uma métrica, este deve satisfazer, além das três propriedades anteriores, a propriedade de desigualdade triangular:

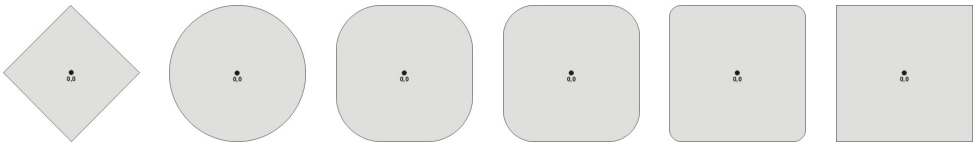
iv. $dist(E_i, E_j) \leq dist(E_i, E_q) + dist(E_q, E_j), \forall(i, j, q)$

Considerando os atributos dos exemplos como dimensões de um espaço multi-dimensional, a descrição de cada exemplo corresponde a um ponto nesse espaço, *i.e.*, $E_i = (x_{i1}, x_{i2}, \ldots, x_{in})$. Minkowsky estabeleceu uma maneira genérica para calcular a distância entre dois pontos no espaço n-dimensional $\Re^n$ de acordo com o valor do parâmetro d, o qual determina a medida utilizada — Equação 4.1.

$$dist(E_i, E_j) = \left(\sum_{l=1}^{n} |x_{il} - x_{jl}|^d \right)^{\frac{1}{d}} \tag{4.1}$$

Conforme o aumento do valor de d, a figura geométrica formada pelos pontos equidistantes do ponto central aproximam-se de um quadrado. Para $d \to \infty$, esses pontos formam exatamente um quadrado, conforme ilustrado na Figura 4.2. Quando $d = 1$, a medida Minkowsky é conhecida como distância de Manhattan/*city-block*, e quando $d = 2$, ela define a distância Euclidiana.

Aggarwal et al. (2001) discutem a adequabilidade da medida de Minkowsky para conjuntos de dados com grandes dimensões. Além disso, sugerem que valores $d = 1$ ou 2 são mais relevantes que valores $d \geq 3$. Neste trabalho, são apresentadas também algumas evidências teóricas e empíricas de que os resultados dessas medidas podem degradar rapidamente ao serem aplicadas



(a) (b) (c) (d) (e) (f)

Figura 4.2: Efeito da variação de d na métrica de Minkowsky. (a) $d = 1$, (b) $d = 2$, (c) $d = 3$, (d) $d = 4$, (e) $d = 20$ e (f) $d \to \infty$

sobre conjuntos de dados com alta dimensão e altos valores de d. A partir desse estudo, observou-se que valores de d menores produzem resultados melhores.

A similaridade entre exemplos pode também ser determinada pelo coeficiente de correlação, o qual é também medido utilizando os valores dos diversos atributos considerando o padrão desses valores e não a magnitude. As medidas de similaridade baseadas em correlação são consideradas semi-métricas, pois não satisfazem a propriedade de desigualdade triangular.

Neste trabalho será usada a distância Euclideana.

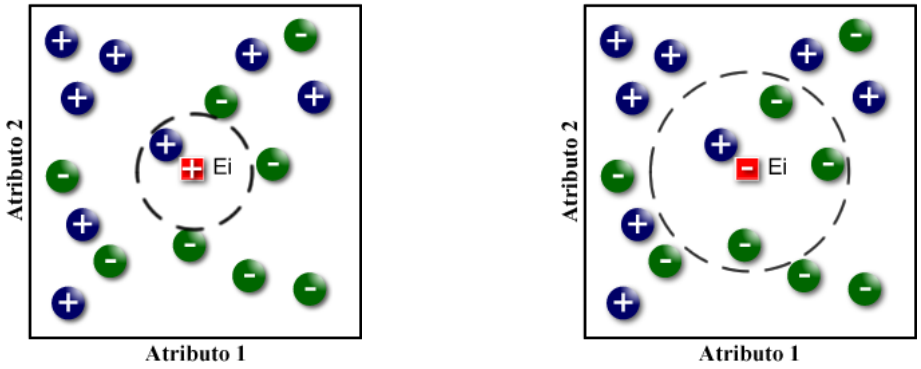
Uma outra questão importante está relacionada aos atributos que representam os exemplos. As medidas de similaridade apresentadas consideram todos os atributos para calcular a distância entre os exemplos. Entretanto, alguns atributos podem não ser representativos e é importante que esses atributos não participem no cálculo de similaridade já que podem influenciar desfavoravelmente o desempenho da classificação. Assim, é importante selecionar inicialmente os atributos relevantes que descrevem os exemplos, o que pode ser realizado utilizando algum algoritmo de seleção de atributos (Liu e Motoda, 2007; Lee, 2005)

4.3.3 Cardinalidade do Conjunto de Vizinhos mais Próximos

O algoritmo k*NN* classifica exemplos considerando a classe dos k vizinhos mais próximos. Se $k = 1$, então o exemplo é classificado com a mesma classe do exemplo mais próximo segundo a medida de similaridade utilizada. Se $k > 1$, então são consideradas as classes dos k exemplos mais próximos para realizar a classificação. Nesse caso, a abordagem mais simples consiste em atribuir ao exemplo a classe majoritária (predominante) dos k exemplos mais próximos.

Na Figura 4.3 são ilustrados ambos os casos na classificação do exemplo E_i, utilizando um conjunto de exemplos positivos (+) e negativos (−) descritos por dois atributos. No primeiro caso, 1*NN*, o exemplo E_i será classificado como positivo. Já no segundo caso, 4*NN*, a maioria dos quatro exemplos mais

próximos é negativo e E_i será classificado como negativo.



(a) Exemplo de $1NN$ (b) Exemplo de $4NN$

Figura 4.3: Exemplos de aplicação do algoritmo kNN, com $k = 1$ e $k = 4$

Como pode ser observado, o número de vizinhos mais próximos a ser considerado na classificação de novos exemplos influencia fortemente a classificação. É importante ressaltar que não existe um único valor de k que seja apropriado para todos os problemas, de modo que esse valor deve ser avaliado para cada problema em particular. No caso de simplesmente utilizar a classe majoritária dos k exemplos mais próximos para classificar exemplos, valores ímpares de k são mais apropriados a fim de não ter situações de empate. Outras abordagens alternativas consistem na atribuição de pesos a cada um dos k vizinhos mais próximos de acordo com a medida de similaridade considerada, *i.e.*, os k vizinhos mais próximos são ordenados em ordem crescente de similaridade com o exemplo a ser classificado, tal que, para a determinação da classificação, a classe dos exemplos de maior similaridade têm peso maior que a classe dos exemplos de menos similaridade.

4.4 Algoritmo kNN-TSP — k-Nearest Neighbor - Time Series Prediction

Como mencionado no Capítulo 3, o problema de previsão é um dos de maior interesse nos estudos relacionados a séries temporais. Esse problema consiste em estimar o valor de x_{t+1} dada uma série temporal[1] $X = (x_1, x_2, \ldots, x_n)$, utilizando os valores anteriores a $t + 1$, isto é, $x_t, x_{t-1}, x_{t-2}, \ldots, x_{t-m+1}$, onde m corresponde ao número de valores prévios da série X utilizados para realizar a previsão.

[1] A notação X é equivalente à notação prévia $Z(t)$ — Página 15.

Na seção anterior foi apresentado o algoritmo de aprendizado supervisionado k-*Nearest Neighbor*, para classificação de exemplos. Na previsão de séries temporais é de interesse prever valores que irão ocorrer no futuro[2]. Para isso, é necessário adaptar a abordagem convencional do algoritmo k*NN* para a previsão de dados contínuos (Kulesh et al., 2008). Neste trabalho, essa adaptação deve também considerar que os dados provêm de séries temporais.

A ideia consiste em, considerando os últimos w registros ocorridos, encontrar as subsequências de tamanho w que apresentaram comportamentos similares no passado. Com base nas informações dessas subsequências, é realizado o cálculo do valor futuro $\hat{x}_{t+1}$, que é uma aproximação do verdadeiro, mas desconhecido, valor x_{t+1} da série temporal. Neste trabalho, o método que implementa essa ideia é denominado k-*Nearest Neighbor - Time Series Prediction* — k*NN-TSP*.

Na Figura 4.4 é apresentado um exemplo da aplicação do algoritmo k*NN-TSP*, com $k = 3$. No gráfico, a linha cinza representa os valores registrados; a linha vermelha, a sequência dos quatro últimos valores ocorridos, isto é, os registros de um período de um ano; e, a linha verde, as sequências similares ao período em questão, encontradas pelo algoritmo. Utilizando os valores sucessores de cada uma das sequências mais similares foi realizada a previsão do valor do quarto trimestre de 2002. Nessa figura, também é ilustrado o valor previsto por uma função de previsão, representado por um círculo vermelho.

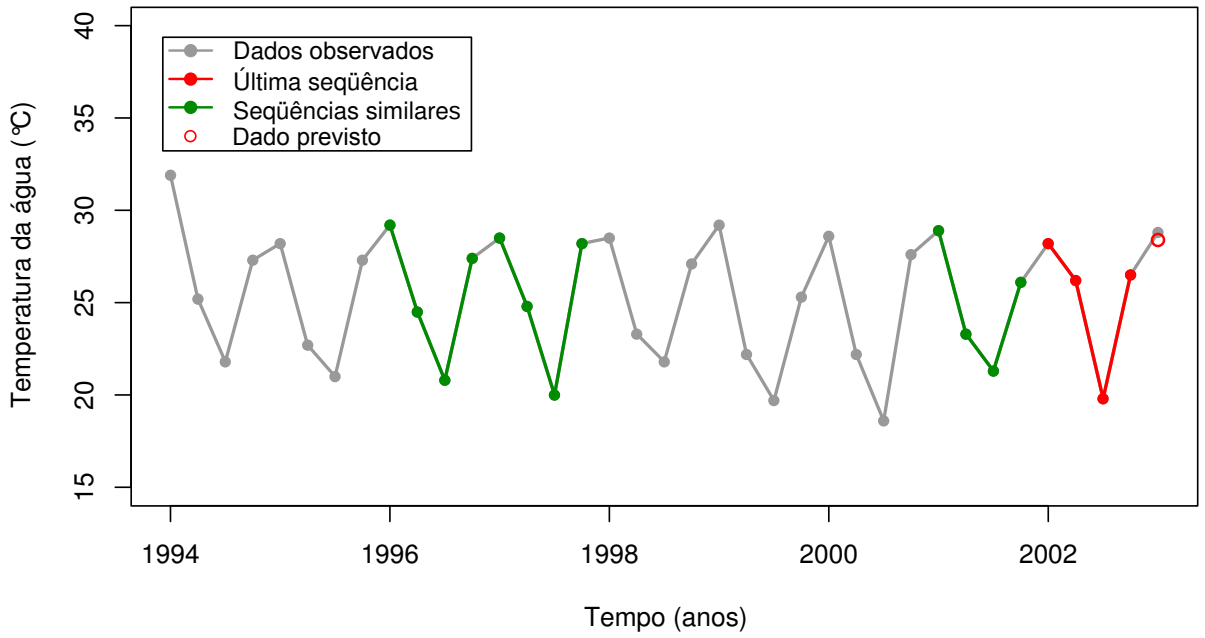


Figura 4.4: Exemplo da aplicação do algoritmo k*NN-TSP*

[2]Nesse caso, os valores são contínuos (regressão), e não discretos como no caso de classificação.

Como pode ser visualizado na Figura 4.4, o valor de w, *i.e.*, o tamanho da janela de procura das sequências similares, no caso $w = 4$, é muito relevante, pois de acordo com o valor dessa variável o conjunto de subsequências mais similares encontradas na série temporal pode ser diferente, sendo, consequentemente, diferente o valor calculado pela função de previsão. Além do valor de w, e de modo análogo ao algoritmo k*NN*, para a execução do algoritmo k*NN-TSP* é necessário determinar: o conjunto de exemplos de treinamento; a medida de similaridade; a cardinalidade do conjunto de séries similares; e a função de previsão, ilustrados na Figura 4.5 e detalhados a seguir.

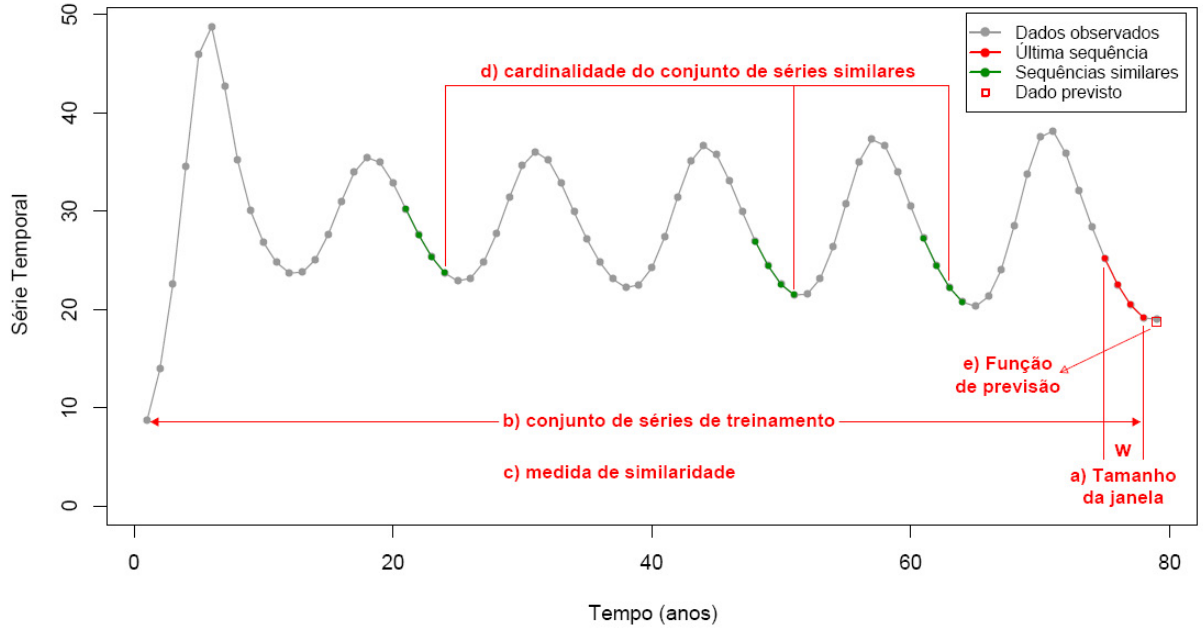


Figura 4.5: Parâmetros do algoritmo k*NN-TSP*

(a) tamanho w da janela para extrair as subsequências: corresponde ao tamanho das subsequências extraídas da série temporal.

(b) conjunto de exemplos (séries) de treinamento: refere-se às subsequências da série temporal que serão consideradas para constituir o conjunto de séries de treinamento. Para diminuir o tamanho desse conjunto, podem ser armazenadas apenas as últimas subsequências de acordo com algum período de tempo, ou as mais representativas de acordo com algum critério de seleção de subsequências.

(c) medida de similaridade: utilizada para quantificar a similaridade entre exemplos — Seção 4.3.2. No k*NN-TSP*, cada subsequência é representada no espaço w-dimensional. Assim, cada ponto da subsequência é considerado como o valor de cada atributo no cálculo de similaridade.

(d) cardinalidade do conjunto de séries similares: define o número de séries mais próximas, $k \geq 1$, que serão consideradas para a previsão do valor futuro. Geralmente, diversos valores de k são utilizados com o objetivo de encontrar o mais adequado para um determinado problema.

(e) função de previsão: determina como serão utilizados os valores das subsequências mais próximas para estimar o valor futuro. Essa estimativa pode ser realizada, por exemplo, de acordo com o valor futuro das subsequências mais similares consideradas. Assim, diferentes critérios podem ser utilizados para definir funções de previsão, no intuito de explorar características específicas dos dados nos requisitos de cada domínio.

Assim, o algoritmo de previsão *kNN-TSP* deve levar em consideração as seguintes três fases:

Fase 1 — Preparação do conjunto de séries de treinamento;

Fase 2 — Obtenção dos vizinhos mais próximos;

Fase 3 — Cálculo do valor futuro,

as quais são descritas a seguir.

4.4.1 Fase 1 — Preparação do Conjunto de Séries de Treinamento

No caso em que a base de dados esteja constituída de ensaios representados por um conjunto de séries temporais, que descrevem o comportamento de determinado fenômeno, o conjunto de séries de treinamento consiste simplesmente das séries temporais correspondentes aos ensaios. Porém, grande parte dos problemas de monitoramento fornecem uma única série temporal. Exemplos desses problemas são o monitoramento do estado de saúde de pacientes, da estabilidade de estruturas civis como barragens e da qualidade de águas em lagos e reservatórios. Nesses casos, informações são coletadas continuamente ao longo do tempo em uma única série temporal, a qual necessita ser desmembrada utilizando algum critério, com o objetivo de construir o conjunto de séries de treinamento.

Para construir esse conjunto pode ser utilizada uma janela de tamanho w, de modo a extrair da série temporal subsequências de tamanho fixo, que representem parte do comportamento da série temporal. Com isso, é possível analisar localmente o fenômeno que está sendo avaliado, no intuito de realizar previsões locais, ou de curto prazo. Assim, a preparação dos dados consiste na

definição do tamanho w dessa janela e na extração de todas as subsequências de tamanho w.

Seja a série temporal $X = (x_1, x_2, \ldots, x_n)$ de tamanho n. Considere que o par ordenado $(\boldsymbol{x_n}, x_{n+1})$ define a sequência de referência, em que $\boldsymbol{x_n}$ corresponde aos últimos w valores de X, $(x_{n-(w-1)}, x_{n-(w-2)}, \ldots, x_n)$, e x_{n+1} ao valor a ser previsto. Utilizando essa notação, cada elemento do conjunto de séries de treinamento $S = \{S_w, S_{w+1}, S_{w+2}, \ldots, S_{n-1}\}$ é definido pela Equação 4.2.

$$
\begin{aligned}
S_w &= (\boldsymbol{x_w}, x_{w+1}) = ((x_1, x_2, \ldots, x_w), x_{w+1}) \\
S_{w+1} &= (\boldsymbol{x_{w+1}}, x_{w+2}) = ((x_2, x_3, \ldots, x_{w+1}), x_{w+2}) \\
&\vdots \\
S_{w+i} &= (\boldsymbol{x_{w+i}}, x_{w+i+1}) = ((x_{i+1}, x_{i+2}, \ldots, x_{w+i}), x_{w+i+1}) \\
&\vdots \\
S_{w+j} &= (\boldsymbol{x_{w+j}}, x_{w+j+1}) = ((x_{j+1}, x_{j+2}, \ldots, x_{w+j}), x_{w+j+1}) \\
&\vdots \\
S_{n-1} &= (\boldsymbol{x_{n-1}}, x_n) = ((x_{n-w}, x_{n-(w-1)}, \ldots, x_{n-1}), x_n)
\end{aligned}
\tag{4.2}
$$

ou seja, cada elemento de $S_i \in S$ é um par ordenado $(\boldsymbol{x_i}, x_{i+1})$, onde a primeira componente corresponde à i-ésima sequência extraída de X, dada por $(x_{i-(w-1)}, x_{i-(w-2)}, \ldots, x_i)$, e a segunda componente refere-se à classe, *i.e.*, ao valor futuro de cada subsequência de treinamento.

Como mencionado, o tamanho w da janela influencia diretamente no formato dos padrões encontrados e, consequentemente, no valor da função de previsão. Ainda, a utilização de todas as subsequências extraídas da série para compor o conjunto de treinamento pode tornar o processo de previsão lento e ineficiente em determinados domínios. Com base nessas considerações, a seguir são descritos métodos para identificar o tamanho w da janela e são apresentados alguns critérios para a seleção de subsequências importantes para compor o conjunto de séries de treinamento.

Identificação do Tamanho da Janela: o tamanho da janela depende do domínio, pois o próprio comportamento do fenômeno avaliado, assim como a frequência de coleta das informações ao longo do tempo, são fatores que podem variar para cada domínio e cada problema. Porém, algumas abordagens permitem identificar aproximações para determinar o tamanho da janela[3], entre essas:

[3] O tamanho w da janela pode ser entendido como a dimensão de imersão da série temporal, a qual possibilita a modelagem da trajetória da dinâmica que gera o comportamento dessas séries (Chun-Hua e Xin-Bao, 2004).

- pela análise visual do especialista do domínio;
- pelo cálculo da dimensão de correlação;
- pelo método de Falsos Vizinhos Próximos (*FNN — False Nearest Neighbor*).

A primeira abordagem consiste na identificação visual da periodicidade dos dados (Kulesh et al., 2008). Por exemplo, no caso de dados de temperatura da água em lagos, ou outros comportamentos biológicos, esse período pode ser identificado visualmente e, assim, a quantidade de dados de cada período corresponde ao tamanho da janela. A segunda abordagem, cálculo da dimensão de correlação, proposta por Grassberger e Procaccia (1983), permite identificar a dimensão de imersão pela avaliação da auto-similaridade entre os pontos. Para isso, a dimensão é calculada para diversos valores do tamanho da janela até que a relação entre essa correlação e w se torne constante. A última abordagem, *FNN*, proposta por Kennel et al. (1992), consiste em definir w de modo que valores de dimensão superiores a w não influenciem na evolução da trajetória da série. Quando é utilizada uma dimensão baixa, comportamentos de trajetórias diferentes podem estar próximos no espaço w-dimensional, sendo que, à medida que o valor de dimensão aumenta, esses comportamentos deixam de ser próximos, caracterizando-os como falsos vizinhos. Com isso, o objetivo do método *FNN* consiste em aumentar a dimensão até minimizar o número de falsos vizinhos próximos.

Outras abordagens têm sido propostas, tais como a utilização da decomposição de valores singulares (Broomhead e King, 1986) e a utilização da medida *Dynamic Time Warping* (Mizuhara et al., 2006). De acordo com Chun-Hua e Xin-Bao (2004), o método *FNN* é o mais comumente utilizado por apresentar dimensões adequadas para vários problemas de imersão de séries temporais.

Seleção de Subsequências Importantes: a utilização do método k*NN-TSP* para previsão depende, em termos de tempo e de espaço, do tamanho do conjunto de treinamento. Por esse motivo, é importante que sejam consideradas apenas subsequências que permitem definir estimativas próximas dos valores a serem previstos. Essa seleção pode ser realizada, por exemplo, de acordo com a representatividade de alguma medida de informação, ou pela definição de uma distância temporal que permita identificar as mais recentes e desconsiderar as mais antigas.

4.4.2 Fase 2 — Obtenção dos Vizinhos mais Próximos

Uma vez definido o conjunto de séries que compõem o conjunto de treinamento S do algoritmo, é iniciada a fase de obtenção dos vizinhos mais próximos. O par $(\boldsymbol{x_n}, x_{n+1})$ contém os últimos w dados coletados, com o objetivo de prever o valor de x_{n+1} usando os vizinhos mais próximos. Como mencionado, para encontrar os vizinhos mais próximos devem ser definidos: a medida de similaridade a ser utilizada e o número k de vizinhos mais próximos a ser considerado. Pela relevância dessas duas questões, a seguir são apresentadas algumas medidas de similaridade para séries temporais e critérios para selecionar os vizinhos mais próximos.

Medida de Similaridade para Séries Temporais: no contexto de séries temporais é relevante a utilização de medidas de similaridade que permitam não apenas quantificar a distância entre os pontos que compõem duas subsequências, mas também considerar o formato, ou comportamento, das subsequências como um aspecto importante para decidir se duas sequências são ou não similares. Para isso, inicialmente, as subsequências podem ser pré-processadas, por exemplo realizando uma normalização dos valores para colocar ambas as subsequências no mesmo nível e, desse modo, permitir uma melhor aproximação do grau de similaridade em relação ao comportamento.

Na Figura 4.6 é ilustrado um exemplo de normalização de duas séries temporais Ei e Ej, as quais se encontram em níveis diferentes. Essa normalização consiste em subtrair de cada subsequência o valor médio da respectiva subsequência. Em (Kulesh et al., 2008) é apresentada uma abordagem que considera não somente a normalização de nível, mas também a normalização de variância das subsequências. Outras abordagens podem ser encontradas em (Illa et al., 2004).

São muitas as medidas de similaridade propostas na literatura. Fabris et al. (2008) propõem um algoritmo para realizar a combinação de várias medidas de similaridade, às quais são atribuídos pesos para valorizar aquelas mais apropriadas. A avaliação experimental dessa proposta mostrou bons resultados. Ainda assim, de acordo com Keogh e Kasetty (2002), a distância Euclidiana apresenta-se como a mais frequentemente utilizada, devido ao fato de ser uma medida intuitiva e do baixo custo computacional para o cálculo da similaridade. Como mencionado, essa medida determina a distância entre dois pontos no espaço $\Re^w$, no qual w corresponde ao tamanho da sequência considerada.

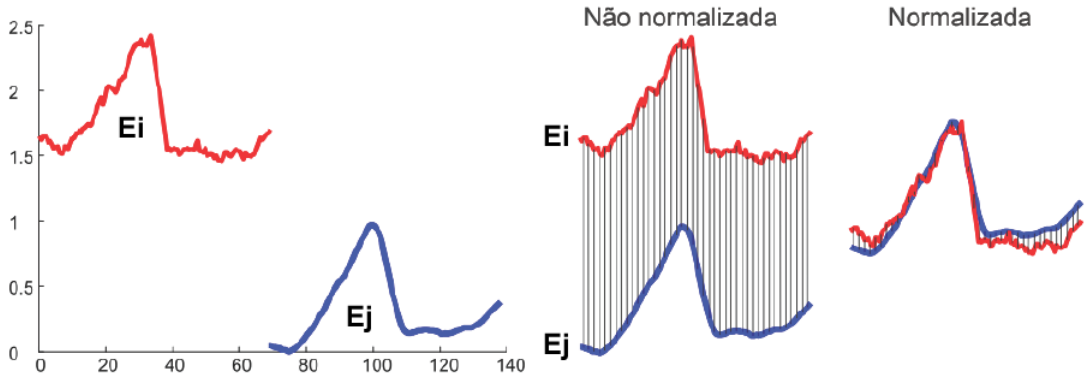


Figura 4.6: Exemplo de normalização de séries temporais (Keogh e Kasetty, 2002)

No contexto de séries temporais, existem situações em que as subsequências a serem comparadas possuem tamanhos diferentes. Nesse caso, as medidas convencionais de similaridade não podem ser aplicadas. Para esses casos têm sido desenvolvidas medidas específicas, como: *Longest Common Subsequence* — LCSS, que utiliza o tamanho da maior subsequência em comum como o critério de similaridade (Hirschberg, 1977); *Edit Distance on Real Sequence* — EDRS, que utiliza como critério de medida o número mínimo de operações de inserção, remoção e substituição para transformar uma sequência em outra (Chen e Ng, 2004); e *Dynamic Time Warping*, que consiste em calcular a distância para vários alinhamentos, por meio de alguma medida convencional de similaridade, com a finalidade de encontrar o melhor alinhamento entre as subsequências (Chu et al., 2002).

Seleção dos Vizinhos mais Próximos: a ideia de escolher os melhores vizinhos mais próximos é intuitiva para permitir a previsão de valores mais precisos (Kulesh et al., 2008). A seleção desses vizinhos mais próximos abrange duas questões importantes:

1. a cardinalidade do conjunto de vizinhos mais próximos;
2. os critérios para selecionar esses vizinhos.

Na primeira questão deve ser definido o número k de séries mais próximas a serem consideradas para estimar o valor futuro. O valor de k pode ser definido *a priori*, ou podem ser experimentados diversos valores de k para encontrar o que apresenta o melhor desempenho para a tarefa de previsão. Outra abordagem consiste em definir um limiar de similaridade l, tal que, todas as subsequências com valor de similaridade maior que

l são consideradas para compor o conjunto de vizinhos mais próximos. Ou seja, neste caso não é definido um valor único para k.

Em relação à segunda questão, no algoritmo *kNN-TSP* os vizinhos mais próximos podem ser selecionados não somente de acordo com a medida de similaridade, mas também considerando o tempo, já que no contexto de dados temporais, o tempo pode ser um fator importante no processo de seleção, no sentido de que os últimos comportamentos similares ocorridos têm, em geral, uma influência maior no valor futuro, em relação àqueles que ocorreram em períodos de tempo mais distantes. Desse modo, podem ser consideradas, basicamente, duas abordagens: utilizar apenas a medida de similaridade, ou combinar a medida de similaridade e a distância temporal. Neste trabalho são utilizadas ambas as abordagens, sendo que, para essa última propomos um procedimento para combinar a similaridade e a distância temporal. As duas abordagens de seleção são apresentadas em maiores detalhes a seguir.

Seleção pela medida de similaridade: as subsequências de maior similaridade são selecionadas para constituir o conjunto de sequências mais próximas. De modo análogo ao algoritmo *kNN*, cada ponto da subsequência é considerado um valor de atributo. Nesse sentido, as medidas apresentadas na Seção 4.3.2 podem ser utilizadas para quantificar a distância entre os pontos. Após, as k subsequências de menor distância, ou todas as subsequências a uma distância menor que um limiar l, podem ser selecionadas para constituir o conjunto de vizinhos mais próximos.

Na Figura 4.7 é ilustrado um exemplo da aplicação do critério de seleção de vizinhos mais próximos utilizando a distância Euclidiana. A série temporal original é registrada pelos pontos delineados em cinza, a última sequência pelos pontos delineados em vermelho e as sequências similares pelos pontos delineados em verde. Os pontos de cor preta apresentam a distância Euclidiana entre as sequências que compõem o conjunto de treinamento e a última sequência de dados registrada, ou seja, quanto mais próximos do valor 0 mais similar é a sequência. As sequências similares são selecionadas conforme a menor distância.

Seleção pela similaridade e a distância temporal: as subsequências de maior similaridade e menor distância temporal são consideradas para constituir o conjunto de sequências mais próximas. Neste trabalho, este procedimento é utilizado e é proposto um método muito simples

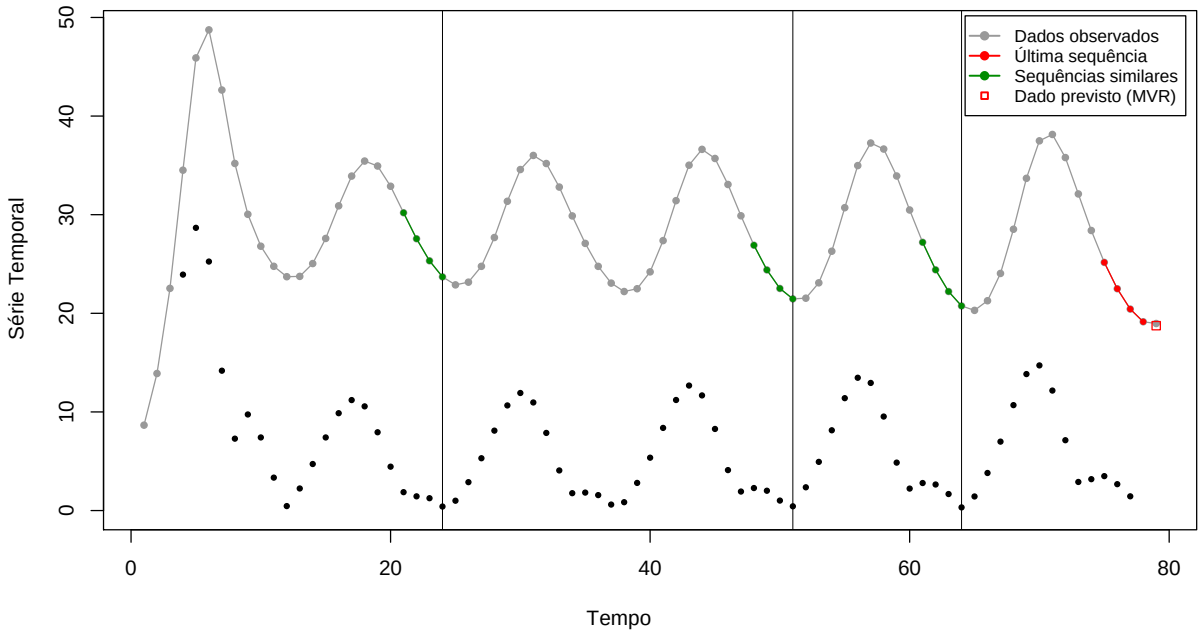


Figura 4.7: Exemplo do critério de seleção dos $k = 3$ vizinhos mais próximos por similaridade

para realizar esse tipo de seleção. O método considera uma reta L definida inicialmente pelo par de pontos $((t_0, 0), (t_n, 0))$. As coordenadas y desses pontos, inicialmente 0, representam o limiar de distância entre as sequências. Caso o número de pontos que pertencem a L for igual a k, eles identificam as k sequências mais próximas. Se for maior que k, então esses pontos são organizados em ordem decrescente de tempo e os k primeiros são selecionados, caso contrário, o ângulo da reta L, inicialmente zero, é incrementado por uma constante β_g, definida pelo usuário, e o número de pontos, tanto abaixo de L quanto os que pertencem a L, é calculado. Se esse número for igual ou maior que k, as k sequências mais próximas são identificadas da mesma maneira que no inicio do processo. Caso contrário, na próxima iteração o ângulo β é novamente incrementado pela constante β_g e o processo é repetido até encontrar os k vizinhos mais próximos.

Na Figura 4.8 é apresentado um exemplo utilizando a mesma série da Figura 4.7 para o critério de similaridade. Como pode ser observado, neste caso é dada prioridade a uma sequência similar mas que se apresenta mais próxima ao valor a ser previsto, o qual, pode fornecer melhores informações relacionadas ao valor a ser previsto.

Qualquer que seja a abordagem utilizada para selecionar os vizinhos mais próximos da última sequência registrada, o resultado da execução da segunda fase do algoritmo *kNN-TSP* consiste no subconjunto $S' = S'_1, S'_2, ..., S'_k \subset S$, tal

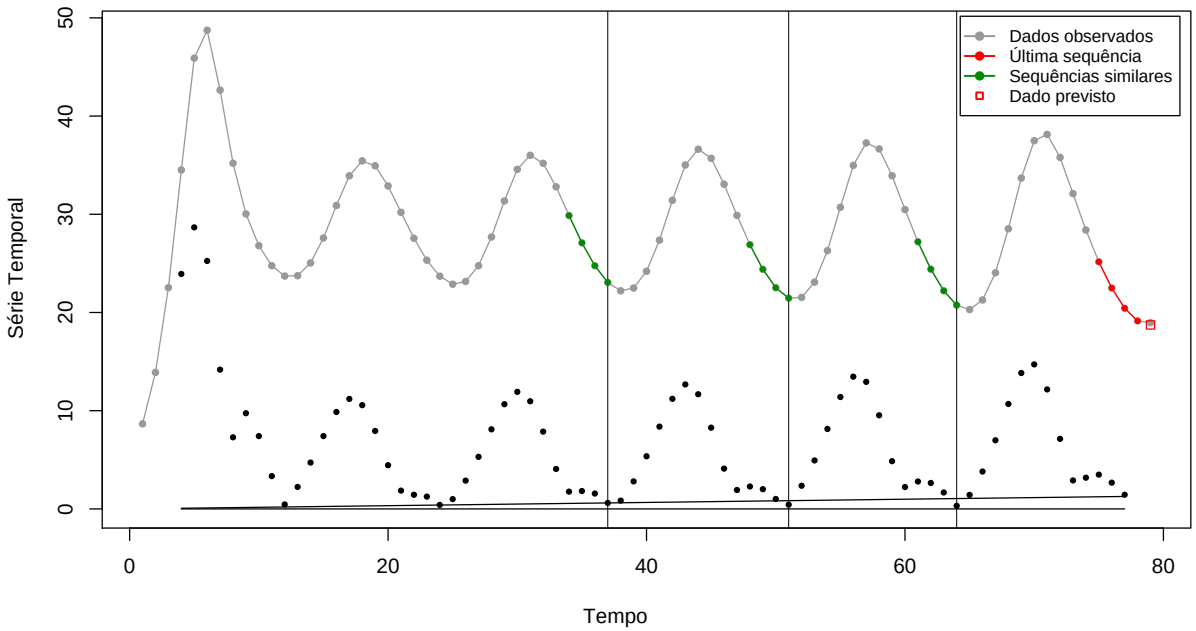


Figura 4.8: Exemplo do critério de seleção dos $k = 3$ vizinhos mais próximos por similaridade e distância temporal

que, a cardinalidade de S' é k e a série S'_i é um par ordenado $(\boldsymbol{s_{i,w}}, s_{i,w+1})$ que corresponde ao exemplo que contém a i-ésima sequência mais similar à última sequência registrada, *i.e.*, a sequência de referência $(\boldsymbol{x_n}, x_{n+1})$.

Por exemplo, considere o conjunto de séries de treinamento definido na Seção 5.2.1, página 37, utilizada para estimar o valor x_{n+1}, para a qual foram encontradas os seguintes três vizinhos mais próximos $\{S_{w+1}, S_i, S_j\}$, então $S' = \{S'_1, S'_2, S'_3\}$, tal que

$$\begin{aligned} S'_1 &= ((s'_{1,1}, s'_{1,2}, \ldots, s'_{1,w}), s'_{1,w+1}) = S_{w+1} = ((x_2, x_3, \ldots, x_{w+1}), x_{w+2}) \\ S'_2 &= ((s'_{2,1}, s'_{2,2}, \ldots, s'_{2,w}), s'_{2,w+1}) = S_{w+i} = ((x_{i+1}, x_{i+2}, \ldots, x_{w+i}), x_{w+i+1}) \\ S'_3 &= ((s'_{3,1}, s'_{3,2}, \ldots, s'_{3,w}), s'_{3,w+1}) = S_{w+j} = ((x_{j+1}, x_{j+2}, \ldots, x_{w+j}), x_{w+j+1}) \end{aligned} \tag{4.3}$$

4.4.3 Fase 3 — Cálculo do Valor Futuro

Nesta fase, a função de previsão $f(S')$ é responsável pela estimativa do valor futuro de x_{n+1} da sequência de referência. As funções de previsão comumente utilizadas na literatura aproximam esse valor pela média local ou pela média ponderada dos valores da classe, dado por $s'_{i,w+1}$ de cada sequência $S'_i \in S'$ (Karunasinghe e Liong, 2006). A função de previsão que utiliza a média local para estimar o valor de x_{n+1}, *i.e.*, $\hat{x}_{n+1}$, é definida pela equação 4.4.

$$f_{MV}(S') = \frac{\sum_{i=1}^{k} s'_{i,w+1}}{k} = \hat{x}_{n+1} \tag{4.4}$$

Quanto à função de previsão que utiliza a média ponderada dos $s'_{i,w+1}$, diversos critérios são propostos na literatura para definir os pesos dos $s'_{i,w+1}$,

os quais podem ser combinados linearmente ou utilizando, por exemplo, uma função exponencial (Solomatine et al., 2006).

Neste trabalho, propomos a função de previsão que denominamos Média de Valores Relativos — MVR — f_{MVR}, a qual aproxima o valor de x_{n+1} pelo valor de x_n mais a média local da diferença dos valores da classe, $s'_{i,w+1}$, e o valor $s'_{i,w}$ de cada sequência $S'_i \in S'$. A função é definida pela Equação 4.5.

$$f_{MVR}(S') = x_n + \frac{\sum_{i=1}^{k} \Delta s'_{i,w+1}}{k} = \hat{x}_{n+1} \tag{4.5}$$

onde $\Delta s'_{i,w+1} = s'_{i,w+1} - s'_{i,w}$.

Uma das vantagens da função f_{MVR} sobre a função f_{MV} é que ela permite prever valores futuros a partir de padrões encontrados em níveis diferentes. Ambas funções são ilustradas na Figura 4.9, para $k = 2$, *i.e.*, $S' = \{S'_1, S'_2\}$.

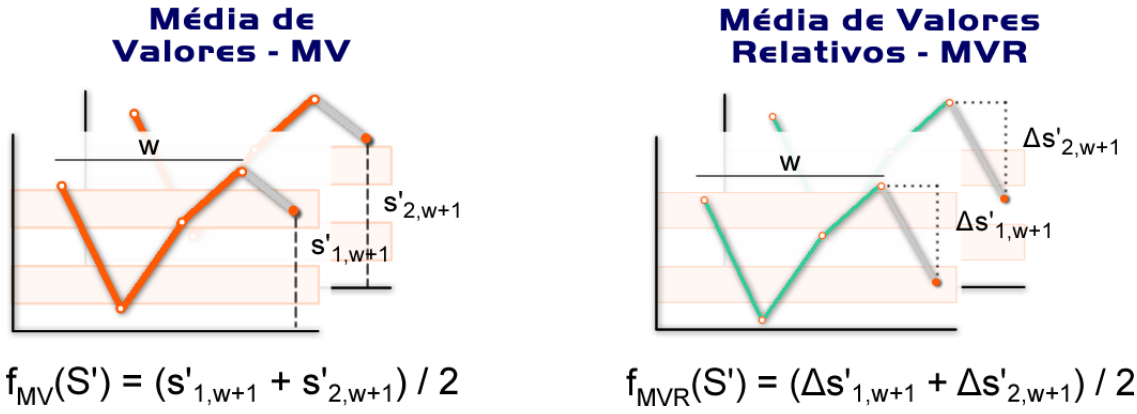


Figura 4.9: Funções de previsão f_{MV} e f_{MVR} para $k = 2$

Na Figura 4.10 é apresentado um exemplo de previsão de um valor futuro utilizando as funções de previsão f_{MV} e f_{MVR} para $k = 2$, considerando o critério de similaridade para a seleção de vizinhos próximos. Os pontos delineados em cinza apresentam a série temporal, os vermelhos os últimos registros ocorridos, os verdes as sequências similares e os valores previstos por f_{MV} e f_{MVR} estão representados, respectivamente, por um círculo e um quadrado vermelho.

4.4.4 Algoritmo kNN-TSP

O Algoritmo 1 descreve o pseudocódigo do algoritmo *kNN-TSP*, onde

- $X = (x_1, x_2, \ldots, x_n)$ é a série de treinamento;
- w é o tamanho da janela para extrair as subsequências;
- M_s é a medida de similaridade;

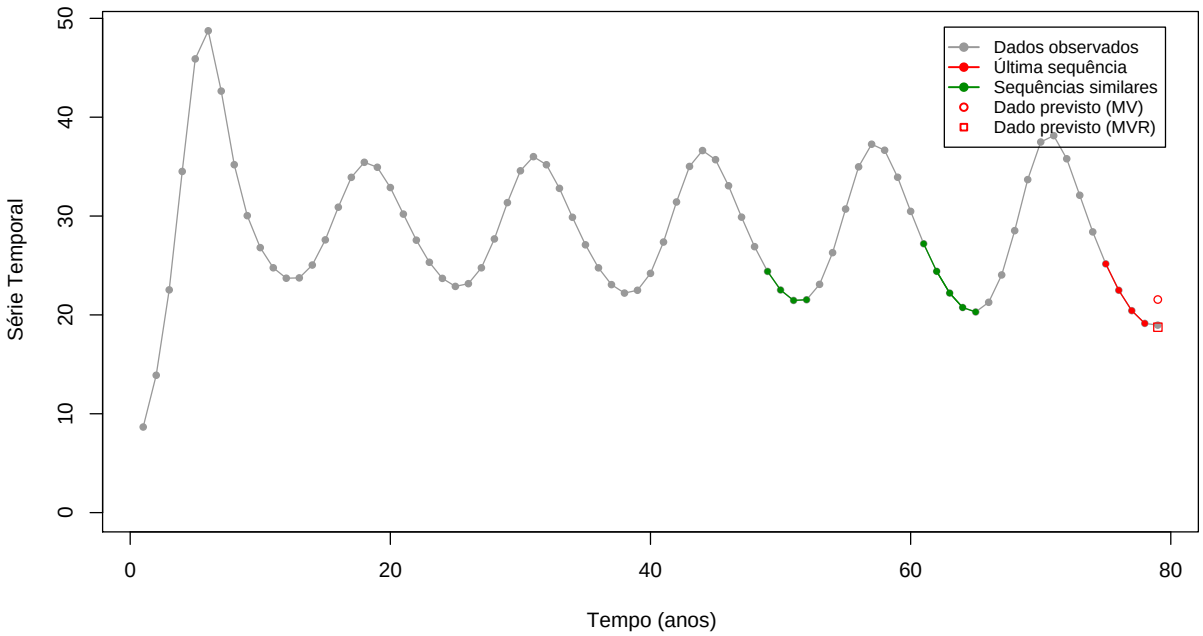


Figura 4.10: Exemplo de previsão utilizando f_{MV} e f_{MVR}

- C_k é o critério de seleção dos vizinhos mais próximos;
- k é o número de vizinhos mais próximos; e
- f é a função de previsão.

Algorithm 1: *kNN-TSP*

Input: X, w, M_s, C_k, k, f
Output: $\hat{x}_{n+1}$
// Construir o conjunto de séries de treinamento S a partir da série X
// e o tamanho de janela w
$S \leftarrow \mathit{séries\_de\_treinamento}(X, w)$;
// Definir a sequência de referência U, para a qual o valor futuro, x_{n+1},
// não é conhecido.
$U \leftarrow (\boldsymbol{x_n}, ?)$;
// Obtenção das k sequências mais próximas a U, contidas em S,
// considerando a medida de similaridade M_s e o critério de seleção
// de vizinhos próximos C_k
$S' \leftarrow \mathit{vizinhos\_próximos}(S, U, M_s, C_k, k)$;
// Cálculo do valor futuro da sequência de referência, utilizando $f(S')$
$\hat{x}_{n+1} \leftarrow f(S')$;
Return $\hat{x}_{n+1}$;

A complexidade da função de preparação do conjunto de séries de treinamento, $\mathit{séries\_de\_treinamento}(X, w)$, é da ordem de $O(n)$, assim como a complexidade da função $\mathit{vizinhos\_próximos}(S, U, M_s, C_k, k)$, responsável por encontrar os k vizinhos mais próximos da sequência de referência U. A função de previsão, $f(S')$, apresenta complexidade $O(k)$. Desse modo, a complexidade do

algoritmo k*NN-TSP* é da ordem de $O(n)$, *i.e.*, é linear em relação ao tamanho da série temporal.

4.5 Considerações Finais

Métodos provenientes da área de aprendizado de máquina têm sido adaptados para a previsão de dados temporais, entre os quais o k*NN*. A adaptação do algoritmo, k*NN-TSP*, permite localizar padrões (subsequências similares) em séries temporais e utilizar esses padrões para prever valores futuros. Neste capítulo foram apresentados alguns conceitos básicos de aprendizado de máquina e o algoritmo k*NN* no contexto de problemas de classificação. Posteriormente, foi apresentado o algoritmo k*NN-TSP* e descritas as fases para a execução do algoritmo. Também são mencionadas as contribuições deste trabalho para esse algoritmo. No próximo capítulo é apresentada a metodologia de avaliação do algoritmo k*NN-TSP* para a realização de experimentos neste trabalho.

CAPÍTULO

5

Metodologia para Avaliação

5.1 Considerações Iniciais

Neste capítulo é apresentada a metodologia que tem como finalidade avaliar o desempenho do algoritmo de previsão *kNN-TSP*, de modo uniformizado, considerando diferentes bases de dados temporais. Nesse sentido, é descrito o processo de construção dos conjuntos de dados de treinamento, assim como as medidas utilizadas para a avaliação, com o objetivo de tornar possível a comparação entre o desempenho de diferentes configurações dos parâmetros do algoritmo. Também é brevemente descrita a ferramenta computacional que implementa a metodologia.

5.2 Metodologia

A metodologia para avaliação proposta tem como objetivo avaliar a qualidade dos métodos de previsão apresentados na Seção 4.4, e consiste de três fases:

Fase 1 — Pré-processamento de séries temporais;

Fase 2 — Configuração de experimentos e previsão; e

Fase 3 — Avaliação de resultados e pós-processamento.

A Fase 1 é responsável pela representação das séries temporais no formato adequado para serem utilizadas por métodos de previsão. A Fase 2

consiste na determinação da quantidade de valores a serem previstos. Com essa informação, são determinados os dados a serem utilizados pelo algoritmo *kNN-TSP* para prever cada valor futuro. Na Fase 3 são calculadas as medidas utilizadas para quantificar o desempenho dos algoritmos de previsão. Na Figura 5.1 são ilustradas, de modo simplificado, as três fases da metodologia.

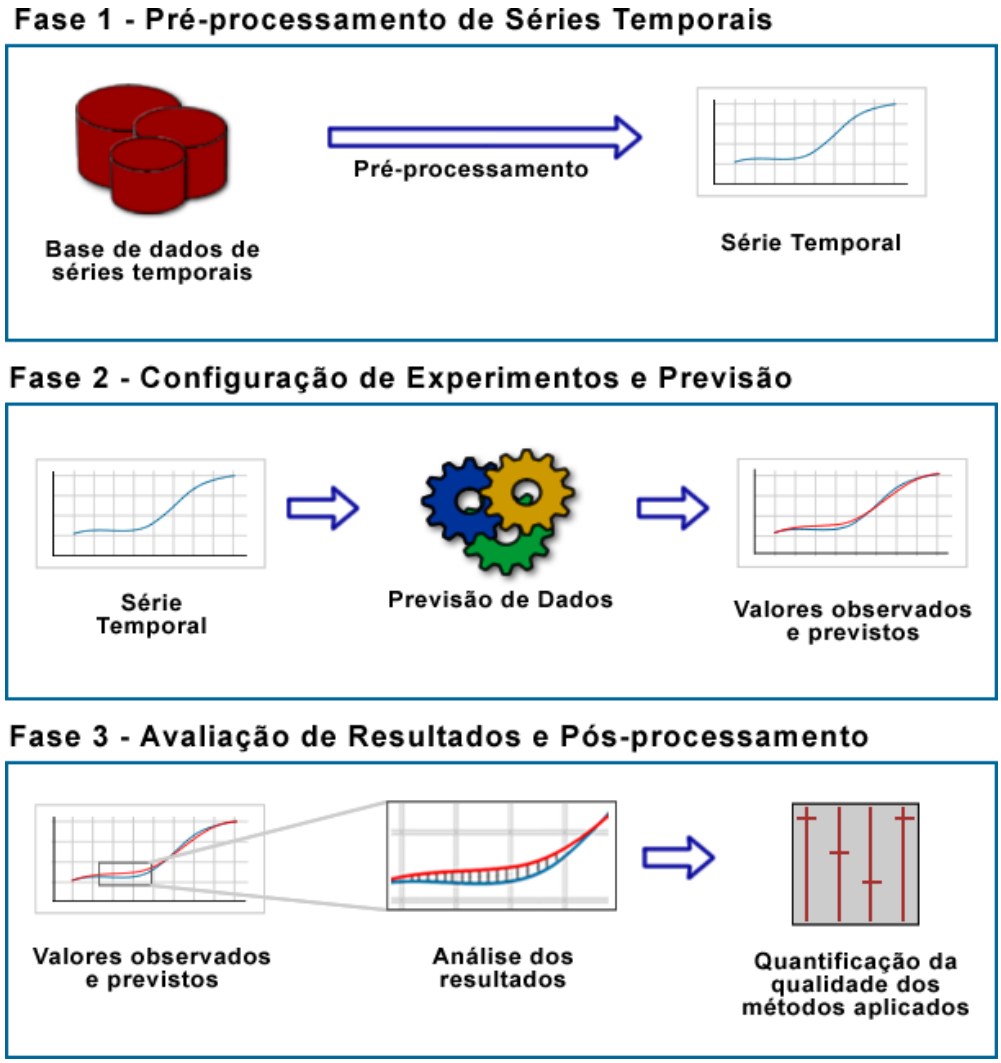


Figura 5.1: Fases da metodologia proposta para a avaliação de métodos de previsão

A seguir, as três fases da metodologia proposta são descritas em maiores detalhes.

5.2.1 Fase 1 — Pré-processamento de Séries Temporais

Como mencionado, esta fase tem como objetivo aplicar métodos de pré-processamento de séries temporais no intuito de preparar essas séries para que possa ser aplicado o algoritmo *kNN-TSP*. Em casos reais, os dados são

coletados e armazenados em diferentes formatos, os quais devem ser transformados para o formato padrão requerido pelo algoritmo de previsão.

Neste trabalho, para realizar essa tarefa foram construídos *scripts*[1], os quais são também responsáveis, entre outros, pela extração de séries temporais de interesse das bases de dados, bem como da identificação de valores faltantes.

5.2.2 Fase 2 — Configuração de Experimentos e Previsão

A abordagem proposta consiste em prever um valor futuro utilizando apenas valores reais no conjunto de treinamento. Ou seja, dada a série temporal $ST = (x_1, x_2, \ldots, x_n)$, o *kNN-TSP* prevê o valor $\hat{x}_{n+1}$ de x_{n+1}. Entretanto, para avaliar o algoritmo é necessário conhecer o valor de x_{n+1} para medir o erro entre esse valor real x_{n+1} e o valor previsto $\hat{x}_{n+1}$. Esse processo deve ser repetido um certo número m de vezes, com o objetivo de obter o erro médio de previsão. Isso pode ser realizado gerando m pares ordenados a partir de ST, como mostra a Equação 5.1.

$$\begin{aligned} ST_1 &= (\boldsymbol{X_1}, x_{n-(m-1)}) = ((x_1, x_2, \ldots, x_{n-m}), x_{n-(m-1)}) \\ ST_2 &= (\boldsymbol{X_2}, x_{n-(m-2)}) = ((x_1, x_2, \ldots, x_{n-m}, x_{n-(m-1)}), x_{n-(m-2)}) \\ &\vdots \\ ST_m &= (\boldsymbol{X_m}, x_n) = ((x_1, x_2, \ldots, x_{n-m}, x_{n-(m-1)}, \ldots, x_{n-1}), x_n) \end{aligned} \tag{5.1}$$

em que ST_i define o i-ésimo conjunto de treinamento e teste, o qual é constituído pelo par ordenado $(\boldsymbol{X_i}, x_{n-(m-i)})$, onde o primeiro termo é a série temporal de treinamento e o segundo termo refere-se à classe.

Assim, o algoritmo *kNN-TSP* deverá ser executado m vezes, utilizando como série de treinamento, em cada iteração i, a série $\boldsymbol{X_i}$, *i.e.*,

---

for $i = 1$ *to* m **do**
  $\hat{x}_{n-(m-1)} \leftarrow$ *kNN-TSP*$(\boldsymbol{X_i}, w, M_s, C_k, k, f)$ — Seção 4.4.4;
end

---

para estimar os m últimos valores da série ST.

[1] *Scripts* são programas de tamanho reduzido normalmente implementados para solucionar problemas específicos.

5.2.3 Fase 3 — Avaliação de Resultados e Pós-processamento

De acordo com cada valor de previsão, é de interesse quantificar a qualidade das estimativas. Por exemplo, a diferença entre o valor previsto e o valor real pode ser entendida como a medida de erro para avaliar métodos de previsão. Porém, quando o número de previsões é grande o suficiente, podem ser extraídas medidas que avaliem de modo mais completo a qualidade da sequência de previsões e, com isso, obter uma estimativa melhor do erro verdadeiro de determinado método de previsão quando aplicado numa série temporal. Em (Hyndman e Koehler, 2006) é apresentada uma discussão a respeito de diversas medidas utilizadas na literatura para a avaliação de métodos de previsão. Neste trabalho são utilizadas as medidas de Erro Médio Absoluto (EMA) e o coeficiente de correlação (r). Considerando $ST = (x_1, x_2, \ldots, x_n)$ a série temporal de valores observados, e $\hat{X} = (\hat{x}_{n-(m-1)}, \hat{x}_{n-(m-2)}, \ldots, \hat{x}_n)$ os m valores estimados utilizando o *kNN-TSP*, como mostrado na Seção 5.2.2, essas medidas são descritas a seguir:

Erro Médio Absoluto (EMA): essa medida permite calcular o erro de previsão a partir das diferenças entre os valores previstos e os observados, *i.e.*, para cada valor previsto é calculada a diferença com o valor observado e, com isso, é calculada a média dos módulos dessas diferenças. A Equação 5.2 define a medida de EMA.

$$EMA = \frac{\sum_{i=1}^{m} |x_{n-(m-i)} - \hat{x}_{n-(m-i)}|}{m} \tag{5.2}$$

Coeficiente de correlação (r): essa medida consiste no cálculo de correlação entre os dados observados e os dados previstos. Assim, no contexto deste trabalho é utilizada para verificar a relação de informação contida nos valores previstos em relação aos valores observados, *i.e.*, se à medida que os valores observados aumentam (ou diminuem) os previstos também aumentam (ou diminuem). Se cada conjunto de dados encontra-se normalmente distribuído, a correlação pode ser medida pelo coeficiente de Pearson, apropriado para correlações paramétricas. Caso contrário, deve ser utilizada uma medida apropriada para correlações não-paramétricas, como o coeficiente de Spearman.

Neste trabalho é utilizado o coeficiente de Spearman, pois grande parte das séries temporais descrevem comportamentos que contêm dados que não estão, naturalmente, distribuídos normalmente. Esse coeficiente é calculado por uma função baseada na diferença (d) entre os postos dos dados previstos e observados. Os postos consistem nas posições que

ocupam os dados previstos quando ordenados de acordo com os dados observados. A Equação 5.3 define a função que calcula o coeficiente de Spearman.

$$r = 1 - \frac{6 \times \sum_{i=1}^{m} (d_i)^2}{m \times (m^2 - 1)} \tag{5.3}$$

O coeficiente consiste em um número puro, entre -1 e $+1$, em que valores menores que zero indicam associações negativas e maiores que zero associações positivas. De acordo com Doria (1999) o valor do coeficiente pode classificar a correlação em:

- perfeita, se $|r| = 1$;
- forte, se $0,75 \leq |r| < 1$;
- média, se $0,50 \leq |r| < 0,75$;
- fraca, se $0 < |r| < 0,50$; e
- inexistente, se $r = 0$.

Utilizando essas medidas, é possível comparar objetivamente diferentes configurações dos parâmetros do algoritmo. Essas comparações podem ser realizadas por meio de testes estatísticos de significância, os quais são selecionados de acordo com a natureza de dados que se deseja comparar. Neste trabalho, foi utilizado o teste estatístico não-paramétrico de Wilcoxon para amostras emparelhadas (Flores, 1989; Freedman et al., 1998).

5.3 Implementação da Metodologia

Dentro do projeto Análise Inteligente de Dados de Séries Temporais já mencionado (Ferrero et al., 2007, 2008; Cherman et al., 2008; Spolaôr et al., 2008) e com o intuito de prover um ambiente computacional para o auxílio de pesquisas relacionadas com dados temporais, está sendo desenvolvido no Laboratório de Bioinformática — LABI — da UNIOESTE, em parceria com o Laboratório de Inteligência Computacional — LABIC — ICMC/USP, um sistema computacional para análise de séries temporais denominado *TimeSSys — Time Series System*. O objetivo desse sistema é disponibilizar, em um mesmo *framework*, conjuntos de ferramentas que auxiliem nas diversas etapas do processo de análise de séries temporais. A ideia desse sistema é favorecer a integração de ferramentas de código livre implementadas, bem como de ferramentas desenvolvidas pelos pesquisadores parceiros, em um mesmo *framework*,

disponibilizando funcionalidades para visualização, pré-processamento, previsão, agrupamento, recuperação de conteúdo, entre outras, de series temporais.

O desenvolvimento do sistema *TimeSSys* é baseado em tecnologias e ferramentas livres. A camada de negócio está sendo implementada na linguagem R (R Development Core Team, 2008), a qual disponibiliza uma grande diversidade de funcionalidades relacionadas à análise estatística, análise de séries temporais e visualização. Para compartilhar as ferramentas integradas no *TimeSSys*, é utilizada a tecnologia Rpad[2], baseada no paradigma cliente-servidor, a qual disponibiliza um ambiente simples e flexível para o desenvolvimento de interfaces *WEB*, além de estar diretamente associada com a linguagem R, o que facilita a integração das ferramentas desenvolvidas e a realização e o compartilhamento de experimentos científicos.

Neste trabalho, além da implementação do algoritmo *kNN-TSP*, foram implementadas, dentro do *framework* do *TimeSSys*, diversas ferramentas para auxiliar na aplicação da metodologia proposta. Essas ferramentas, para as quais serão futuramente construídas interfaces com Rpad, foram executadas por meio de *scripts* específicos e incluem:

Módulo de Experimentos: permite executar automaticamente experimentos a partir de uma série temporal e do número de valores a serem previstos;

Módulo de Extração de Medidas de Avaliação: permite extrair e calcular medidas de avaliação dos resultados dos experimentos, como erro médio quadrático, erro médio absoluto, desvio-padrão absoluto, erro médio absoluto percentual e coeficiente de correlação, entre outros;

Módulo de Comparação de Resultados: possibilita a comparação de resultados entre diferentes configurações do algoritmo de previsão por meio da aplicação de testes de hipótese e da criação de gráficos;

Módulo de Geração de Gráficos de Previsão: produz arquivos gráficos para visualizar os dados previstos pelo algoritmo de previsão considerando as diferentes configurações; e

Módulo de Análise de Correlação: para pares de séries temporais correlacionadas, permite verificar se os valores previstos por algoritmos de previsão permitem manter a correlação dos dados observados.

[2]http://www.rpad.org/Rpad/.

5.4 Considerações Finais

Neste capítulo foi apresentada a metodologia utilizada para a avaliação das diversas configurações do algoritmo k*NN-TSP*. As três fases que constituem a metodologia foram descritas detalhadamente e a implementação computacional foi apresentada. Essa metodologia é aplicada a conjuntos de dados artificiais, no Capítulo 6, e a conjuntos de dados reais, no Capítulo 7.

CAPÍTULO

6

Avaliação Experimental

6.1 Considerações Iniciais

Como descrito no Capítulo 4, a previsão de dados pelo algoritmo *kNN-TSP* requer a definição dos seguintes parâmetros, os quais podem influenciar na qualidade das previsões realizadas pelo algoritmo:

(a) tamanho w da janela para extrair as subsequências;

(b) conjunto de exemplos (séries) de treinamento;

(c) medida de similaridade;

(d) cardinalidade do conjunto de séries similares; e

(e) função de previsão.

Neste capítulo, é descrita a avaliação de *kNN-TSP* utilizando conjuntos de dados que consistem de séries temporais artificiais, *i.e.*, geradas por funções matemáticas. O fato de conhecer o modelo que gera os dados permite realizar uma avaliação controlada dos experimentos. Neste caso, o parâmetro w (a) pode ser determinado considerando a periodicidade do modelo, se for o caso. Quanto ao parâmetro (b), será usado todo o conjunto de exemplos gerados pelo modelo em um intervalo fixo de tempo. Assim, o foco da avaliação apresentada está na avaliação dos parâmetros (c), (d) e (e), os quais serão analisados utilizando diferentes critérios de seleção de vizinhos mais próximos e número

de séries similares consideradas na previsão (vizinhos próximos), bem como diversas funções de avaliação.

A seguir são apresentadas as séries temporais artificiais utilizadas, a configuração dos experimentos e a discussão dos resultados.

6.2 Descrição dos Conjuntos de Dados

Os dados foram gerados utilizando modelos (funções) pertencentes a duas famílias:

(a) série temporais de modelos sazonais; e

(b) sistemas caóticos.

Modelos pertencentes à família (a) permitem avaliar a capacidade do algoritmo em séries com comportamento razoavelmente previsível, enquanto as que pertencem à família (b) geram séries com comportamentos pouco previsíveis, com ciclos que não se repetem (Gandur, 1999, p. 60).

As série temporais foram geradas utilizando três modelos pertencentes à família (a) e dois modelos pertencentes à família (b), os quais são descritos nas próximas seções. Os modelos da família (b) foram selecionados após realizar uma pesquisa bibliográfica que mostrou que eles são frequentemente utilizados na literatura para avaliar algoritmos de modelagem de comportamentos biológicos.

6.2.1 Séries Temporais de Modelos Sazonais

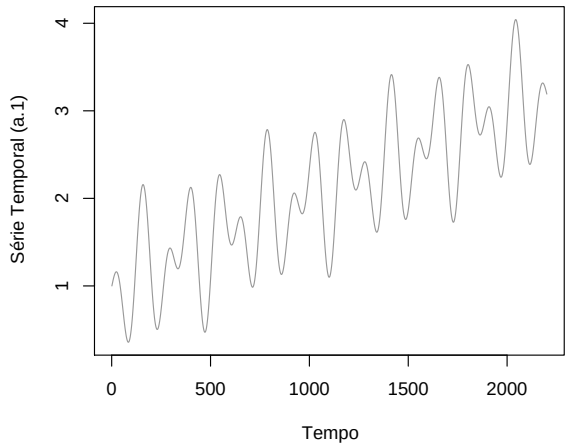
Os modelos pertencentes a esta família geram séries sazonais que contêm tendência e/ou mudança de amplitude de série. Em (Kulesh et al., 2008) são apresentadas três séries temporais que permitem avaliar os métodos de previsão nesse sentido:

(a.1) série temporal de dependência sazonal;

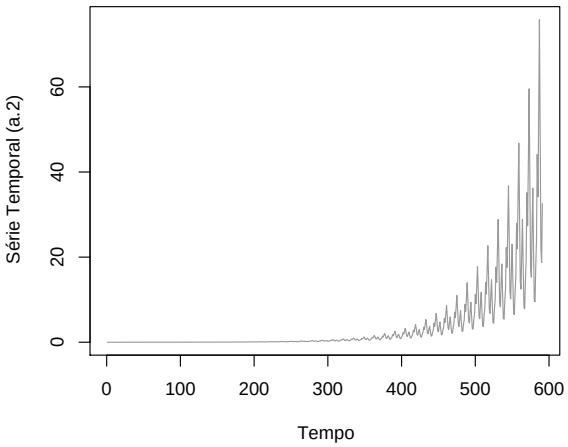
(a.2) série temporal de sazonalidade multiplicativa; e

(a.3) série temporal de alta frequência.

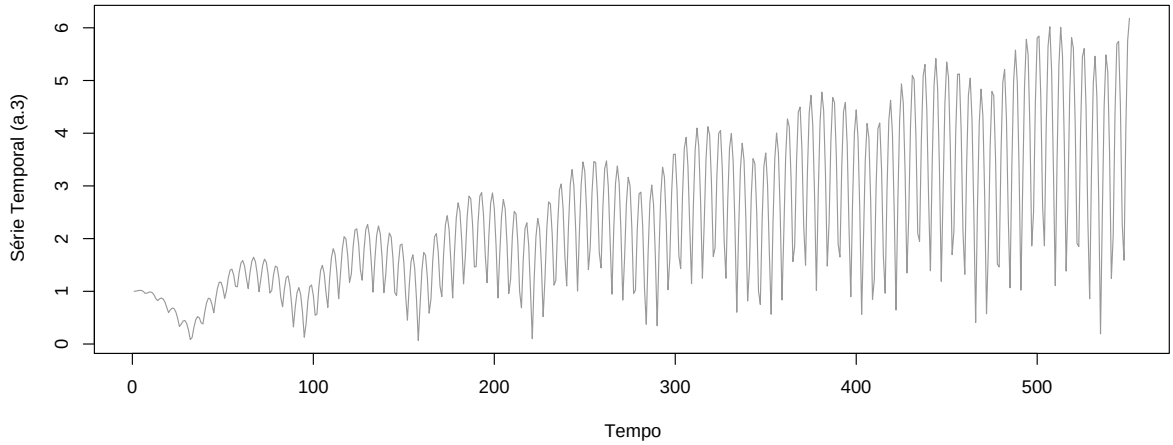
A representação gráfica desses três modelos é apresentada na Figura 6.1. A seguir são descritas as características dos modelos e as expressões matemáticas que os geram.



(a) Série Temporal (a.1)



(b) Série Temporal (a.2)



(c) Série Temporal (a.3)

Figura 6.1: Séries temporais artificiais sazonais geradas a partir de funções matemáticas

(a.1) Série temporal de dependência sazonal: os dados são gerados por uma função que considera sazonalidade constante e tendência linear, definida pela Equação 6.1.

$$ST_{a.1}(t) = cos\left(\frac{t}{25}\right) \times sin\left(\frac{t}{100}\right) + \frac{t}{1000} + 1, t \in [0, 2200] \tag{6.1}$$

A série temporal gerada pode ser visualizada na Figura 6.1(a), na qual é possível observar a variação da tendência ao longo do tempo, assim como a sazonalidade constante.

(a.2) Série temporal de sazonalidade multiplicativa: os dados são gerados por uma função, proposta por Kulesh et al. (2008), que considera variação de tendência não-linear e sazonalidade multiplicativa, pela qual as oscilações crescem ao longo do tempo, definida pela Equação 6.2.

$$ST_{a.2}(t) = \begin{cases} R(t), & t \in [0; 79] \\ \frac{ST_{a.2}(t-t_0)^2}{ST_{a.2}(t-t_0)}, & t \in [80; 590], t_0 = 14 \end{cases} \tag{6.2}$$

onde $R(t) = \frac{t}{(7*10^4)} \times \left(sin\left(\frac{t}{350}\right) \times cos\left(\frac{9t}{7}\right) + 10\right)$

A série temporal gerada pode ser visualizada na Figura 6.1(b), na qual é possível observar a variação da tendência e a sazonalidade crescente ao longo do tempo.

(a.3) Série temporal de frequência alta: os dados são gerados por uma função que considera sazonalidade multiplicativa e suave aumento de amplitude, definida pela Equação 6.3.

$$ST_{a.3}(t) = \frac{t}{100} \times \left|sin\left(\frac{t}{2}\right)\right| + cos\left(\frac{t}{20}\right), t \in [0; 550] \tag{6.3}$$

A série temporal gerada pode ser visualizada na Na Figura 6.1(c), na qual é possível observar essas características.

6.2.2 Séries Temporais de Sistemas Caóticos

Como mencionado, os modelos que pertencem a esta família geram séries com comportamento pouco previsível, com ciclos que não se repetem. A avaliação por meio desse tipo de séries é importante, devido ao fato de que parte das séries temporais que representam eventos naturais apresentam comportamento caótico. Evidências desse tipo de comportamento têm sido encontradas em hidrologia (Sivakumar, 2000), na análise de eletrocardiogramas obtidos em períodos de sono (Ferri et al., 1996), na previsão de índices de bolsas de valores (Espinosa et al., 2005), entre outros. Neste trabalho, são utilizados os sistemas de:

(b.1) Lorenz; e

(b.2) Mackey-Glass,

para gerar séries temporais com comportamento caótico.

(b.1) Sistema de Lorenz: consiste em um sistema de equações diferenciais que permitem gerar séries temporais de comportamento aperiódico e imprevisível. No estudo de modelos de previsão de tempo, Lorenz percebeu que essa dinâmica apresenta uma característica interessante, a qual consiste de que dados dois pontos próximos no sistema, eles podem pertencer a trajetórias ou rotas divergentes (Gandur, 1999, p. 56). Em (McNames, 1999) podem ser encontrados detalhes a respeito das equações

que permitem gerar as séries de Lorenz. Neste trabalho, é utilizada a série temporal criada utilizando esse sistema disponibilizada por James McNames[1], ilustrada na Figura 6.2.

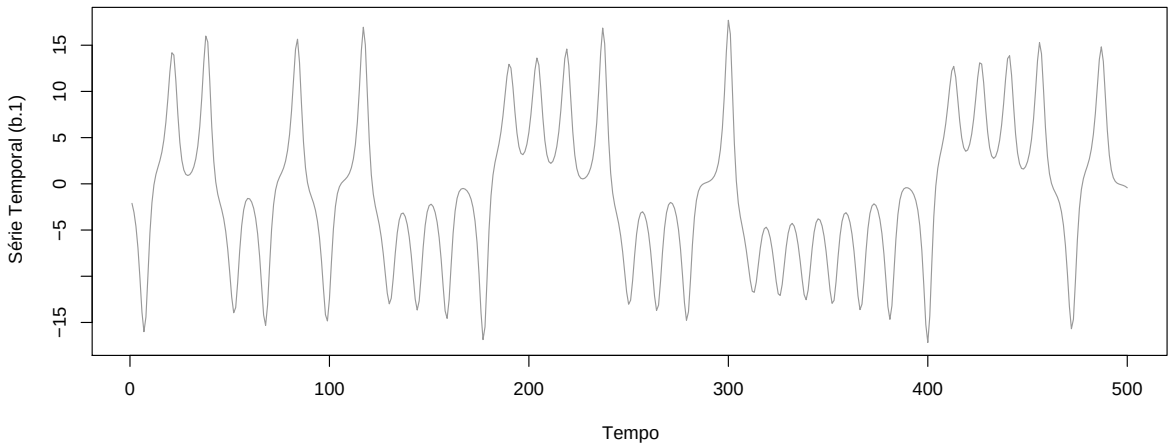


Figura 6.2: Série temporal referente aos primeiros 500 pontos experimentais de uma série de Lorenz

(b.2) Sistema de Mackey-Glass: consiste em um sistema de equações que modela a formação de células sanguíneas brancas (linfócitos) (Mackey e Glass, 1977). Atualmente, é um dos sistemas de equações mais utilizado para gerar séries temporais para avaliar a capacidade de métodos de previsão para a modelagem de comportamentos caóticos. O sistema de equações e os valores dos parâmetros utilizados para a geração dessas séries podem ser encontrados em (McNames, 1999, p. 134–135). Neste trabalho, é utilizada a série temporal, criada utilizando esse sistema de equações, também disponibilizada por James McNames, ilustrada na Figura 6.3.

6.3 Configuração dos Experimentos

Cada um dos cinco conjuntos de dados anteriores foi submetido à metodologia apresentada na Seção 5.2, considerando o algoritmo *kNN-TSP* para diferentes configurações de parâmetros. Para cada conjunto de dados é utilizado um valor, definido por Kulesh et al. (2008), de parâmetro (a) — tamanho w da janela para extrair subsequências. Deve ser observado que não é foco deste trabalho a investigação da influência da dimensão de imersão da série

[1] http://web.cecs.pdx.edu/~mcnames/DataSets/index.html. Último acesso em: 22 de janeiro de 2008.

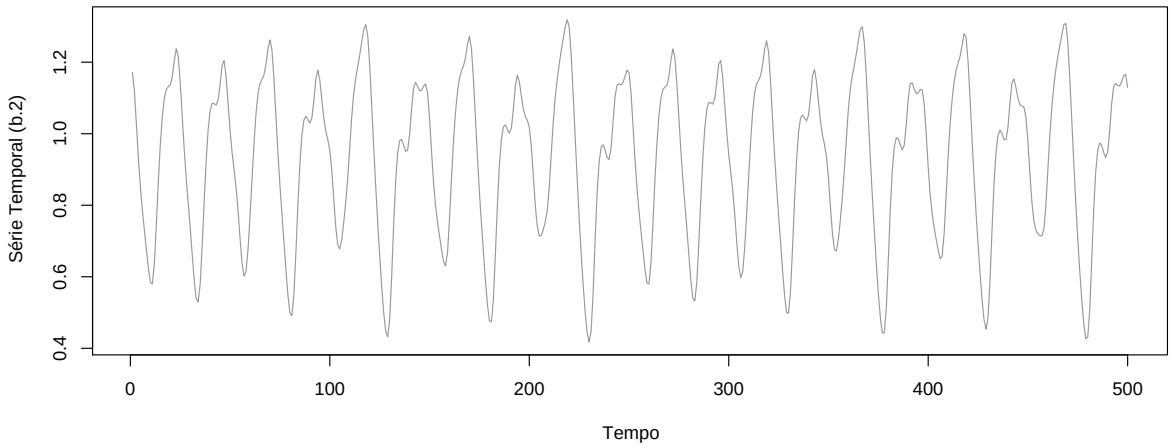


Figura 6.3: Série temporal referente aos primeiros 500 pontos experimentais de uma série de Mackey-Glass

temporal na previsão de dados. Quanto ao parâmetro (b), conjunto de séries de treinamento, a totalidade da série é utilizada. Na Tabela 6.1 são apresentados os valores do parâmetro (a) e a quantidade de valores previstos, m, para avaliar o algoritmo, identificando cada base de dados considerada pela variável i.

Tabela 6.1: Configuração dos parâmetros (a) e (b)

i	**Id**	**Série Temporal**	**(a)**	**#Valores previstos**
1	(a.1)	Dependência sazonal	100	220
2	(a.2)	Sazonalidade multiplicativa	15	88
3	(a.3)	Alta frequência	70	55
4	(b.1)	Lorenz	25	100
5	(b.2)	Mackey-Glass	7	100

Neste trabalho são avaliados os parâmetros (c), (d) e (e). Em relação ao parâmetro (c), medida de similaridade, são considerados os critérios de seleção de vizinhos próximos por similaridade e similaridade e tempo, apresentados na Seção 4.4.2. Em relação ao parâmetro (d), cardinalidade do conjunto de séries similares, são considerados os valores de cardinalidade $k = 1, 2, 3, 4$ e 5, no intuito de verificar a influência do número de vizinhos mais próximos para a previsão dos dados. Em relação ao parâmetro (e), função de previsão, são utilizadas as funções apresentadas na Seção 4.4.3, média de valores e média de valores relativos. Na Tabela 6.2 são descritas as configurações experimentais utilizadas, as quais são identificadas pelo valor do índice j. Cada linha da tabela representa uma configuração experimental, a qual será realizada variando o valor de k de 1 a 5.

Tabela 6.2: Configuração dos parâmetros (c) e (e) com parâmetro (d) definido por $k = 1, 2, 3, 4, 5$

j	(c)	(e)
1	Similaridade não normalizada	f_{MV}
2	Similaridade não normalizada e tempo	f_{MV}
3	Similaridade normalizada	f_{MVR}
4	Similaridade normalizada e tempo	f_{MVR}

Desse modo, cada configuração experimental é identificada por $CE_{i,j,k}$, onde i corresponde à base de dados utilizada, j corresponde à configuração e k corresponde ao número de vizinhos próximos considerados, totalizando 100 experimentos.

A seguir, os resultados dos experimentos são mostrados graficamente. Os detalhes correspondentes aos resultados numéricos de cada configuração experimental, bem como os gráficos que mostram os valores previstos para cada série temporal, nas diversas configurações, encontram-se no documento *kNN-TSP* Resultados Experimentais (Ferrero, 2009).

6.4 Análise dos Resultados e Discussão

Para a avaliação do desempenho do *kNN-TSP*, com base nos conjuntos de dados apresentados, a análise dos resultados é realizada, primeiramente, em relação ao critério de seleção de vizinhos mais próximos e, posteriormente, em relação às funções de previsão. Em ambos os casos, para cada uma das bases de dados artificiais, é utilizado o teste não-paramétrico de Wilcoxon (Freedman et al., 1998) para dados emparelhados, considerando nível de significância de 95%, no intuito de verificar diferença estatisticamente significativa — **d.e.s** — nos resultados. Os $p - valores$ obtidos encontram-se no Apêndice A.

6.4.1 Seleção dos Vizinhos mais Próximos: similaridade versus similaridade e tempo

Seguem os resultados para os critérios de seleção dos vizinhos mais próximos usando: (1) similaridade e (2) similaridade e tempo. Neste caso, o teste de Wilcoxon é utilizado para dados emparelhados segundo o critério de similaridade e similaridade e tempo, de maneira independente para as funções f_{MV} e f_{MVR}, utilizando como medida o EMA e o coeficiente de correlação r de Spearman. Os $p - valores$ obtidos encontram-se na Tabela A.1 no Apêndice A.

Série de dependência sazonal: os resultados usando esta série são apresen-

tados na Figura 6.4.

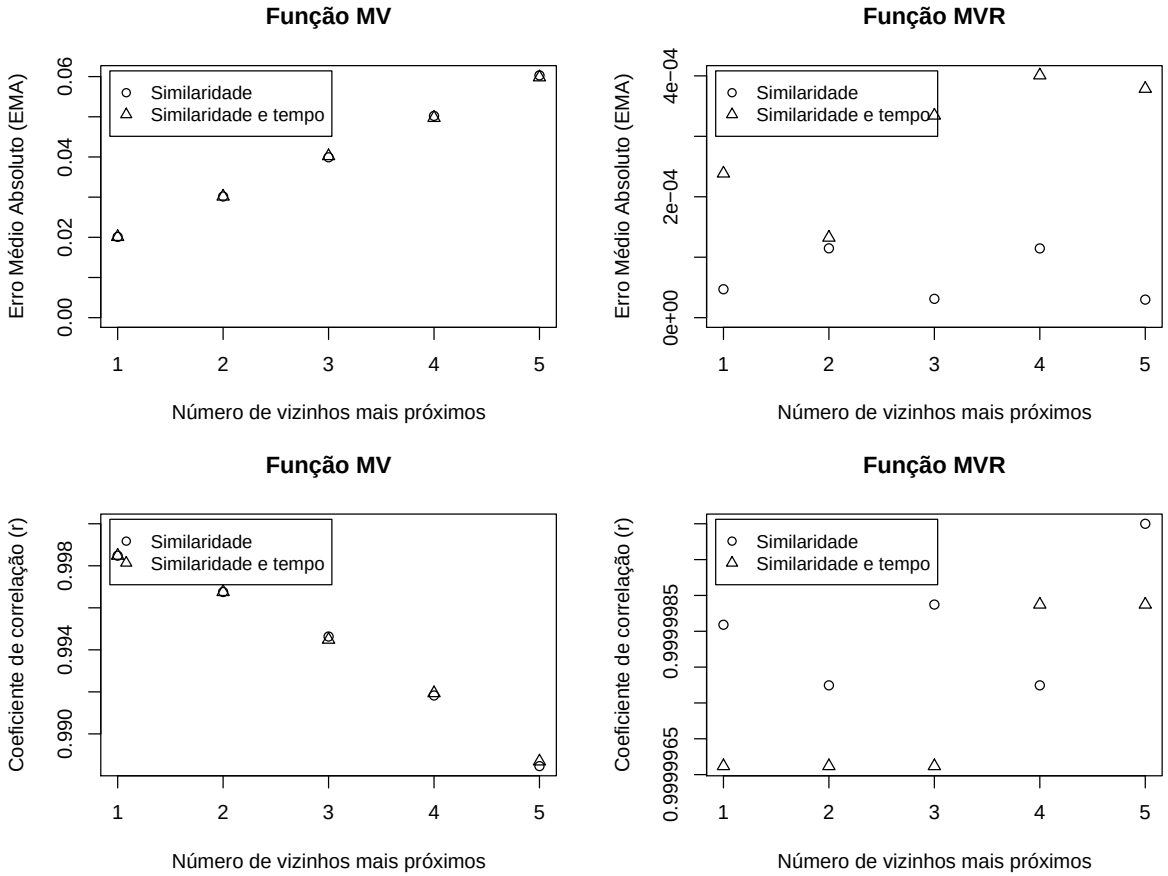


Figura 6.4: Gráficos referentes às medidas de EMA e r de previsão da série de dependência sazonal: similaridade *versus* similaridade e tempo

Como pode ser observado, o aumento do valor de k influenciou negativamente no desempenho da função f_{MV} (o erro aumenta com k e a correlação diminui). Entre os critérios de seleção de vizinhos próximos foi possível constatar que:

- Para a função f_{MV} **não** houve **d.e.s** em relação ao EMA e ao r;
- Para a função f_{MVR} houve **d.e.s** em relação ao EMA, em que o critério de similaridade teve melhor desempenho. Em relação ao r **não** houve **d.e.s**.

Série de sazonalidade multiplicativa: os resultados usando esta série são apresentados na Figura 6.5.

Semelhante à série de dependência sazonal, o aumento do valor k influenciou negativamente no desempenho da função f_{MVR}. Entre os critérios de seleção de vizinhos próximos foi possível constatar que:

- Para a função f_{MV} **não** houve **d.e.s** em relação ao EMA e ao r;

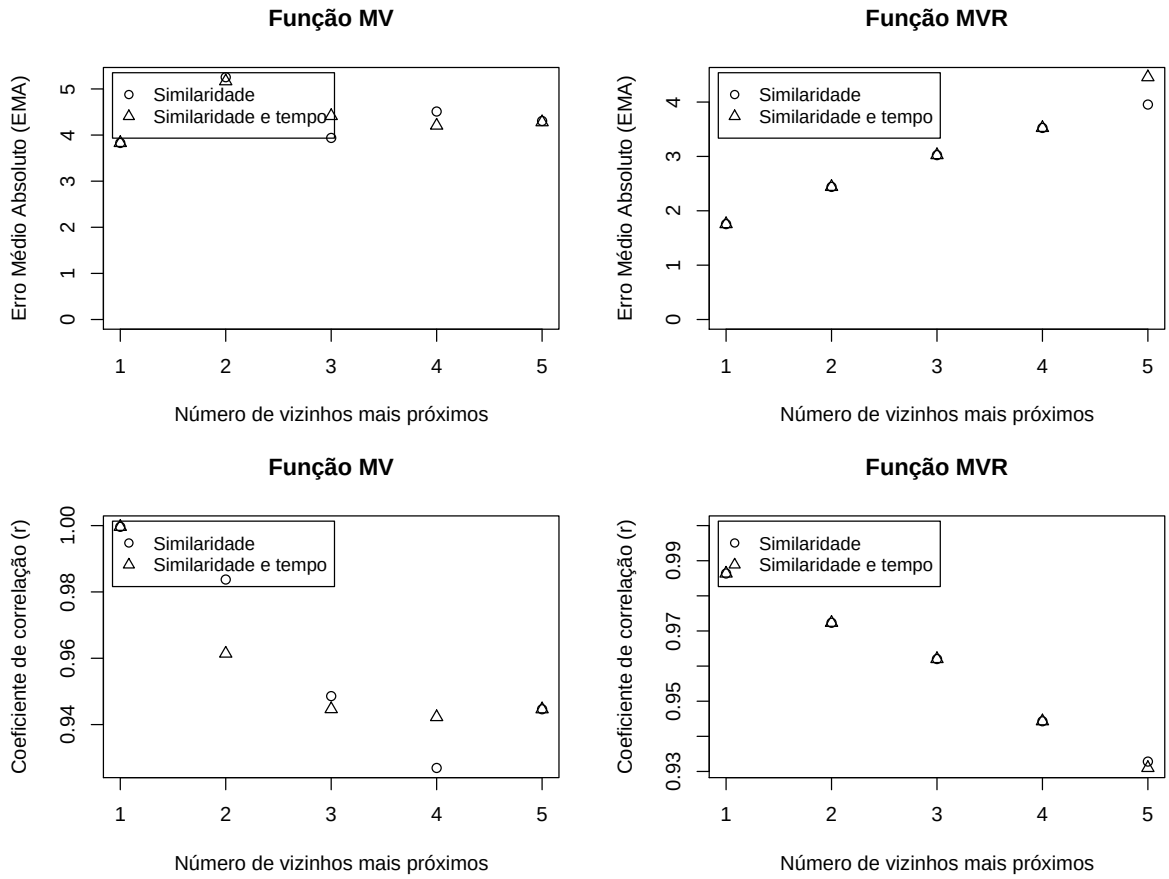


Figura 6.5: Gráficos referentes às medidas de EMA e r de previsão da série de sazonalidade multiplicativa: similaridade *versus* similaridade e tempo

- Para a função f_{MVR} **não** houve **d.e.s** em relação ao EMA e ao r.

Série de alta frequência: os resultados usando esta série são apresentados na Figura 6.6.

Pode ser observado que o valor $k = 1$ apresentou pior desempenho em relação aos outros valores de k. Entre os critérios de seleção de vizinhos próximos foi possível constatar que:

- Para a função f_{MV} **não** houve **d.e.s** em relação ao EMA e ao r;
- Para a função f_{MVR} **não** houve **d.e.s** em relação ao EMA e ao r.

Série de Lorenz: os resultados usando esta série são apresentados na Figura 6.7.

Pode ser observado que não existe um padrão de desempenho de previsão relacionado com o valor de k. Entre os critérios de seleção de vizinhos próximos foi possível constatar que:

- Para a função f_{MV} houve **d.e.s** em relação ao EMA e ao r, em que o critério de similaridade teve melhor desempenho;

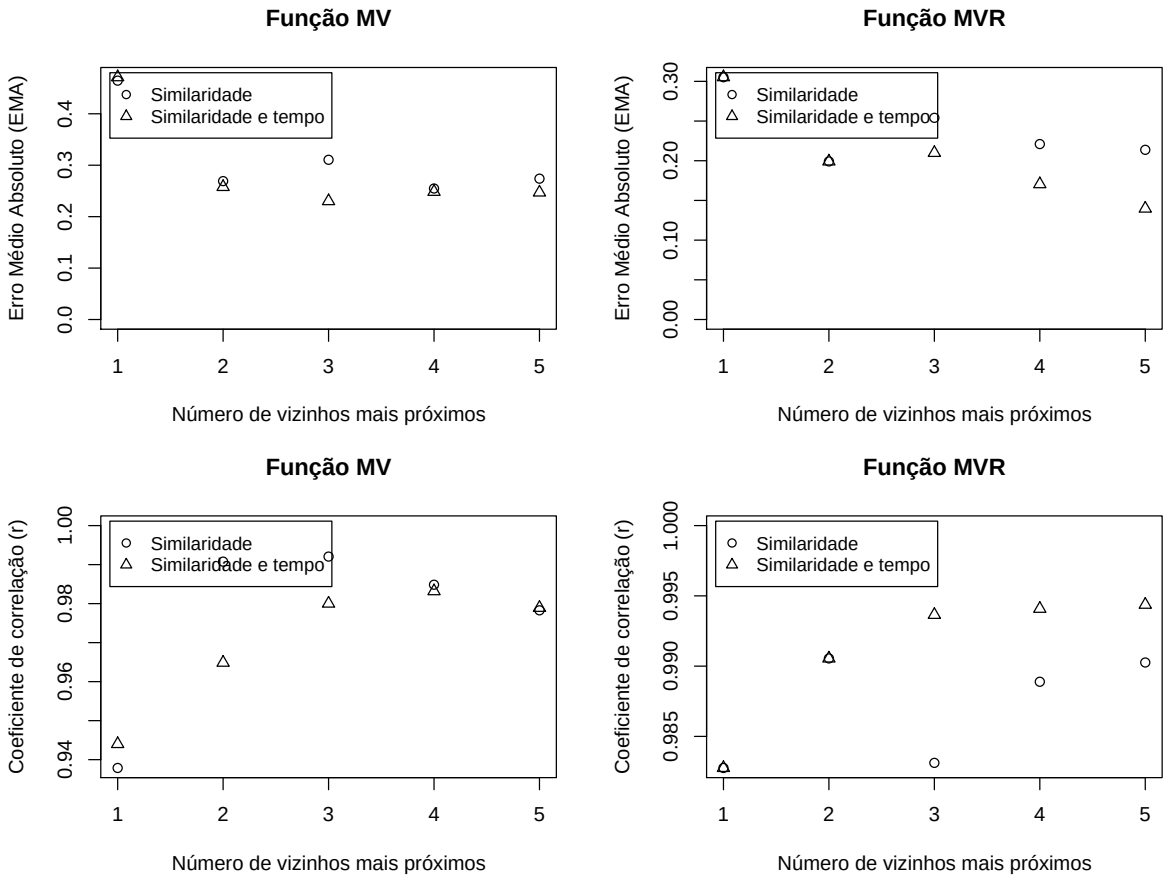


Figura 6.6: Gráficos referentes às medidas de EMA e r de previsão da série de alta frequência: similaridade *versus* similaridade e tempo

- Para a função f_{MVR} houve **d.e.s** em relação ao EMA e ao r, em que o critério de similaridade teve melhor desempenho.

Série de Mackey-Glass: os resultados usando esta série são apresentados na Figura 6.8.

Como pode ser observado, o aumento do valor de k utilizando o critério de similaridade e tempo também influenciou negativamente no desempenho da função f_{MV}. Entre os critérios de seleção de vizinhos próximos foi possível constatar que:

- Para a função f_{MV} houve **d.e.s** em relação ao EMA e ao r, em que o critério de similaridade teve melhor desempenho;
- Para a função f_{MVR} houve **d.e.s** em relação ao EMA e ao r, em que o critério de similaridade teve melhor desempenho.

Resumo das Comparações

Na Tabela 6.3 é apresentado o resumo das comparações para a avaliação dos critérios de seleção de vizinhos próximos. Nessa tabela, as colunas EMA

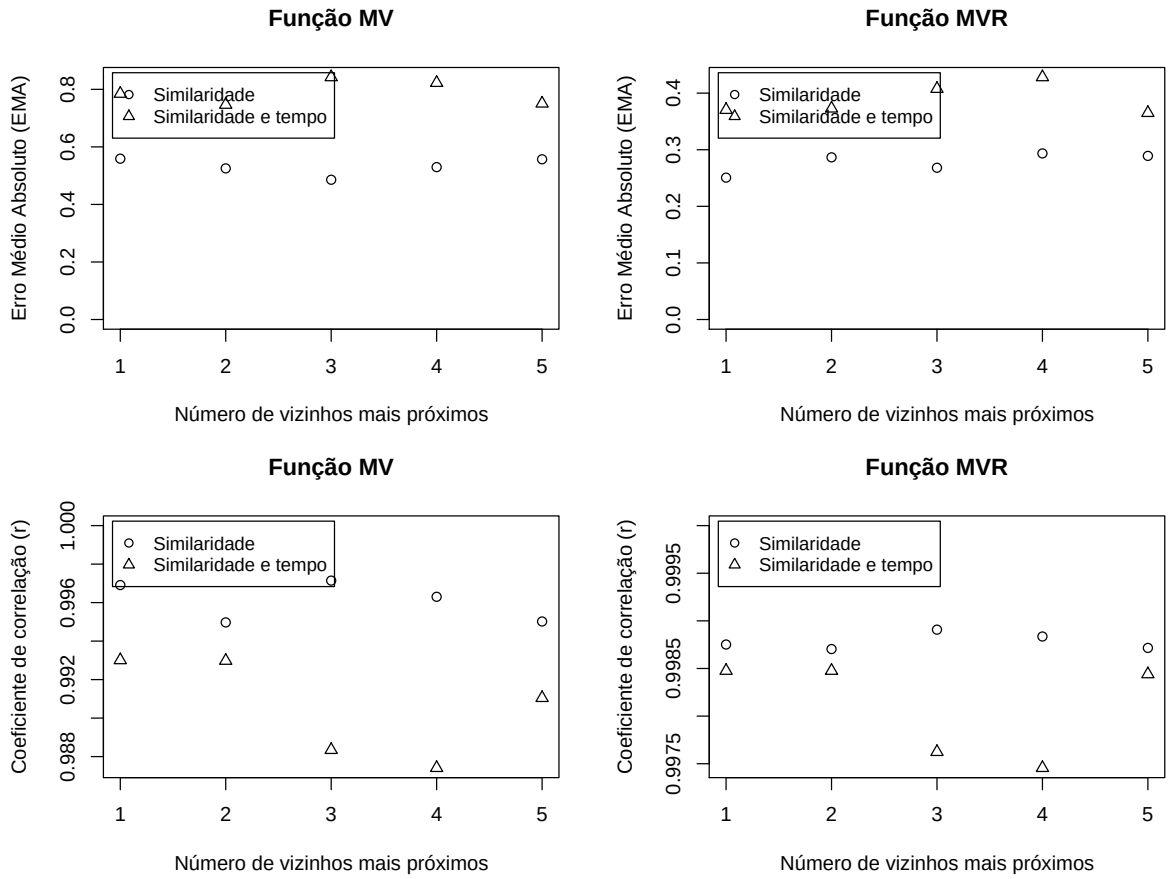


Figura 6.7: Gráficos referentes às medidas de EMA e r de previsão da série de de Lorenz: similaridade *versus* similaridade e tempo

e r indicam os resultados com **d.e.s** entre os critérios de seleção de vizinhos próximos para as funções de previsão f_{MV} e f_{MVR}. O símbolo ✓ indica os resultados com **d.e.s**.

Tabela 6.3: Resumo das comparações entre os critérios de similaridade e similaridade e tempo

		f_{MV}			f_{MVR}		
i	**Série Temporal**	EMA	r	Melhor desempenho	EMA	r	Melhor desempenho
1	Dependência sazonal				✓	✓	similaridade
2	Sazonalidade multiplicativa						
3	Alta frequência						
4	Lorenz	✓	✓	similaridade	✓	✓	similaridade
5	Mackey-Glass	✓	✓	similaridade	✓	✓	similaridade

Apesar do critério de seleção de vizinhos próximos proposto neste trabalho ser interessante do ponto de vista de considerar o tempo uma característica importante para complementar a medida de similaridade no processo de seleção desses vizinhos próximos, os resultados mostraram que, para ambas as funções de previsão f_{MV} e f_{MVR}, o critério de similaridade apresentou melhor

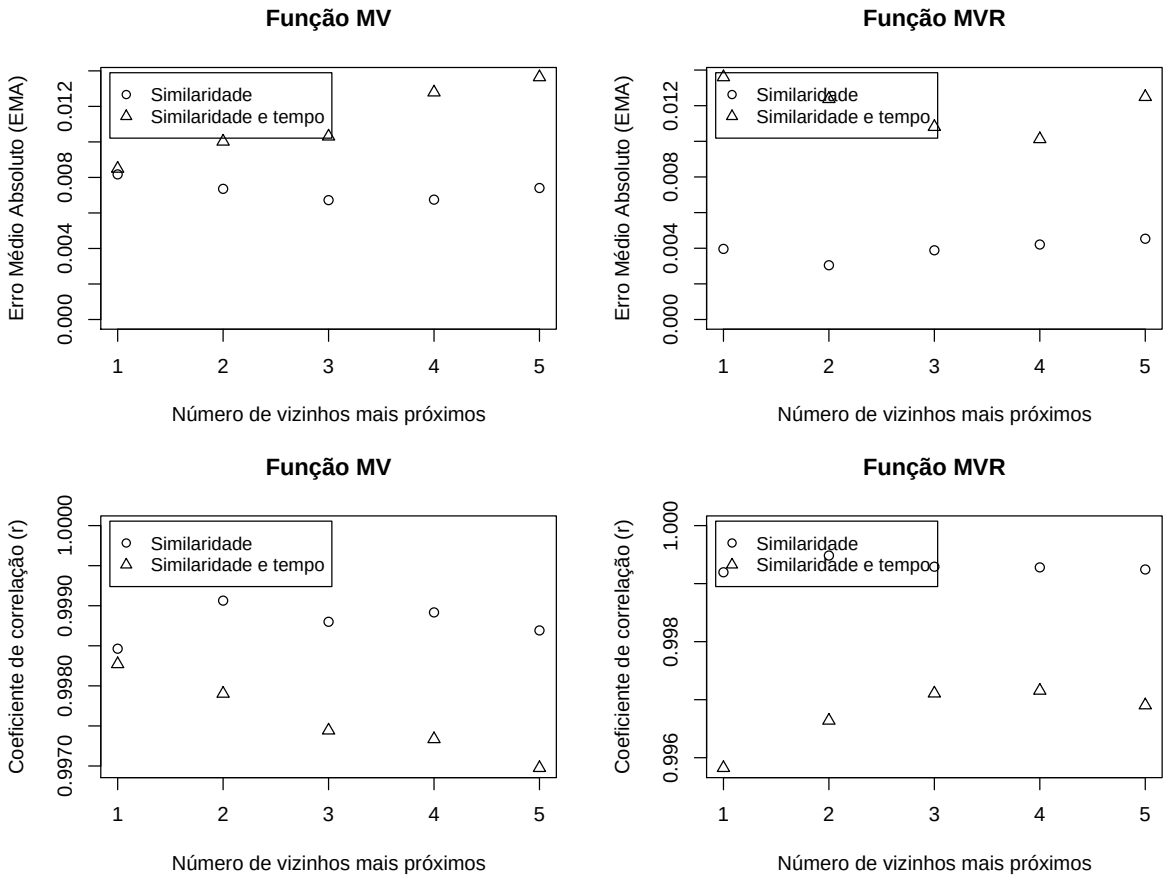


Figura 6.8: Gráficos referentes às medidas de EMA e r de previsão da série de Mackey-Glass: similaridade *versus* similaridade e tempo

desempenho. O critério proposto neste trabalho considera importantes aqueles padrões mais próximos do estado atual. Entretanto, foi observado que, quando uma subsequência próxima é encontrada, as subsequências adjacentes são comumente também similares, devido ao fato de que a diferença entre o padrão encontrado e a subsequência imediatamente adjacente em relação padrão procurado tende e ser baixa. Devido a isso, o critério de similaridade e tempo proposto tende a selecionar subsequências superpostas, o que pode influenciar negativamente no desempenho de k*NN-TSP*.

6.4.2 Funções de Previsão: f_{MV} versus f_{MVR}

Nesta seção é apresentado o estudo comparativo entre as funções f_{MV} e f_{MVR} para a previsão de valores futuros. Para cada base de dados gerada é avaliada a capacidade de previsão de cada uma dessas funções em relação à medida de erro médio absoluto e coeficiente de correlação. Também nesse caso, o teste de Wilcoxon é utilizado para dados emparelhados segundo os valores das funções f_{MV} e f_{MVR}, de maneira independente para os dois critérios de seleção de vizinhos mais próximos, utilizando como medidas o EMA e o

coeficiente de correlação r de Spearman. Os $p - valores$ obtidos encontram-se na Tabela A.2 no Apêndice A.

Série de dependência sazonal: os resultados usando esta série são apresentados na Figura 6.9.

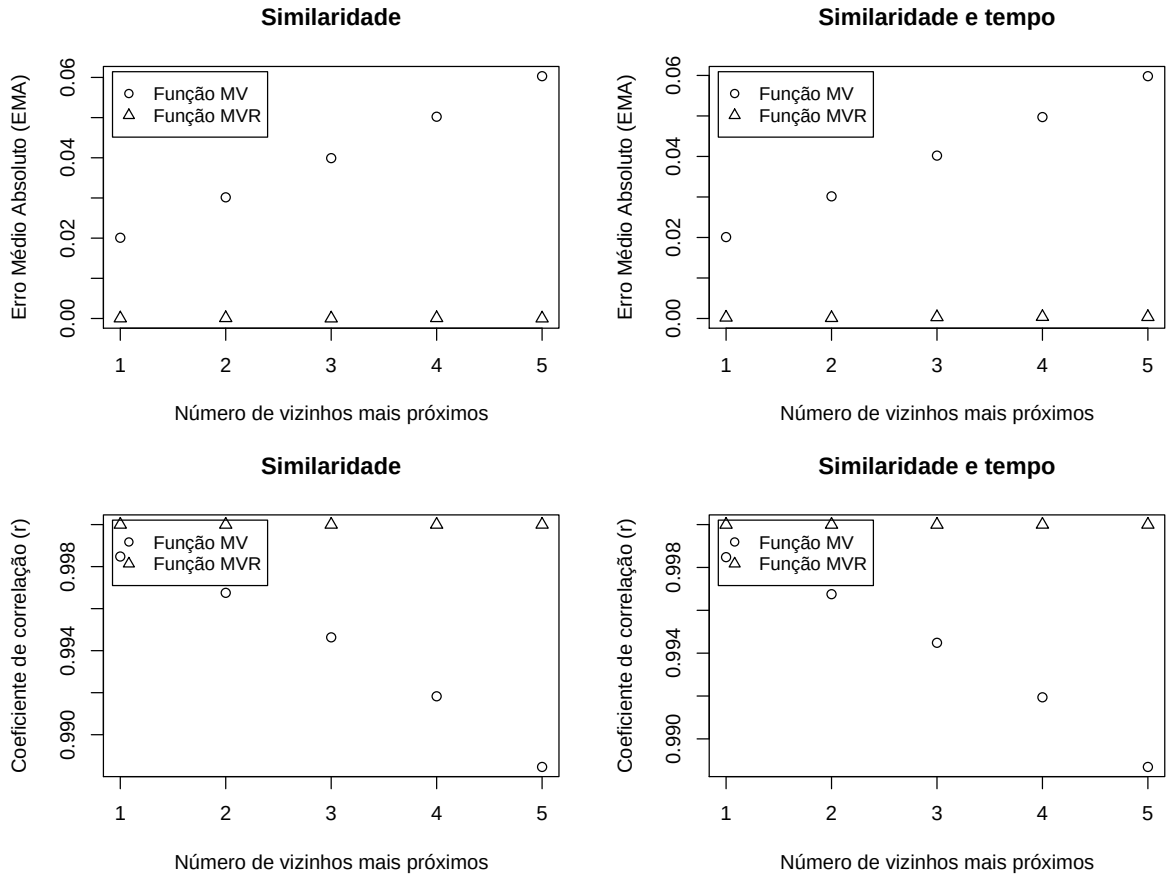


Figura 6.9: Gráficos referentes às medidas de EMA e r de previsão da série de dependência sazonal: f_{MV} *versus* f_{MVR}

Como pode ser observado, o aumento do valor de k influencia negativamente no desempenho da função f_{MV}. Entre as funções de previsão foi possível constatar que:

- Para o critério de similaridade houve **d.e.s** em relação ao EMA e ao r, em que a função f_{MVR} teve melhor desempenho;
- Para o critério de similaridade e tempo houve **d.e.s** em relação ao EMA e ao r, em que a função f_{MVR} teve melhor desempenho.

Série de sazonalidade multiplicativa: os resultados usando esta série são apresentados na Figura 6.10.

Pode ser observado que o aumento do valor de k influencia negativamente no desempenho da função f_{MVR}. Entre as funções de previsão foi possível constatar que:

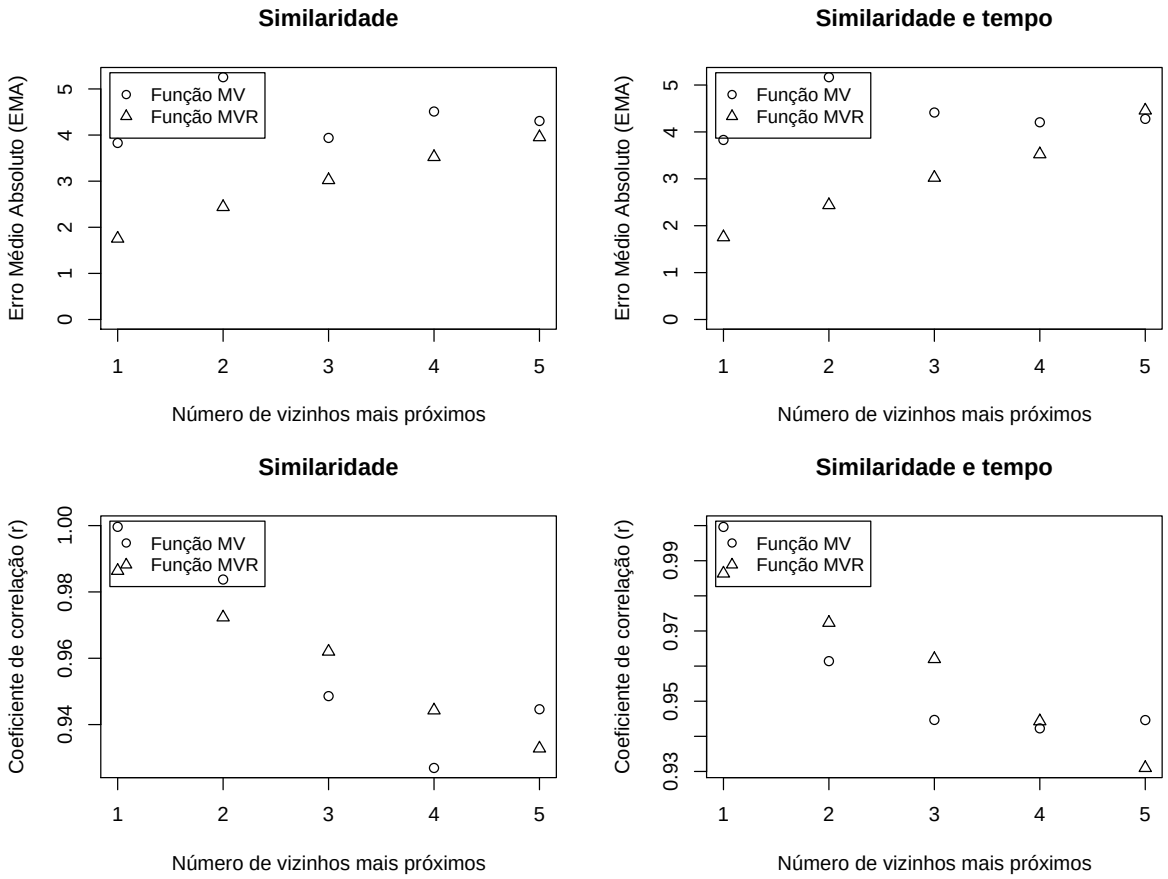


Figura 6.10: Gráficos referentes às medidas de EMA e r de previsão da série de sazonalidade multiplicativa: f_{MV} *versus* f_{MVR}

- Para o critério de similaridade houve **d.e.s** em relação ao EMA, em que a função f_{MVR} teve melhor desempenho. Em relação ao r **não** houve **d.e.s**;
- Para o critério de similaridade e tempo **não** houve **d.e.s** em relação ao EMA e ao r.

Série de alta frequência: os resultados usando esta série são apresentados na Figura 6.11.

Pode ser observado que o valor $k = 1$ apresentou pior desempenho em relação aos outros valores de k. Entre as funções de previsão foi possível constatar que:

- Para o critério de similaridade **não** houve **d.e.s** em relação ao EMA e ao r;
- Para o critério de similaridade e tempo houve **d.e.s** em relação ao r, em que a função f_{MVR} teve melhor desempenho. Em relação ao EMA **não** houve **d.e.s**.

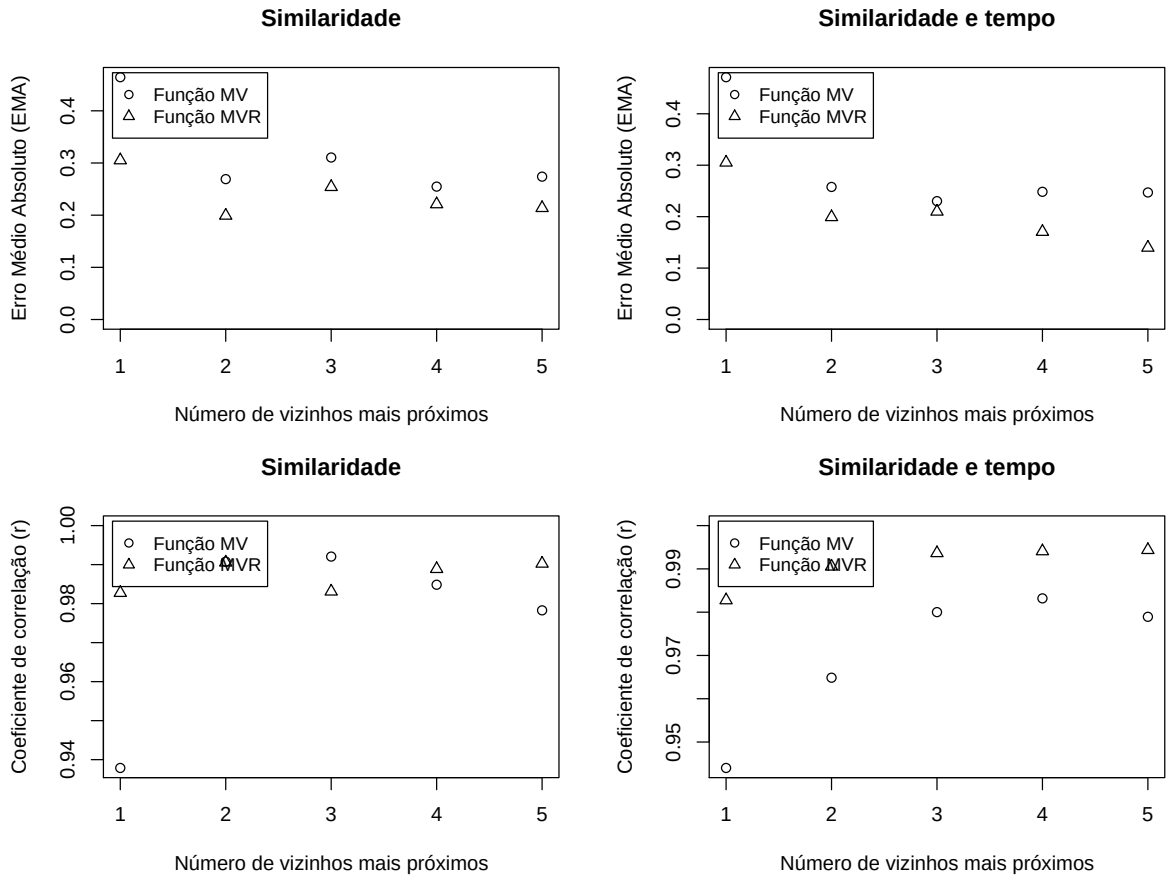


Figura 6.11: Gráficos referentes às medidas de EMA e r de previsão da série de alta frequência: f_{MV} *versus* f_{MVR}

Série de Lorenz: os resultados usando esta série são apresentados na Figura 6.12.

Pode ser observado que não existe um padrão de desempenho de previsão relacionado com o valor de k. Entre as funções de previsão foi possível constatar que:

- Para o critério de similaridade houve **d.e.s** em relação ao EMA e ao r, em que a função f_{MVR} teve melhor desempenho;
- Para o critério de similaridade e tempo houve **d.e.s** em relação ao EMA e ao r, em que a função f_{MVR} teve melhor desempenho.

Série de Mackey-Glass: os resultados usando esta série são apresentados na Figura 6.13.

Pode ser observado que o aumento do valor de k utilizando o critério de similaridade e tempo também influenciou negativamente no desempenho da função f_{MV}. Entre as funções de previsão foi possível constatar que:

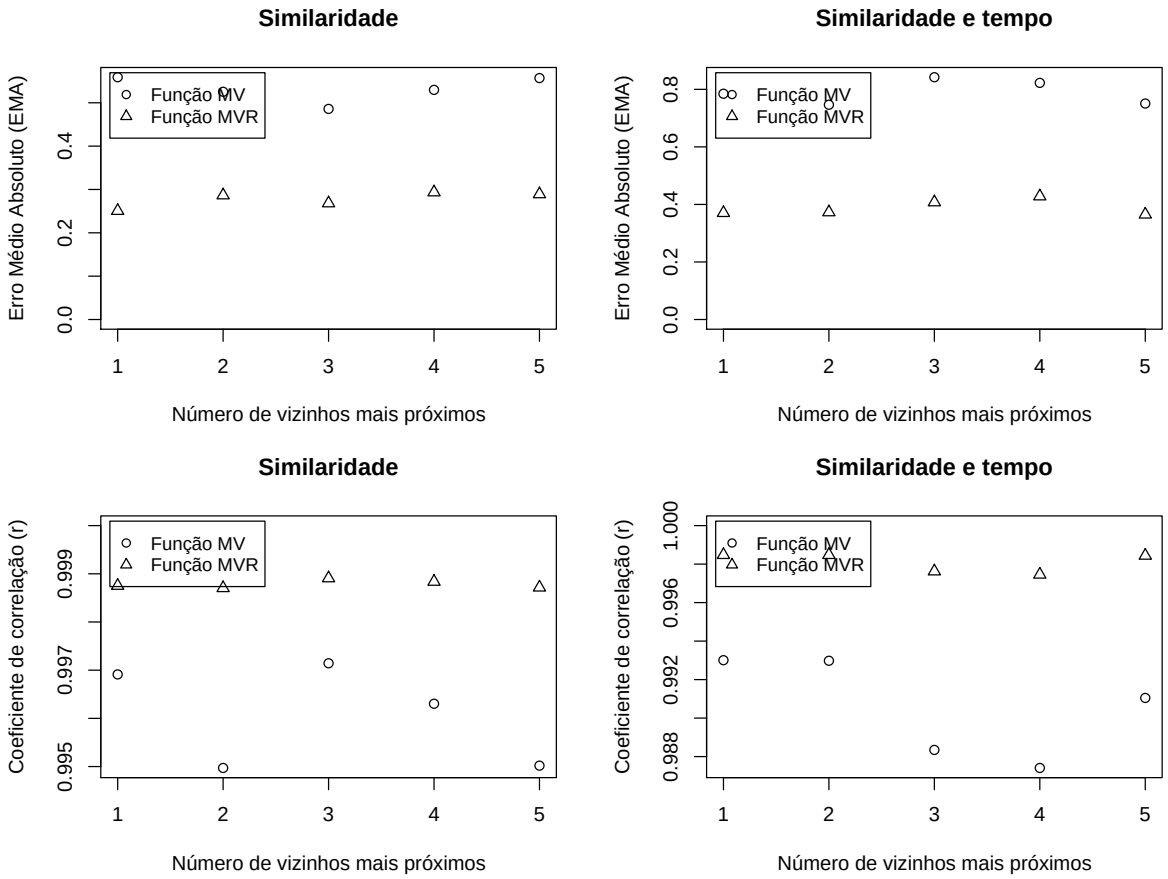


Figura 6.12: Gráficos referentes às medidas de EMA e r de previsão da série de Lorenz: f_{MV} *versus* f_{MVR}

- Para o critério de similaridade houve **d.e.s** em relação ao EMA e ao r, em que a função f_{MVR} teve melhor desempenho;
- Para o critério de similaridade e tempo houve **d.e.s** em relação ao r, em que a função f_{MV} teve melhor desempenho. Em relação ao EMA **não** houve **d.e.s**.

Resumo das Comparações

Na Tabela 6.4 é apresentado o resumo das comparações de avaliação das funções f_{MV} e f_{MVR}. Nessa tabela, as colunas indicam os resultados com **d.e.s** entre as funções de previsão f_{MV} e f_{MVR} para os dois critérios de seleção de vizinhos mais próximos. O símbolo ✓ indica os resultados com **d.e.s**.

Como pode ser observado, das oito ocorrências de **d.e.s**, a função f_{MVR} teve melhor desempenho em sete delas.

Na série temporal mais controlada, dependência sazonal, a função f_{MVR} permitiu aproveitar de modo adequado a característica dessa série temporal, de padrões em vários níveis, em relação à função f_{MV}, a qual limita-se a encontrar padrões que estejam no mesmo nível. Para a série temporal de sazonalidade

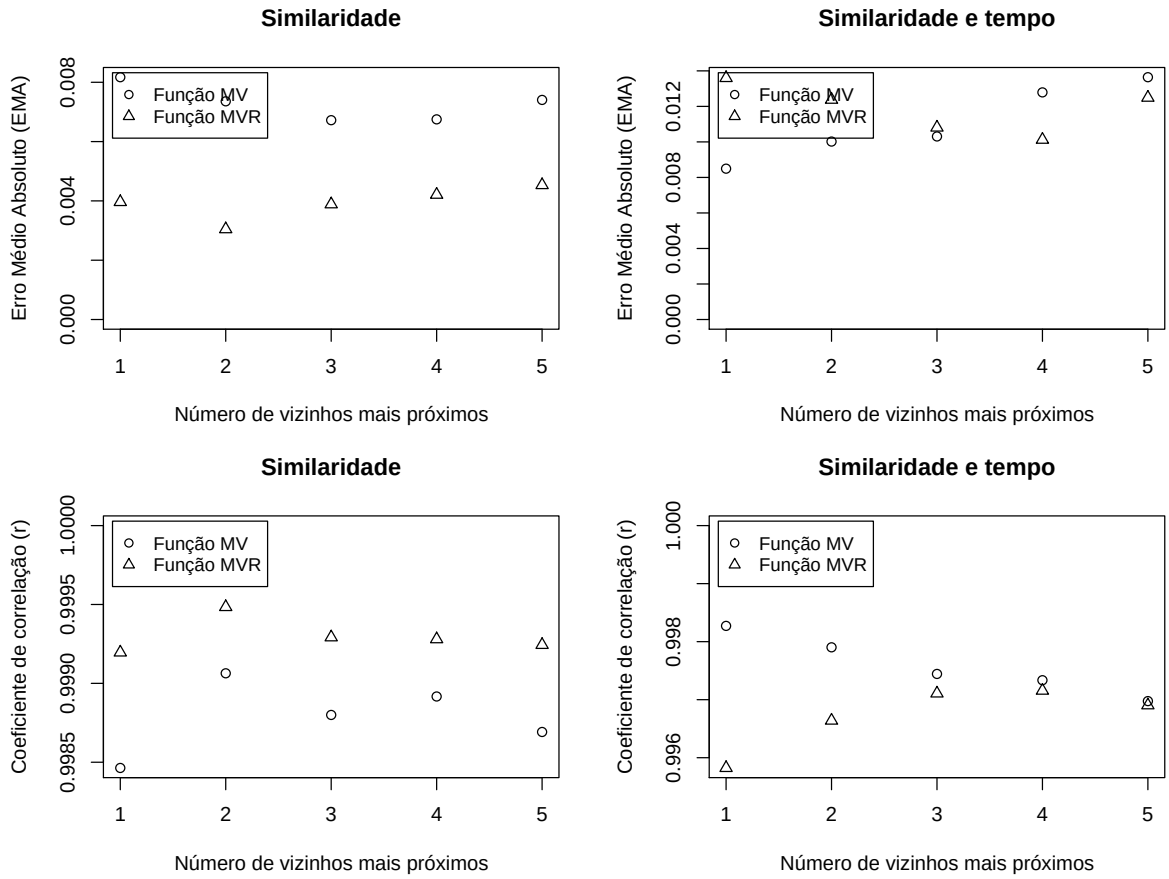


Figura 6.13: Gráficos referentes às medidas de EMA e r de previsão da série de Mackey-Glass: f_{MV} *versus* f_{MVR}

Tabela 6.4: Resumo das comparações entre as funções f_{MV} e f_{MVR}

		Similaridade			**Similaridade e Tempo**		
i	**Série Temporal**	EMA	r	Melhor desempenho	EMA	r	Melhor desempenho
1	Dependência sazonal	✓	✓	f_{MVR}	✓	✓	f_{MVR}
2	Sazonalidade multiplicativa	✓		f_{MVR}			
3	Alta frequência					✓	f_{MVR}
4	Lorenz	✓	✓	f_{MVR}	✓	✓	f_{MVR}
5	Mackey-Glass	✓	✓	f_{MVR}		✓	f_{MV}

multiplicativa houve **d.e.s** quando usado o critério de similaridade, em que f_{MVR} apresentou-se mais apropriada. Por outro lado, para a série temporal de alta frequência, a função f_{MVR} apresentou melhor desempenho quando usado o critério de similaridade e tempo para a seleção de vizinhos próximos.

Em relação à série temporal de Lorenz, em ambos os critérios de seleção de vizinhos próximos, houve **d.e.s** entre o desempenho das funções f_{MV} e f_{MVR}, em que a segunda apresentou melhor desempenho. Os padrões encontrados pela função f_{MVR} permitiram descrever melhor o comportamento caótico. Em relação à série temporal Mackey-Glass, para o critério de seleção de vizinhos

próximos por similaridade, a função f_{MVR} apresentou melhor desempenho e, para o critério por similaridade e tempo, a função f_{MV} apresentou melhor desempenho.

Pode ser observado que em todos os experimentos realizados, os erros são relativamente baixos, o que pode ser constatado nos gráficos apresentados em (Ferrero, 2009) que mostram que os valores previstos em cada caso foram muito próximos aos valores observados.

6.5 Considerações Finais

Neste capítulo foi apresentada a avaliação experimental do desempenho do algoritmo k*NN-TSP* utilizando (1) similaridade e (2) similaridade e tempo, como critérios de seleção de vizinhos próximos; e as funções de previsão (1) média de valores e (2) média de valores relativos para realizar o cálculo do valor futuro. Essa avaliação foi realizada considerando cinco séries temporais artificiais, no intuito de avaliar, sob diversas características, as variações do algoritmo de previsão k*NN-TSP*. As comparações realizadas em relação a ambos os critérios de seleção de vizinhos próximos mostraram que, de modo geral, o critério de similaridade apresentou melhor desempenho e, em relação as funções de previsão, a função f_{MVR} apresentou, na maioria das vezes, melhor desempenho que f_{MV}.

De acordo com os resultados apresentados pode ser constatado que a eficiência do algoritmo de previsão depende da característica de cada série temporal. Porém, devido às séries temporais artificiais utilizadas apresentarem padrões similares em diferentes níveis, de modo geral, a função f_{MVR} apresentou melhor desempenho em relação à função f_{MV}. Especificamente as séries de dependência sazonal, sazonalidade multiplicativa e alta frequência apresentam padrões ao longo de tendências, motivando a aplicação da função f_{MVR} no intuito de reproduzir esses padrões mais adequadamente. As séries temporais de Lorenz e Mackey-Glass não apresentam padrões ao longo de uma linha de tendência específica, porém, as sequências similares foram encontradas, em geral, em diferentes níveis, o que torna viável a aplicação da função f_{MVR}. Por outro lado, em relação ao critério de seleção de vizinhos próximos, o de similaridade apresentou melhor desempenho, possivelmente devido ao fato do critério de similaridade e tempo não considerar as melhores sequências para realizar previsões, além de incrementar a seleção de sequências superpostas, o que poderia influenciar negativamente no desempenho da previsão.

CAPÍTULO

7

Estudo de Caso

7.1 Considerações Iniciais

Foi realizado um estudo de caso da aplicação do algoritmo k*NN-TSP* a dados reais. Esses dados referem-se a informações ambientais, especificamente limnológicas, do reservatório da Hidrelétrica Itaipu Binacional. A limnologia é a ciência que estuda as águas continentais e essas informações referem-se a variáveis físicas, químicas e biológicas (fitoplâncton e zooplâncton) a respeito dessas águas (Esteves, 1998). No estudo de caso foram realizadas previsões de variáveis ambientais coletadas em uma estação próxima à barragem. Neste capítulo é apresentado o monitoramento ambiental na usina Hidrelétrica de Itaipu, as variáveis envolvidas nesse monitoramento, as etapas envolvidas neste estudo de caso, o desenvolvimento dessas etapas e os resultados experimentais.

7.2 Usina Hidrelétrica de Itaipu: Monitoramento Ambiental

A Usina de Itaipu é resultado de um trabalho conjunto entre Brasil e Paraguai, os quais, na década de 70, assinaram o tratado de Itaipu, dando inicio à construção da usina hidrelétrica. Atualmente, é considerada a maior em produção energética do mundo, responsável por 95% do abastecimento do Paraguai e 20% de toda a energia consumida no Brasil. A Itaipu é um marco da

engenharia moderna devido às suas dimensões, considerada em 1995 pela revista *Popular Mechanics* como uma das sete maiores obras da engenharia moderna. Pelas suas características, constitui fonte de geração de conhecimento e pesquisa nos temas referentes à construção, à manutenção e à segurança de barragens.

A avaliação de riscos em uma barragem deve ser capaz de identificar problemas e recomendar soluções, tais como estratégias corretivas e operacionais (Pan e He, 2000). Para auxiliar nesse processo de avaliação de riscos, é necessário que sejam realizadas coletas de dados através de monitoramentos frequentes, com o objetivo de manter a integridade de todas as áreas relacionadas à barragem. Como os dados são temporais, uma solução adequada consiste em representar o problema por meio de séries temporais e analisá-las utilizando técnicas apropriadas para esse tipo de dados.

O monitoramento ambiental é definido como o conjunto de dados físicos, químicos e biológicos de um ecossistema em estudo, que permite obter informações a respeito da qualidade das águas, um tema de importância para a segurança de barragens (Ribeiro Filho, R. A., 2006). O monitoramento consiste de repetidas observações, medidas, registros ambientais e parâmetros operacionais em um período de tempo. A Itaipu Binacional possui uma equipe de especialistas e uma rede de monitoramento de qualidade de água que está distribuída entre o reservatório e os seus afluentes.

Os dados coletados pela equipe envolvem a medição de variáveis físicas, químicas e biológicas da água e do ar. As amostras de água são coletadas na superfície e em diferentes profundidades. Um total de 12 (doze) estações de coleta foram estrategicamente distribuídas entre o reservatório e os afluentes, as quais estão descritas na Tabela 7.1. Na Figura 7.1 é apresentado o mapa com a localização de cada estação de coleta.

7.3 Descrição do Conjunto de Dados

Os dados coletados pelo Instituto Ambiental do Paraná (IAP) foram cedidos pela Superintendência de Meio Ambiente, Divisão de Reservatórios da Itaipu Binacional e pelo Centro de Estudo Avançados em Segurança de Barragens — CEASB — FPTI-BR, que além de ceder os dados, participaram ativamente das reuniões de trabalho. Dentre as doze estações, os especialistas consideraram que os dados coletados na estação E5 seriam mais apropriados para a previsão neste estudo inicial, por ser uma estação localizada no corpo central do Reservatório da Itaipu, a 15 Km a montante da barragem. Assim, foi decidido considerar as informações limnológicas coletadas em superfície na estação E5,

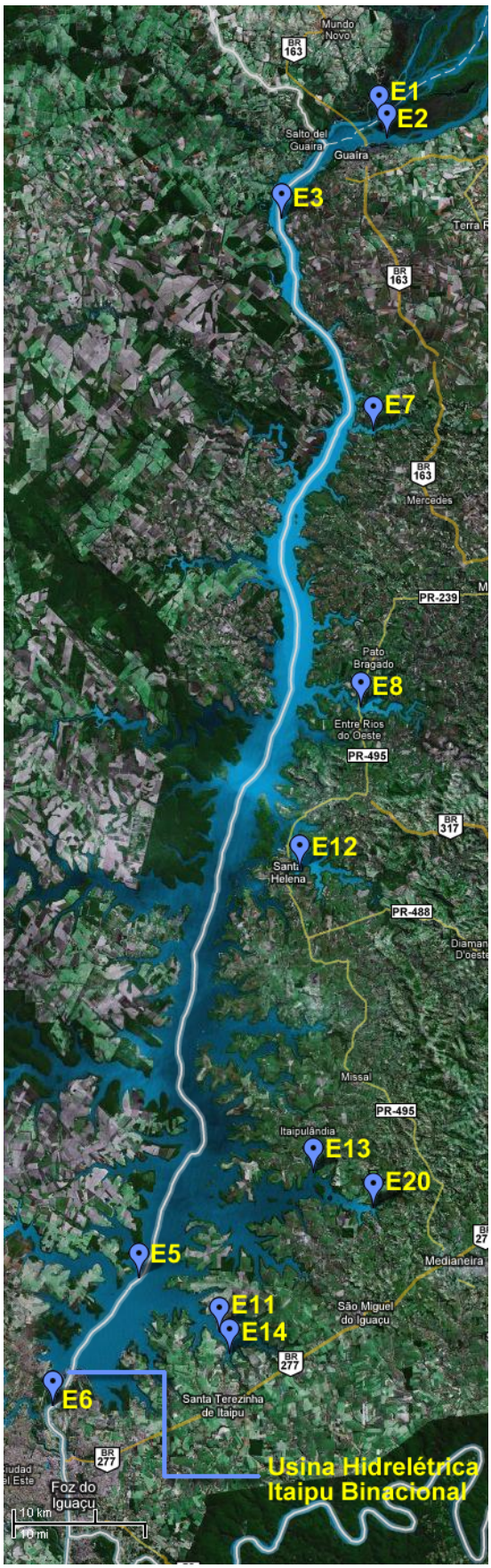


Figura 7.1: Mapa indicando as doze estações de coleta. Fonte: Itaipu Binacional

Tabela 7.1: Estações de coleta das amostras e suas respectivas localizações

Código da estação	Corpo da água	Localização
E1	Rio Paraná	Canal direito do Rio Paraná, a montante do Reservatório de Itaipu, em Guairá
E2	Rio Paraná	Canal esquerdo do Rio Paraná, a montante do Reservatório de Itaipu, em Guairá
E3	Reservatório de Itaipu, corpo central	Corpo central do Reservatório, próximo a Oliveira Castro
E5	Reservatório de Itaipu, corpo central	Corpo central do Reservatório de Itaipu, 15 Km a montante da barragem
E6	Rio Paraná	Rio Paraná, quatro Km a jusante da Barragem
E7	Reservatório de Itaipu, braço da margem esquerda	Braço formado pelo Arroio Guaçu.
E8	Reservatório de Itaipu, braço da margem esquerda	Braço formado pelo Rio São Francisco Verdadeiro
E11	Reservatório de Itaipu, braço da margem esquerda	Braço formado pelo Rio Passo Cuê
E12	Reservatório de Itaipu, braço da margem esquerda	Braço formado pelo Rio São Francisco Falso
E13	Reservatório de Itaipu, braço da margem esquerda	Braço formado pelo Rio Ocoí
E14	Reservatório de Itaipu, braço da margem esquerda	Braço formado pelo Rio Passo Cuê
E20	Reservatório de Itaipu, braço da margem esquerda	Braço formado pelo Rio Ocoí

medidas pelas variáveis apresentadas na Tabela 7.2 juntamente com as unidades dessas medidas, onde DQO refere-se à demanda química de oxigênio, e DBO_5 à demanda bioquímica de oxigênio (5 dias).

7.4 Desenvolvimento do Estudo de Caso

Os objetivos do trabalho consistem em, primeiramente, verificar a capacidade do *kNN-TSP* de prever valores futuros de variáveis ambientais sobre diferentes configurações de parâmetros e, posteriormente, verificar a capacidade do algoritmo de prever dados que mantenham a propriedade de correlação entre as variáveis. O estudo foi desenvolvido por meio das seguintes quatro etapas:

1. Formatação dos dados;
2. Seleção dos dados;
3. Configuração dos parâmetros do algoritmo *kNN-TSP*; e
4. Avaliação das previsões.

Tabela 7.2: Variáveis coletadas na estação E5 com suas respectivas unidades de medida

Variável	**Unidade**
Temperatura da água	°C
Temperatura do ar	°C
Turbidez	NTU[1]
Transparência da água	m
Alcalinidade total	mg/L
Nitrato	mg/L
Nitrito	mg/L
Nitrogênio amoniacal	mg/L
Nitrogênio Kjeldahl	mg/L
Sólidos suspensos	mg/L
DQO	mg/L
DBO_5	mg/L
Dureza total	mg/L
Saturação de oxigênio	%
Fósforo total	mg/L
pH	unidade
Oxigênio dissolvido	mg/L
Condutividade	μS/cm
Sólidos totais	mg/L

Na primeira etapa, os dados contidos na base de dados original, relacionados à estação E5, foram transformados para o formato adequado para a análise computacional. Posteriormente, foi realizada a seleção das variáveis ambientais a serem usadas para previsão. Na terceira etapa, foi aplicado o algoritmo *kNN-TSP* nos dados selecionados. Na quarta etapa, as previsões foram avaliadas de acordo com o objetivo do trabalho. Cada etapa é descrita nas seções subsequentes.

7.5 Etapa 1 — Formatação dos Dados

Esta etapa demandou muito tempo para ser realizada, devido, entre outros, à organização dos dados na base de dados original[2] ser diferente da requerida para a aplicação do algoritmo de previsão. Após várias reuniões com os especialistas, as informações referentes ao monitoramento ambiental da estação E5, registradas no Sistema de Informações Ambientais — SIA —, sistema utilizado para o registro das informações limnológicas das doze estações foram extraídas para um arquivo sequencial, cuja estrutura é ilustrada na Tabela 7.3.

Essas informações, armazenadas de modo sequencial, dificultam a aplicação de métodos de análise de dados para a extração de informações relevantes. Assim, foram implementados algoritmos na linguagem Perl[3] para

[2]Essa base armazena 30 MB de informações a respeito do monitoramento ambiental no reservatório e seus afluentes.

[3]http://www.perl.com/.

Tabela 7.3: Estrutura do arquivo sequencial de dados

Data de coleta	Estação	Variável	Valor
10/05/1985	E5	temperatura da água	$23,5$
10/05/1985	E5	temperatura do ar	$21,0$
		⋮	
10/05/1985	E5	condutividade	47
24/07/1985	E5	temperatura da água	$18,2$
24/07/1985	E5	temperatura do ar	$17,5$
		⋮	
24/07/1985	E5	condutividade	48
⋮			
24/02/2003	E5	temperatura da água	$28,8$
24/02/2003	E5	temperatura do ar	$28,0$
		⋮	
24/02/2003	E5	condutividade	49

transformar automaticamente os dados armazenados nesse formato para o formato atributo-valor — Tabela 4.1, página 28. Na Tabela 7.4 é apresentada a estrutura do arquivo no formato atributo-valor gerado a partir do arquivo descrito na Tabela 7.3.

Tabela 7.4: Estrutura do arquivo de dados no formato atributo-valor

Data de coleta	Estação	temperatura da água	temperatura do ar	...	condutividade
10/05/1985	E5	$23,5$	$21,0$		47
24/07/1985	E5	$18,2$	$17,5$		48
⋮					
24/02/2003	E5	$28,8$	$28,0$		49

7.6 Etapa 2 — Seleção dos Dados

A partir dos dados armazenados no formato atributo-valor, nesta etapa foram selecionadas, conjuntamente com os especialistas, as informações a serem utilizadas para a previsão de dados temporais. Foram definidos o período de coleta e as variáveis ambientais, com o objetivo de selecionar o conjunto de variáveis mais representativas para este estudo, descritos a seguir:

Período de coleta: em muitas estações, a frequência das coletas tem sido alterada de acordo com as necessidades de um melhor monitoramento. Entretanto, especificamente na estação E5, a frequência trimestral foi mantida desde o inicio de 1994 até o final de 2004. Assim, o período de

coleta selecionado corresponde ao compreendido entre o primeiro trimestre de 1994 e o último trimestre de 2004. Durante esses 11 anos foram realizadas 44 coletas trimestrais.

Variáveis ambientais: na estação selecionada foram coletadas 19 variáveis físico-químicas, no período entre 1994 a 2004. Os registros dessas variáveis apresentaram dois problemas, os quais estão relacionados a valores faltantes e registro de medidas. O primeiro problema afeta diretamente a análise das variáveis, *i.e.*, das séries temporais que cada variável representa. Ainda que vários métodos tenham sido propostos para contornar este problema (Batista e Monard, 2003), foi decidido desconsiderar as variáveis com valores faltantes, a fim de utilizar somente os valores verdadeiros dessas variáveis (e não valores estimados), de modo a não incluir mais um fator que pudesse influenciar o foco deste trabalho.

O segundo problema está relacionado a valores de variáveis nos quais aparece o símbolo menor (<). Esse símbolo indica que o valor coletado é menor que o mínimo valor (*threshold*) que pode ser mensurado pelo instrumento de coleta. As variáveis com esse tipo de dados também foram desconsideradas.

Na Tabela 7.5 são apresentadas as 19 variáveis registradas na estação E5. As doze variáveis que não apresentam nenhum dos problemas mencionados estão sinalizadas com o símbolo •.

Tabela 7.5: Seleção de variáveis físico-químicas

	Variável	**Valores faltantes**	**Presença de (<)**
•	Temperatura da água	não	não
•	Temperatura do ar	não	não
	Turbidez	sim	não
•	Transparência da água	não	não
•	Alcalinidade total	não	não
•	Nitrato	não	não
	Nitrito	sim	sim
	Nitrogênio amoniacal	não	sim
•	Nitrogênio Kjeldahl	não	não
•	Sólidos suspensos	não	não
	DQO	não	sim
	DBO_5	não	sim
	Dureza total	sim	não
•	Saturação de oxigênio	não	não
•	Fósforo total	não	não
•	pH	não	não
•	Oxigênio dissolvido	não	não
•	Condutividade	não	não
	Sólidos totais	sim	não

Um dos interesses da área é verificar, para variáveis correlacionadas, se a correlação, entre os valores observados de duas das séries, se mantém para os valores previstos dessas séries. A fim de encontrar os pares de variáveis mais fortemente correlacionados, foi construída uma matriz de correlação com as doze variáveis previamente selecionadas. Na Tabela 7.6 são apresentados os dez pares de variáveis que apresentaram maior coeficiente de correlação de Spearman (r).

Tabela 7.6: Dez maiores coeficientes de correlação entre pares de variáveis

	Par de variáveis	Coef. de correlação (r)
$\star$	(temperatura da água, temperatura do ar)	$+0,76$
$\star$	(temperatura da água, oxigênio dissolvido)	$-0,74$
	(saturação de oxigênio, oxigênio dissolvido)	$+0,60$
	(transparência da água, sólidos suspensos)	$-0,50$
	(temperatura do ar, oxigênio dissolvido)	$-0,44$
	(oxigênio dissolvido, pH)	$+0,43$
	(temperatura da água, transparência da água)	$-0,42$
	(nitrogênio Kjeldahl, pH)	$+0,41$
	(saturação de oxigênio, pH)	$+0,39$
	(sólidos suspensos, nitrato)	$+0,37$

Os pares de variáveis com r próximo ao limiar de correlação forte foram selecionados e indicados pelo símbolo $\star$ na Tabela 7.6. Assim, as séries temporais que foram utilizadas para previsão referem-se às variáveis:

- Temperatura da água;
- Temperatura do ar; e
- Oxigênio dissolvido.

Na Figura 7.2, são apresentados os gráficos referentes às séries temporais das três variáveis selecionadas.

7.7 Etapa 3 — Configuração dos Parâmetros do Algoritmo kNN-TSP

Uma vez definidas as variáveis limnológicas utilizadas no trabalho, o algoritmo de previsão *kNN-TSP*, foi aplicado sobre essas séries temporais, no intuito de verificar a eficiência dos métodos de previsão, de acordo com diferentes configurações do algoritmo. A metodologia apresentada na Seção 5.2, para avaliação de métodos de previsão de dados temporais, foi utilizada da mesma maneira que no caso dos dados artificiais — Capítulo 6. Devido ao

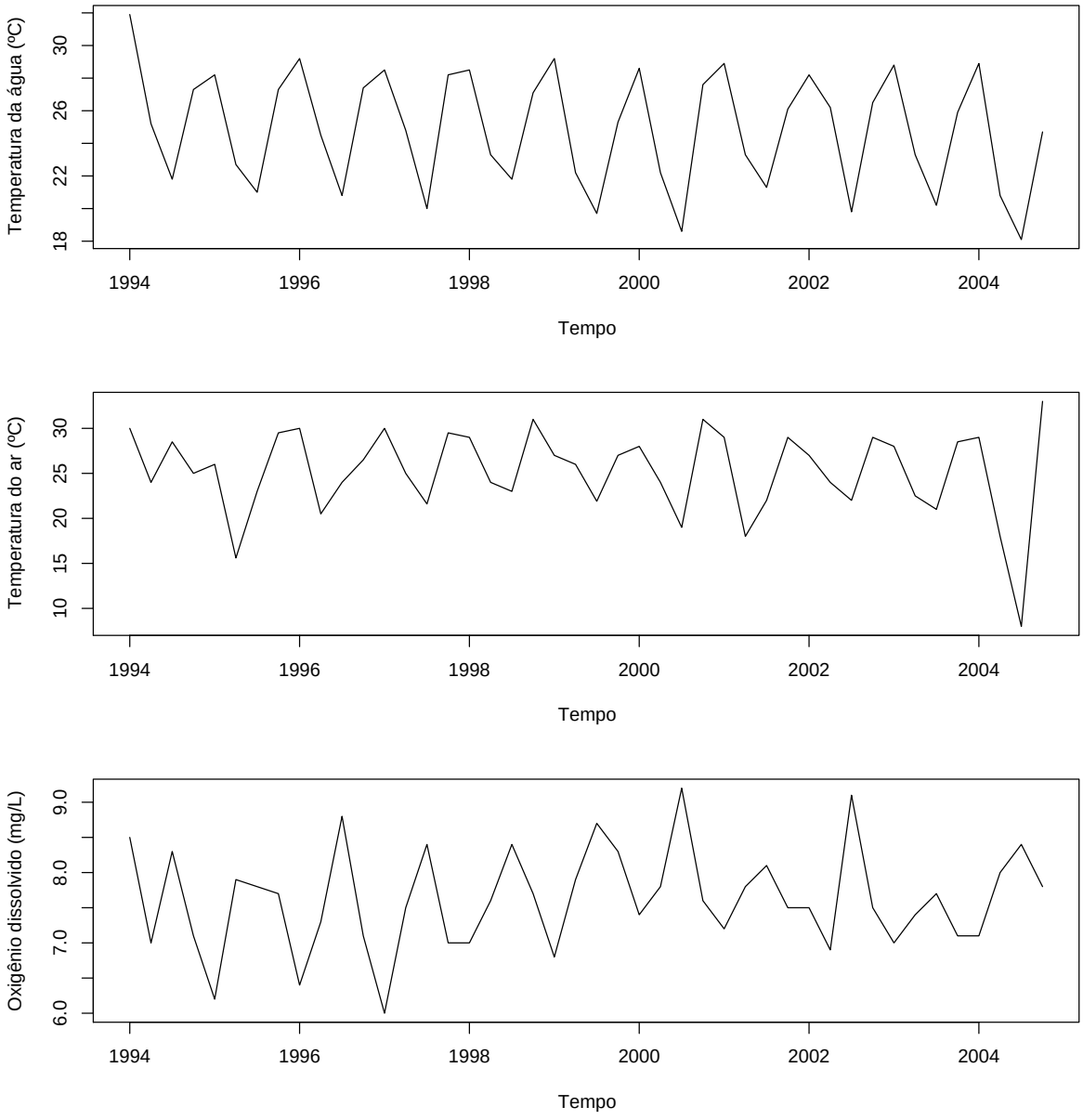


Figura 7.2: Séries temporais das variáveis limnológicas selecionadas: temperatura da água, temperatura do ar e oxigênio dissolvido

reduzido volume de informações de cada série temporal, foi definida a previsão dos últimos dez valores da série temporal, de acordo com a metodologia mencionada.

Quanto aos valores dos parâmetros — Seção 4.4, página 33 — o (a) tamanho da janela w para extração de subsequências foi definido $w = 4$, que corresponde às medidas realizadas em um ano (ciclo completo). Quanto ao parâmetro (b), conjunto de exemplos de treinamento, a totalidade dos exemplos de treinamento foi utilizada em cada previsão. Com relação aos parâmetros (c), (d) e (e) foram definidos os mesmos valores utilizados na avaliação experimental utilizando séries artificiais.

7.8 Etapa 4 — Avaliação da Previsão: Resultados e Discussão

A avaliação do desempenho do algoritmo k*NN-TSP* sobre as diferentes configurações foi realizada de duas maneiras, de acordo com os objetivos do estudo de caso. Primeiramente foi realizada a análise da precisão de cada variável individualmente, de acordo com as medidas de avaliação de erro médio absoluto e coeficiente de correlação. Posteriormente, foi avaliada a capacidade dos métodos de previsão manterem a propriedade de correlação entre os pares de variáveis selecionados na Etapa 2, quando consideramos os dez valores previstos dessas variáveis.

7.8.1 Análise de Precisão

A seguir é apresentada a análise das previsões realizas sobre as séries temporais de temperatura da água, temperatura do ar e oxigênio dissolvido. Para isso, cada configuração do algoritmo k*NN-TSP* foi avaliada de acordo com as medidas erro médio absoluto, EMA, e coeficiente de correlação de Spearman, r. As comparações entre a capacidade de previsão das funções f_{MV} e f_{MVR} utilizando o critério de similaridade e de similaridade e tempo para determinar os k vizinhos mais próximos, são realizadas, como anteriormente, por meio do teste não-paramétrico Wilcoxon para dados emparelhados, considerando nível de confiança de 95%.

A seguir, para cada uma das três variáveis selecionadas, é apresentada uma tabela que mostra o EMA e o coeficiente de correlação (r) dos valores previstos da série para cada configuração.

Temperatura da água: os resultados são mostrados na Tabela 7.7, junto com

o desvio-padrão. Pode ser observado que em todos os casos o valor do EMA é menor que $2,0$ ºC e a correlação entre os dez últimos valores observados da série e os valores previstos é alta. Entre os critérios de seleção de vizinhos próximos foi possível constatar que:

- Para a função f_{MV} **não** houve **d.e.s** em relação ao EMA e ao r;
- Para a função f_{MVR} **não** houve **d.e.s** em relação ao EMA e ao r.

Entre as funções de previsão foi possível constatar que:

- Para o critério de similaridade **não** houve **d.e.s** em relação ao EMA e ao r;
- Para o critério de similaridade e tempo **não** houve **d.e.s** em relação ao EMA e ao r.

Tabela 7.7: Resultado da previsão da série de temperatura da água

Medida de desempenho		Similaridade		Similaridade e tempo	
	k	f_{MV}	f_{MVR}	f_{MV}	f_{MVR}
EMA (ºC)	1	$1,5 \pm 1,6$	$1,2 \pm 0,6$	$1,8 \pm 1,6$	$1,4 \pm 1,1$
	2	$1,0 \pm 1,0$	$1,4 \pm 1,1$	$1,1 \pm 1,0$	$1,3 \pm 1,4$
	3	$1,0 \pm 1,0$	$1,2 \pm 0,9$	$1,2 \pm 0,9$	$1,1 \pm 1,2$
	4	$1,1 \pm 0,9$	$1,2 \pm 1,1$	$1,0 \pm 1,0$	$1,0 \pm 1,1$
	5	$1,1 \pm 1,0$	$1,1 \pm 1,2$	$0,9 \pm 0,9$	$1,1 \pm 1,1$
Coef. de correlação r	1	$+0,92$	$+0,95$	$+0,80$	$+0,87$
	2	$+0,95$	$+0,90$	$+0,97$	$+0,93$
	3	$+0,97$	$+0,98$	$+0,99$	$+0,98$
	4	$+0,98$	$+0,98$	$+0,96$	$+0,98$
	5	$+0,97$	$+0,98$	$+0,98$	$+0,96$

Na Figura 7.3 é apresentada a série temporal de temperatura da água e os valores previstos pelo algoritmo, considerando a configuração que apresentou o menor EMA, dada por $k = 5$, a função f_{MV} e o critério de similaridade e tempo para determinar os k vizinhos mais próximos, onde pode ser observado que os dez valores previstos foram similares aos valores observados.

Temperatura do ar: os resultados são mostrados na Tabela 7.8, junto com o desvio-padrão. Pode ser observado que em todos os casos o valor do EMA é menor ou igual a $5,0$ ºC, apontando um alto desvio-padrão. Quanto à correlação entre os dez valores observados da série e os dez valores previstos, é possível observar uma correlação fraca para f_{MVR}. Quanto a f_{MV}, em geral, a correlação é forte quando é usado o critério de similaridade para determinar os k vizinhos mais próximos. Entretanto, usando o

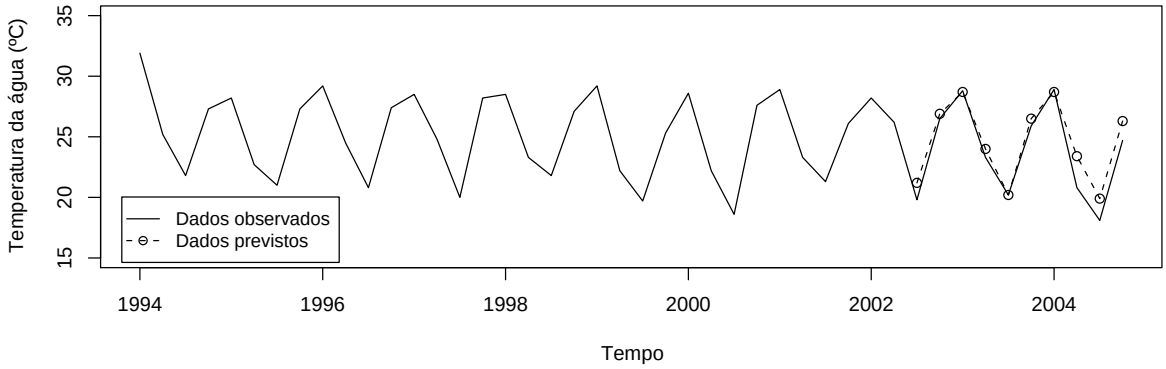


Figura 7.3: Série temporal de previsão da temperatura da água

critério de similaridade e tempo, é possível observar uma variação muito acentuada para $k = 2$. Entre os critérios de seleção de vizinhos próximos foi possível constatar que:

- Para a função f_{MV} **não** houve **d.e.s** em relação ao EMA e ao r;
- Para a função f_{MVR} **não** houve **d.e.s** em relação ao EMA e ao r.

Entre as funções de previsão foi possível constatar que:

- Para o critério de similaridade houve **d.e.s** em relação ao EMA e ao r, em que f_{MV} teve melhor desempenho;
- Para o critério de similaridade e tempo houve **d.e.s** em relação ao EMA, em que f_{MV} teve melhor desempenho. Em relação ao r **não** houve **d.e.s**;

Tabela 7.8: Resultado da previsão da série de temperatura do ar

Medida de desempenho		Similaridade		Similaridade e tempo	
	k	f_{MV}	f_{MVR}	f_{MV}	f_{MVR}
EMA (ºC)	1	$3,6 \pm 4,7$	$4,9 \pm 6,0$	$3,2 \pm 4,1$	$4,2 \pm 5,4$
	2	$3,4 \pm 4,6$	$4,2 \pm 5,7$	$3,7 \pm 5,6$	$4,4 \pm 5,8$
	3	$2,9 \pm 4,4$	$4,2 \pm 5,3$	$3,5 \pm 4,9$	$5,0 \pm 7,3$
	4	$3,1 \pm 4,3$	$4,1 \pm 5,5$	$3,1 \pm 4,5$	$4,5 \pm 7,0$
	5	$3,0 \pm 4,2$	$4,0 \pm 5,5$	$3,1 \pm 4,7$	$4,1 \pm 6,5$
Coef. de correlação r	1	$+0,73$	$+0,27$	$+0,79$	$+0,35$
	2	$+0,73$	$+0,30$	$+0,33$	$+0,35$
	3	$+0,73$	$+0,34$	$+0,65$	$+0,34$
	4	$+0,70$	$+0,28$	$+0,80$	$+0,33$
	5	$+0,80$	$+0,35$	$+0,69$	$+0,31$

Na Figura 7.4 é apresentada a série temporal de temperatura da água e valores previstos, considerando a configuração que apresentou o menor

EMA, dada por $k = 3$, a função f_{MV} e o critério de similaridade para determinar os k vizinhos mais próximos. Pode ser observado que os dois últimos valores medidos dessa série apresentam um comportamento fora do padrão, pelo registro de uma temperatura muito baixa (mínima temperatura na série) seguido de uma temperatura muito alta (máxima temperatura da série), o que dificulta a previsão desses dois últimos valores. Assim, é importante observar que ainda que f_{MV} apresente em geral melhores resultados que f_{MVR}, isso não permite afirmar que a previsão realizada é boa para os dez valores previstos. Isso pode também ser observado nos resultados da Tabela 7.8 levando em conta o valor do EMA e o alto desvio-padrão dos resultados.

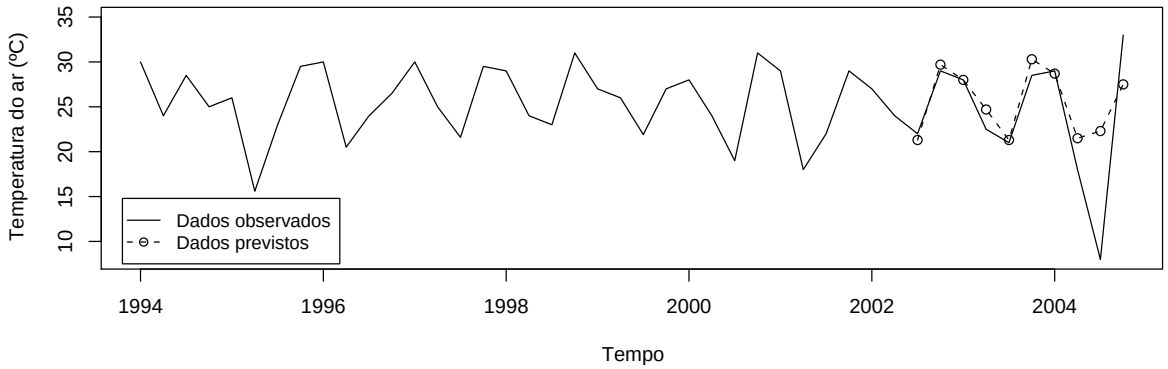


Figura 7.4: Série temporal de previsão da temperatura do ar

Série temporal de oxigênio dissolvido: os resultados são mostrados na Tabela 7.9. Pode ser observado que em todos os casos o valor do EMA é menor ou igual a $1,0$ mg/L, e o valor do coeficiente de correlação é fraco, e assumiu sinal negativo em algumas experiências. Entre os critérios de seleção de vizinhos próximos foi possível constatar que:

- Para a função f_{MV} houve **d.e.s** em relação ao EMA e ao r, em que o critério de similaridade teve melhor desempenho;
- Para a função f_{MVR} houve **d.e.s** em relação ao EMA e ao r, em que o critério de similaridade teve melhor desempenho.

Entre as funções de previsão foi possível constatar que:

- Para o critério de similaridade houve **d.e.s** em relação ao r, em que f_{MV} teve melhor desempenho. Em relação ao EMA **não** houve **d.e.s**;
- Para o critério de similaridade e tempo houve **d.e.s** em relação ao EMA, em que f_{MV} teve melhor desempenho. Em relação ao r **não** houve **d.e.s**;

Tabela 7.9: Resultado da previsão da série de oxigênio dissolvido

Medida desempenho		Similaridade		Similaridade e tempo	
	k	f_{MV}	f_{MVR}	f_{MV}	f_{MVR}
EMA (mg/L)	1	$0,6 \pm 0,6$	$0,4 \pm 0,6$	$0,7 \pm 0,6$	$0,7 \pm 0,9$
	2	$0,6 \pm 0,6$	$0,6 \pm 0,6$	$0,7 \pm 0,6$	$1,0 \pm 0,6$
	3	$0,5 \pm 0,5$	$0,5 \pm 0,6$	$0,6 \pm 0,6$	$0,8 \pm 0,6$
	4	$0,6 \pm 0,5$	$0,6 \pm 0,5$	$0,7 \pm 0,6$	$0,8 \pm 0,6$
	5	$0,5 \pm 0,5$	$0,5 \pm 0,3$	$0,6 \pm 0,6$	$0,7 \pm 0,6$
Coef. de correlação r	1	$-0,03$	$+0,46$	$-0,28$	$-0,09$
	2	$-0,07$	$+0,26$	$-0,27$	$-0,14$
	3	$+0,17$	$+0,30$	$-0,08$	$-0,02$
	4	$+0,12$	$+0,35$	$-0,57$	$-0,10$
	5	$+0,30$	$+0,36$	$-0,13$	$+0,01$

Para ilustrar, na Figura 7.5 é apresentada a série temporal de oxigênio dissolvido e os valores previstos pelo algoritmo, considerando a configuração que apresentou o menor EMA, dada por $k = 1$, a função f_{MVR} e o critério de similaridade para determinar os k vizinhos mais próximos. Esta série temporal apresenta alguns problemas semelhantes à série anterior. Como pode ser observado, a previsão de três dos dez valores deixar a desejar. Mas neste caso, isso seria facilmente detectado pelos baixos e variados valores do coeficiente de correlação obtidos — Tabela 7.9.

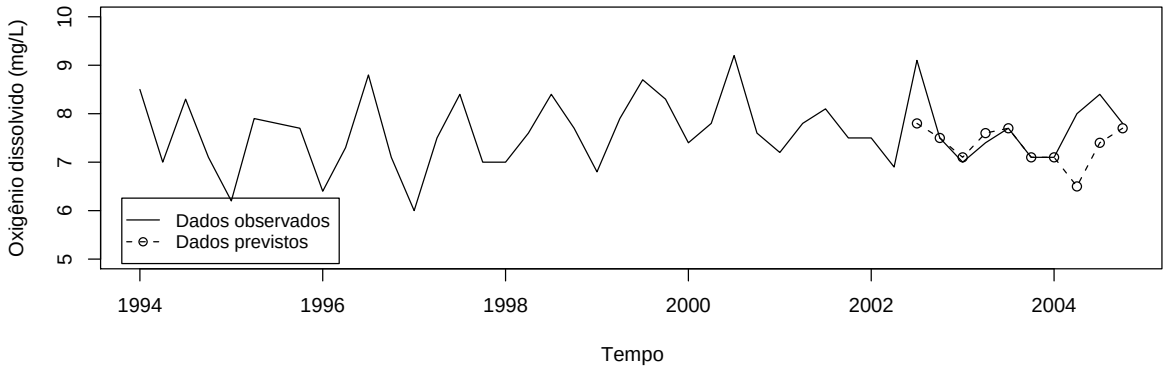


Figura 7.5: Série temporal de previsão do oxigênio dissolvido

7.8.2 Análise de Correlação entre Variáveis

Como mencionado, há interesse dos especialistas em verificar como varia a correlação entre os dez valores previstos para cada par de variáveis considerado, com relação aos dez valores observados dessas variáveis. A seguir são apresentados os coeficientes de correlação obtidos para cada um dos dois pares de variáveis.

Temperatura da água e temperatura do ar: na Figura 7.6 é apresentada a correlação entre os dez valores previstos, para as distintas configurações de k*NN-TSP*. A linha contínua representa a correlação entre os dez valores observados ($r = 0,77$).

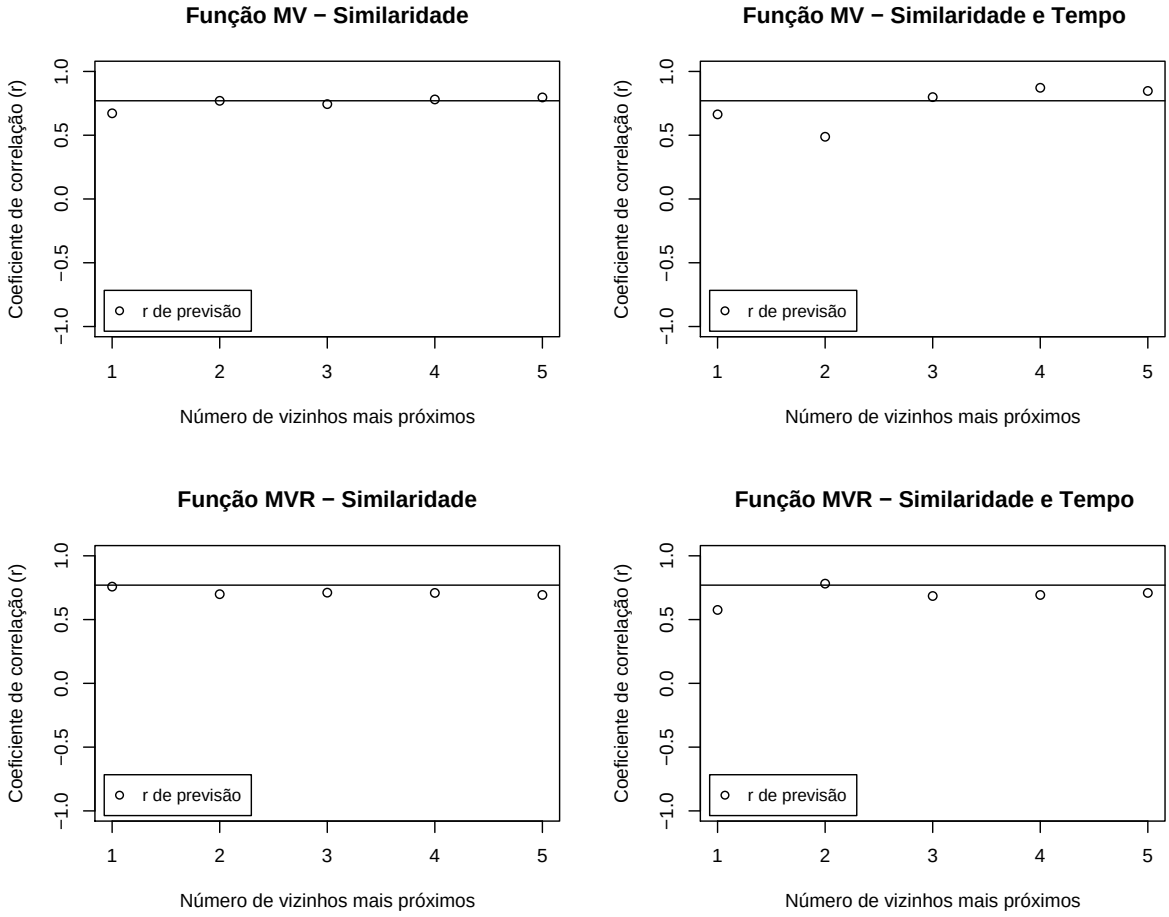


Figura 7.6: Correlação entre as dez previsões das variáveis temperatura da água e temperatura do ar. A correlação entre os correspondentes valores observados é $r = 0,77$

Pode ser observado nesses gráficos que as previsões realizadas com essas duas variáveis apresentam coeficientes de correlação similares ao obtido com os dados observados, para ambas as funções de previsão f_{MV} e f_{MVR}, utilizando o critério de similaridade para determinar os vizinhos mais próximos. Quando o critério é similaridade e tempo, esse resultado não é observado para todos os valores de k.

Temperatura da água e oxigênio dissolvido: na Figura 7.7 é apresentada a correlação entre os dez valores previstos, para as distintas configurações. A linha contínua representa a correlação entre os dez valores observados ($r = -81$).

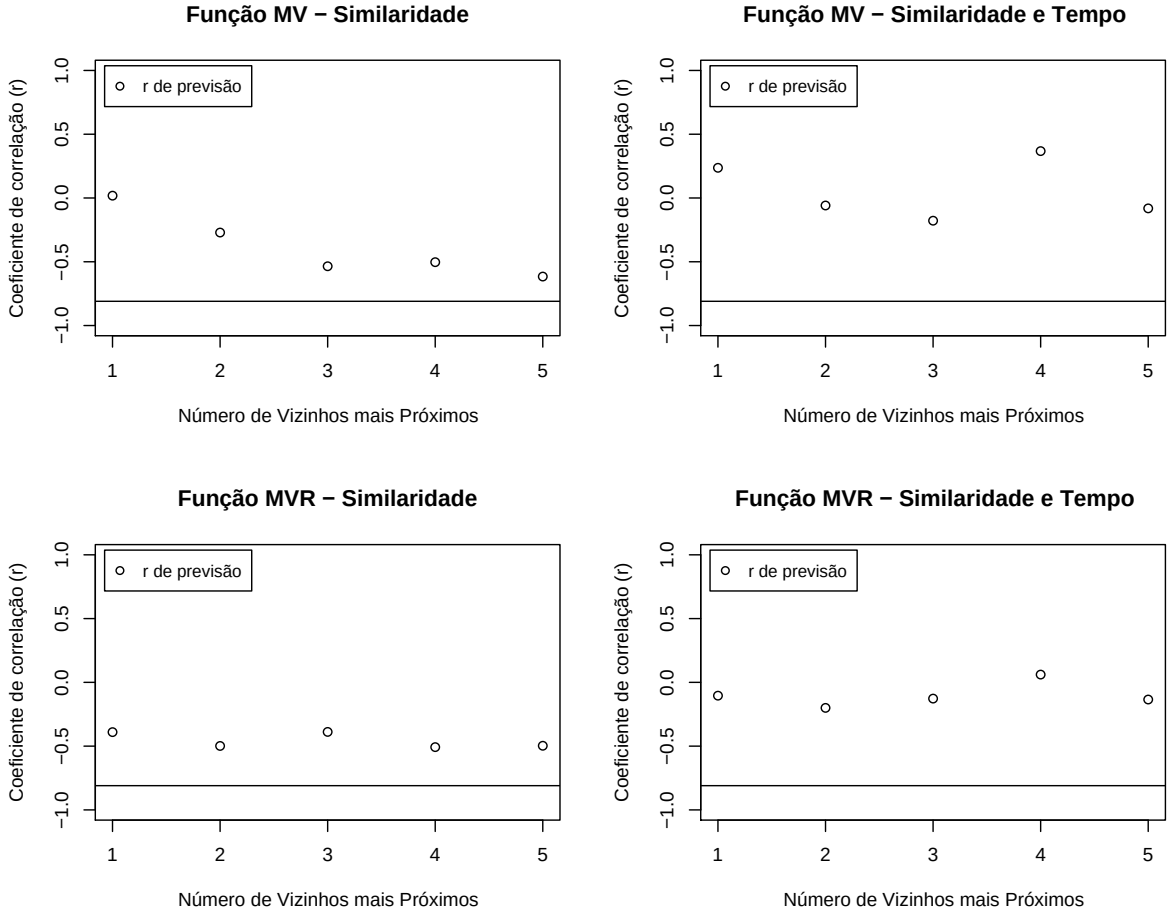


Figura 7.7: Correlação entre as dez previsões das variáveis temperatura da água e oxigênio dissolvido. A correlação entre os correspondentes valores observados é $r = -0,81$

Nos gráficos pode ser observado que o critério de similaridade e tempo para determinar os k vizinhos mais próximos apresentou valores de correlação distantes de -0.81.

7.9 Considerações Finais

Neste capítulo foi apresentada a aplicação do algoritmo *kNN-TSP* em séries temporais de variáveis ambientais. O estudo de caso permitiu identificar quais configurações do algoritmo apresentaram melhor desempenho para cada variável. Além disso, foi verificada a importância da correlação como medida de qualidade de previsão quando o objetivo consiste em prever valores de variáveis que mantenham a propriedade de correlação entre as mesmas.

CAPÍTULO

8

Conclusão

Neste trabalho foi apresentado o algoritmo *kNN-TSP* e contextualizado como uma adaptação do algoritmo de aprendizado de máquina *kNN* — Capítulo 4. Assim, várias questões relacionadas ao problema de aprendizado de máquina foram abordadas para o problema de previsão de dados. O algoritmo *kNN-TSP* apresenta cinco parâmetros: tamanho da janela para extrair subsequências; conjunto de exemplos de treinamento; medida de similaridade; cardinalidade do conjunto de séries similares e função de previsão. Em relação à medida de similaridade para a seleção de vizinhos próximos e à função de previsão, foram propostas abordagens alternativas para contornar dois problemas presentes na abordagem básica do algoritmo *kNN-TSP*.

O primeiro problema está relacionado à seleção de vizinhos próximos, a qual, na abordagem convencional, não leva em consideração a distância temporal em que as sequências mais similares se apresentam, em relação ao valor a ser previsto. É intuitivo pensar que, quanto mais perto do valor a ser previsto estejam as sequências similares, mais informação essas sequências podem conter a respeito do valor a ser previsto. Para contornar essa questão foi proposto um critério muito simples que leva em conta, não apenas a similaridade entre o formato das subsequências, mas também a distância temporal das subsequências em relação ao valor a ser previsto, de modo que sejam selecionadas, preferencialmente, aquelas subsequências mais próximas e mais similares. Esse critério de seleção de vizinhos próximos foi denominado de similaridade e tempo.

O segundo problema está relacionado à tendência presente em grande parte

dos problemas de previsão de dados temporais. A presença de tendência pode posicionar padrões em diferentes níveis ao longo de uma série temporal. A abordagem convencional do método *kNN-TSP* não leva em consideração essa questão, utilizando apenas os padrões mais similares localizados no mesmo nível para estimar o valor futuro. Para isso, foi proposta uma função de previsão que permite, além de encontrar padrões nos diversos níveis, prever o valor futuro levando em consideração o último valor observado. Essa função de previsão foi denominada função de média de valores relativos, ou f_{MVR}.

Essas variações do algoritmo *kNN-TSP* foram avaliadas experimentalmente com cinco conjuntos de séries temporais artificiais — Capítulo 6 — para verificar o desempenho do algoritmo, considerando diferentes configurações, em um ambiente controlado. Os resultados mostraram que, em três séries temporais, o critério de similaridade apresentou melhor desempenho que o critério de similaridade e tempo proposto, enquanto para as outras duas não foi possível evidenciar diferença estatisticamente significativa entre o desempenho de ambos os critérios. Isso mostrou que, embora a ideia de selecionar subsequências mais similares e mais próximas temporalmente seja intuitiva, o critério proposto não consegue realizar bem essa seleção. Quanto à funções de previsão f_{MV} e f_{MVR}, para quatro das séries temporais a função f_{MVR} proposta apresentou melhor desempenho, sendo que, em duas delas, a significância desse resultado foi dependente do critério de seleção de vizinhos mais próximos utilizado. Entretanto, em uma das séries temporais, quando utilizado o critério de similaridade e tempo, a função f_{MV} teve melhor desempenho, e quando utilizado o critério de similaridade, a função f_{MVR} apresentou melhor desempenho. Embora o desempenho da função de previsão possa depender da natureza da série temporal, em geral, a função f_{MVR} adaptou-se melhor que a função f_{MV} para as séries temporais artificiais utilizadas.

Para verificar o desempenho do algoritmo *kNN-TSP* em séries temporais reais, foi realizado um estudo de caso considerando observações de variáveis ambientais (temperatura da água, temperatura do ar e oxigênio dissolvido), coletadas no corpo central do reservatório da Hidrelétrica Itaipu Binacional no período de 1994 a 2004. O estudo foi avaliado, primeiramente, pela previsão individual de cada variável e, posteriormente, pela capacidade do algoritmo de manter a correlação entre pares de variáveis correlacionadas.

Na primeira avaliação, entre o desempenho dos critérios de seleção de vizinhos próximos, somente para a variável oxigênio dissolvido, o critério de similaridade apresentou melhor desempenho. Entre o desempenho das funções de previsão, para a variável temperatura da água, a função f_{MV} apresentou melhor desempenho, enquanto para a variável oxigênio dissolvido, a função

f_{MVR} teve melhor desempenho quando utilizado o critério de similaridade. Na segunda avaliação, foi constatado que a correlação entre os valores previstos das variáveis temperatura da água e temperatura do ar, foi similar à dos dados observados. Por outro lado, os valores previstos para as variáveis temperatura da água e oxigênio dissolvido não apresentaram correlação similar à dos dados observados. Em geral, o desempenho do algoritmo para as variáveis temperatura da água e temperatura do ar, individualmente, foi melhor, em relação ao desempenho do algoritmo para a previsão da variável oxigênio dissolvido.

Considerando os resultados da avaliação experimental do algoritmo *kNN-TSP* usando séries temporais artificiais e séries temporais de variáveis ambientais, podemos concluir que:

1. embora o critério de similaridade e tempo para a seleção de vizinhos próximos não apresentou uma melhoria significativa em relação ao critério de similaridade, em muitos casos mostrou desempenho tão bom quanto esse critério, além de valorizar os padrões mais recentes;

2. a função f_{MVR} proposta neste trabalho mostrou, em vários casos, melhor desempenho em relação à função de previsão f_{MV}, apresentando-se como uma boa alternativa para previsão de dados temporais, principalmente, na presença de tendência nos dados; e

3. o algoritmo *kNN-TSP*, nas suas diferentes configurações, apresentou-se adequado para a previsão de dados artificiais e reais, porém, não existe uma única configuração que mostre os melhores resultados.

8.1 Principais Contribuições

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

- Proposta de uma abordagem de seleção de vizinhos próximos para séries temporais que combina a questão de similaridade com distância temporal, de modo a considerar as sequências mais similares e mais recentes;

- Proposta de uma função de previsão para o algoritmo *kNN-TSP* que permite utilizar padrões contidos em níveis de tendência diferentes, para prever valores futuros e atenuar a influência da tendência, tanto na identificação de padrões, quanto na previsão do valor futuro; e

- Desenvolvimento de um conjunto de ferramentas computacionais que incluem a implementação do algoritmo *kNN-TSP* para várias configurações

e os módulos de: experimentos; extração de medidas de avaliação; comparação de resultados; geração de gráficos de resultados e análise de correlação entre pares de variáveis observadas e previstas.

Deve ser observado que as abordagens tradicionais relacionadas à escolha da medida de similaridade para o algoritmo *kNN-TSP* não levam em consideração a distância temporal em que as subsequências mais similares se encontram. Como mencionado no Capítulo 3, alguns métodos estatísticos, como o método de previsão exponencial simples, atribuem pesos aos últimos valores da série temporal em ordem de importância, por exemplo, considerando os mais recentes mais importantes. A ideia de utilizar a distância temporal difere da atribuição de pesos, pelo simples fato que a valorização dos padrões mais recentes ocorre somente na etapa de seleção de vizinhos próximos, sendo que, posteriormente, para o cálculo do valor futuro, qualquer função de previsão pode ser utilizada.

As funções de previsão consideram, comumente, a atribuição de pesos aos vizinhos próximos para o cálculo do valor futuro. Kulesh et al. (2008) propuseram recentemente a utilização de funções de previsão, com o ajuste de parâmetros, para adaptar-se a diferentes níveis de tendência, bem como à variação de amplitude ao longo do tempo. Nesse contexto, a função f_{MVR}, proposta neste trabalho, consiste em uma função que não necessita de parâmetros adicionais e, como mencionado, permite prever valores futuros na presença de tendência.

O algoritmo, e outras contribuições deste trabalho, estão implementadas dentro de um *framework* de um sistema computacional, denominado *TimeSSys*, para realizar diversas tarefas de interesse em séries temporais, como pré-processamento, previsão, agrupamento, recuperação de conteúdo, entre outras. Esse *framework* implementa conjuntos de ferramentas que auxiliam nas diversas etapas do processo de análise de dados. Como já mencionado, o presente trabalho contribui com a agregação de diversos módulos que auxiliam no desenvolvimento de varias etapas relacionadas a essas tarefas de interesse.

8.2 Limitações

Como mencionado, o trabalho apresentou as seguintes limitações:

- A seleção de vizinhos próximos permite que sejam selecionadas subsequências próximas que apresentam superposição, tanto para o critério de similaridade como para o de similaridade e tempo;

- A função de previsão f_{MVR} não é adequada para a previsão de dados que apresentem variação de amplitude ao longo do tempo; e
- O reduzido número de amostras utilizado no estudo de caso, referente a variáveis ambientais coletadas em períodos regulares, para a avaliação do algoritmo de previsão.

Os critérios de seleção de vizinhos próximos considerados no trabalho permitem a ocorrência de *trivial matches*, *i.e.*, seleção de subsequências similares as quais se sobrepõem. À medida que o tamanho da janela w aumenta, a possibilidade de que duas subsequências extraídas a partir de pontos adjacentes tenham valores de similaridade muito próximos, é alta. Desse modo, a utilização dessas subsequências para calcular o valor futuro pode estar influenciando negativamente no desempenho das funções de previsão.

Duas das séries temporais artificiais utilizadas na avaliação experimental apresentam a característica de variação de amplitude da série temporal ao longo do tempo. A função de previsão f_{MVR} não foi implementada no intuito de prever valores em séries temporais dessa natureza, embora tenha apresentado, dependendo da medida de avaliação, melhor desempenho em relação à função f_{MV}. Desse modo, o fato da amplitude da série crescer conforme o tempo, pode tornar difícil a previsão valores de maior amplitude com base, somente, em valores de menor amplitude.

No contexto de previsão de séries temporais, é desejável que uma grande quantidade de dados representem o comportamento da variável avaliada no tempo. Assim, a frequência de amostragens é um fator importante na representação de grande parte dos fenômenos. Como já mencionado, o estudo de caso foi realizado com dados provenientes do monitoramento ambiental e a frequência amostral desses dados foi trimestral. Considerando a variação possível de cada variável ao longo do tempo, os resultados foram considerados satisfatórios. Embora não seja possível confirmar experimentalmente, consideramos que com uma frequência menor de amostragem e um número maior de dados, os resultados poderiam ser ainda mais promissores.

8.3 Trabalhos Futuros

Ao longo do desenvolvimento do trabalho foram identificadas diversas questões que podem ser investigadas em trabalho futuros, no intuito de solucionar algumas das limitações mencionadas, bem como de investigar outros temas relevantes no contexto de previsão em séries temporais. Esses trabalhos futuros incluem:

- A seleção das sequências mais similares desconsiderando os *trivial matches*, tanto para o critério de similaridade como de similaridade e tempo;
- A utilização de um critério de dispersão de séries temporais, assim como uma medida de similaridade que leve em consideração essa questão, no intuito de adaptar a função f_{MVR} para a previsão em séries temporais com variação de amplitude. Uma abordagem poderia consistir na utilização do coeficiente de variação, que expressa o desvio-padrão como percentagem do valor da média de cada subsequência. Outra abordagem poderia ser o cálculo de medidas baseadas em dimensão fractal.
- A utilização do cálculo de média ponderada dos valores futuros das subsequências mais similares, no intuito de melhorar o desempenho da função f_{MVR};
- A aplicação de técnicas de seleção de exemplos para diminuir o conjunto de séries de treinamento e, assim, diminuir o custo computacional para procurar os vizinhos próximos; e
- A aplicação do algoritmo *kNN-TSP* para previsão de outras séries temporais reais, relacionadas ao tema de monitoramento ambiental e de outras áreas do conhecimento.

Referências Bibliográficas

Aggarwal, C. C., Hinneburg, A., e Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *Lecture Notes in Computer Science*, páginas 420–434. Springer. Citado na página 31.

Aha, D. W., Kibler, D., e Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning 6*, páginas 37–66. Citado nas páginas 2, 29 e 30.

Aitkenhead, M. J. e Cooper, R. J. (2008). Neural network time series prediction of environmental variables in a small upland headwater in NE Scotland. *Hydrological Processes*, 22:1–11. Citado nas páginas 25 e 27.

Alpaydin, E. (2004). *Introduction to Machine Learning*. MIT Press, Cambridge — MA, England. Citado na página 30.

Batista, G. E. A. P. A. e Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5):519–533. Citado na página 79.

Box, G. E. P., Jenkins, C. M., e Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. Prentice hall, Englewood — NJ —, USA, 3ª edição. Citado na página 23.

Bray, M. e Han, D. (2004). Identification of support vector machines for runoff modelling. *Journal of Hydroinformatics*, 6(4):265–280. Citado na página 27.

Broomhead, D. S. e King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica D Nonlinear Phenomena*, 20:217–236. Citado na página 38.

Camilleri, M. (2004). Forecasting using non-linear techniques in time series analysis: An overview of techniques and main issues. Relatório técnico, De-

partamento de Ciência da Computação e Inteligência Artificial da Universidade de Malta, Malta. Disponível em: http://www.cs.um.edu.mt/~csaw/CSAW04/Proceedings/02.pdf. Citado na página 20.

Chan, N. H. (2002). *Time Series: Applications to Finance*. John Wiley and Sons, New York — NY, USA. Citado na página 1.

Chapelle, O., Schölkopf, B., e Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA. Citado na página 28.

Chen, L. e Ng, R. T. (2004). On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, páginas 792–803. Citado na página 40.

Cherman, E. A., Lee, H. D., Maletzke, A. G., Ferrero, C. A., Fagundes, J. J., Coy, C. S. R., e Wu, F. C. (2008). Estudo da influência da redução de dimensionalidade em recuperação de conteúdo: Aplicação em dados temporais de exames de manometria ano-retal. In *III Congresso da Academia Trinacional de Ciências*, páginas 1–10, Foz do Iguaçu — PR, Brasil. Citado nas páginas 13 e 51.

Chu, S., Keogh, E. J., Hart, D., e Pazzani, M. J. (2002). Iterative deepening dynamic time warping for time series. In *Second SIAM International Conference on Data Mining*, páginas 1–18, Arlington, Virginia, USA. Citado na página 40.

Chun-Hua, B. e Xin-Bao, N. (2004). Determining the minimum embedding dimension of nonlinear time series based on prediction method. *Chinese Physics*, 13(5):633–636. Citado nas páginas 37 e 38.

Cortés, A. L. M. e Zimmermann, F. J. P. (2006). Análisis estadístico de series temporales. Relatório técnico, Universidad de La Sabana, Chía, Colombia. Disponível em: http://sabanet.unisabana.edu.co/admon/docencia/zimmerman/Series%20de%20tiempo.pdf. Citado nas páginas 9 e 12.

Daliakopoulosa, I. N., Coulibalya, P., e Tsanis, I. K. (2004). Groundwater level forecasting using artificial neural networks. *Journal of Hydrology*, 309(1–4):229–240. Citado na página 24.

Doria, U. (1999). *Introdução à Bioestatística: para Simples Mortais*. Elseiver, São Paulo — SP, Brasil. Citado na página 51.

Ehlers, R. S. (2005). Análise de séries temporais. Relatório técnico, Departamento de Estatística, Universidade Federal do Paraná, Curitiba — PR, Brasil. Disponível em: http://leg.est.ufpr.br/~ehlers/notas/stemp.pdf. Citado nas páginas 17 e 20.

Ereira Souto, D., Amaro Baldeon, R., e Leitao Russo, S. (1999). Estudo dos modelos exponenciais na previsão. *Industrial Data*, 9(1):97–103. Citado na página 17.

Espinosa, C., Parisi, F., e Parisi, A. (2005). Evidencia de comportamiento caótico en Índices bursátiles americanos. MPRA Paper 2794, University Library of Munich, Germany. Citado na página 58.

Esteves, F. (1998). *Fundamentos de Limnologia*. Interciência, Rio de Janeiro — RJ, Brasil, 2ª edição. Citado na página 73.

Everitt, B. S. (1993). *Cluster Analysis*. Edward Arnold, 3ª edição. Citado na página 31.

Fabris, F., Drago, I., e Varejão, F. M. (2008). A multi-measure nearest neighbor algorithm for time series classification. In *IBERAMIA '08: Proceedings of the 11th Ibero-American conference on AI*, páginas 153–162, Berlin, Heidelberg. Springer-Verlag. Citado na página 39.

Ferrero, C. A. (2009). Documento interno – *kNN-TSP* Resultados Experimentais. Disponível em: http://labic.icmc.usp.br/resultados/anfer/. Citado nas páginas 61 e 72.

Ferrero, C. A., Lee, H. D., Monard, M. C., Wu, F. C., Coy, C. S. R., Fagundes, J. J., e Góes, J. R. N. (2007). Aplicação de métodos de séries temporais para a identificação de seções em exames de manometria anorretal. In *II Congresso da Academia Trinacional de Ciências*, páginas 1–10, Foz do Iguaçu — PR, Brasil. Citado na página 51.

Ferrero, C. A., Monard, M. C., Lee, H. D., Benassi, S. F., e Wu, F. C. (2008). Previsão da temperatura da água no reservatório de Itaipu utilizando o método não-linear *k-Nearest Neighbor*. In *III Congresso da Academia Trinacional de Ciências*, páginas 1–10, Foz do Iguaçu — PR, Brasil. Citado na página 51.

Ferri, R., Alicata, F., Del Gracco, S., Elia, M., Musumeci, S., e Stefanini, M. (December 1996). Chaotic behavior of eeg slow-wave activity during sleep. *Electroencephalography and Clinical Neurophysiology*, 99:539–543(5). Citado na página 58.

Flores, B. E. (1989). The utilization of the wilcoxon test to compare forecasting methods: A note. *International Journal of Forecasting*, 5(4):529–535. Citado na página 51.

Frank, R. J., Davey, N., e Hunt, S. P. (2001). Time series prediction and neural networks. *J. Intell. Robotics Syst.*, 31(1-3):91–103. Citado na página 27.

Freedman, D., Pisani, R., e Purves, R. (1998). *Statistics*. Norton, New York — NY, USA, 3ª edição. Citado nas páginas 51 e 61.

Gandur, M. C. (1999). *Comportamento Dinâmico Complexo em Despelamento de Fitas Adesivas*. Tese de doutorado, Universidade Estadual de Campinas. Disponível em: http://biq.iqm.unicamp.br/arquivos/teses/ficha48910.htm. Citado nas páginas 56 e 58.

Grassberger, P. e Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1–2):189–208. Citado na página 38.

Guerra, J., Sánchez, G., e Reyes, B. (1997). Modelos de series de tiempo para predecir la inflación en Venezuela. Relatório técnico, Gerencia de Investigaciones Económicas, Banco Central de Venezuela. Disponível em: http://www.bcv.org.ve/Upload/Publicaciones/doc13.pdf. Citado na página 23.

Hadad, A., Evin, D., Drozdowicz, B., e Chiotti, O. (2007). Integración de modelos de seguimiento temporal en el monitoreo de pacientes críticos. *Simposio de Informática y Salud — SIS 2007 — 36 Jornadas Argentinas de Informática — JAIIO*, páginas 1–15. Citado na página 13.

Hirschberg, D. S. (1977). Algorithms for the longest common subsequence problem. *Journal of the Association for Computing Machinery*, 24(4):664–675. Citado na página 40.

Hyndman, R. J. e Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688. Citado na página 50.

Illa, J. M. G., Alonso, J. B., e Marré, M. S. (2004). Nearest-neighbours for time series. *Applied Intelligence*, 20(1):21–35. Citado nas páginas 2 e 39.

Jain, A. K. e Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. Citado na página 31.

Karunasinghe, D. S. K. e Liong, S.-Y. (2006). Chaotic time series prediction with a global model: Artificial neural network. *Journal of Hydrology*, 323(1-4):92–105. Citado nas páginas 2, 21, 22, 24, 27 e 43.

Kennel, M. B., Brown, R., e Abarbanel, H. D. I. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A*, 45:3403–3411. Citado na página 38.

Keogh, E. e Kasetty, S. (2002). On the need for time series data mining benchmarks: a survey and empirical demonstration. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, páginas 102–110, New York, USA. Citado nas páginas 39 e 40.

Kulesh, M., Holschneider, M., e Kurennaya, K. (2008). Adaptive metrics in the nearest neighbours method. *Physica D: Nonlinear Phenomena*, 237(3):283–291. Citado nas páginas 24, 34, 38, 39, 40, 56, 57, 59 e 92.

Lee, H. D. (2005). *Seleção de Atributos Importantes para a Extração de Conhecimento de Bases de Dados*. Tese de doutorado, ICMC-USP. Disponível em: http://www.teses.usp.br/teses/disponiveis/55/55134/tde-08032004-164855/. Citado na página 32.

Liu, H. e Motoda, H. (2007). *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC. Citado na página 32.

Mackey, M. e Glass, L. (1977). Oscillation and chaos in physiological control systems. *Science*, 197(4300):287–289. Citado na página 59.

McNames, J. (1998). A nearest trajectory strategy for time series prediction. In *Proceedings of the International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling*, páginas 112–128, Leuven, Belgium. K.U. Leuven. Citado nas páginas 2 e 24.

McNames, J. (1999). *Innovations in Local Modeling for Time Series Prediction*. Tese de doutorado, Stanford. Disponível em: www.ece.pdx.edu/~mcnames/Publications/Dissertation.pdf. Citado nas páginas 58 e 59.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York — NY, USA. Citado na página 27.

Mizuhara, Y., Hayashi, A., e Suematsu, N. (2006). Embedding of time series data by using dynamic time warping distances. *Syst. Comput. Japan*, 37(3):1–9. Citado na página 38.

Morettin, P. A. (2008). Econometria financeira: Um curso em análise de séries temporais financeiras. Relatório técnico, Departamento de Estatística, Universidade Federal do Paraná, São Paulo — SP, Brasil. Citado na página 17.

Morettin, P. A. e Toloi, C. M. (1989). *Modelos de Função de Transferência*. ABE, São Paulo — SP, Brasil. Citado na página 13.

Morettin, P. A. e Toloi, C. M. (2006). *Análise de Séries Temporais*. Edgard Blücher, São Paulo — SP, Brasil. Citado nas páginas 8 e 10.

Nygård, J. F. e Glattre, E. (2003). Fractal analysis of time series in epidemiology: Is there information hidden in the noise? *Norwegian Journal of Epidemiology*, 13(2):303–308. Citado na página 13.

Pan, J. e He, J. (2000). *Large Dams in China: a fifty–year review*. China WaterPower Press, Beijing, China. Citado na página 74.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Citado na página 52.

Ribeiro Filho, R. A. (2006). *Relações Tróficas e Limnológicas no Reservatório de Itaipu: uma Análise do Impacto da Biomassa Pesqueira nas Comunidades Planctônicas*. Tese de doutorado, EESC-USP. Disponível em: http://www.teses.usp.br/teses/disponiveis/18/18139/tde-22012007-154056/. Citado na página 74.

Sáfadi, T. (2004). Uso de séries temporais na análise de vazão de Água na represa de Furnas. *Ciência e Agrotecnologia, Universidade Federal de Lavras*, 28(1):142–148. Citado nas páginas 13 e 23.

Shiki, S. B., Lee, H. D., Burin, E. L. K., Niz, M. A. K., Coy, C. S. R., Fagundes, J. J., e Wu, F. C. (2008). Desenvolvimento de um sistema para a análise de curvas provenientes de exames de manometria ano-retal. In *III Congresso da Academia Trinacional de Ciências*, páginas 1–10, Foz do Iguaçu — PR, Brasil. Citado na página 13.

Shumway, R. H. e Stoffer, D. S. (2006). *Time Series Analysis and Its Applications with R Examples*. Springer, 2ª edição. Citado nas páginas 8, 9 e 13.

Sivakumar, B. (2000). Chaos theory in hydrology: important issues and interpretations. *Journal of Hydrology*, 227(1-4):1 – 20. Citado na página 58.

Slini, T., Karatzas, K., e Moussiopoulos, N. (2002). Statistical analysis of environmental data as the basis of forecasting: an air quality application. *The Science of the Total Environment*, 288:227–237. Citado na página 23.

Solomatine, D., Maskey, M., e Shrestha, D. (2006). Eager and lazy learning methods in the context of hydrologic forecasting. *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, 1(1):4847–4853. Citado na página 44.

Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., e Lendasse, A. (2007). Methodology for long-term prediction of time series. *Neurocomput.*, 70(16-18):2861–2869. Citado na página 15.

Spolaôr, N., Lee, H. D., Ferrero, C. A., Coy, C. S. R., Fagundes, J. J., e Wu, F. C. (2008). Um estudo da aplicação de clustering de séries temporais em dados médicos. In *III Congresso da Academia Trinacional de Ciências*, páginas 1–10, Foz do Iguaçu — PR, Brasil. Citado na página 51.

Tang, Z. e Fishwick, P. (1993). Feed-forward neural nets as models for time series forecasting. *ORSA Journal of Computing*, 5(91–008):374–386. Disponível em: http://citeseer.ist.psu.edu/tang93feedforward.html. Citado na página 24.

Todos os endereços eletrônicos foram acessados em: 5 de janeiro de 2009.

APÊNDICE

A

Resultados da Avaliação Experimental

As tabelas a seguir apresentam os $p - valores$ referentes à avaliação experimental — Capítulo 6. A Tabela A.1 apresenta os resultados da comparação entre os critérios de seleção de vizinhos próximos (1) de similaridade e (2) de similaridade e tempo. A Tabela A.2 apresenta os resultados da comparação entre as funções de previsão f_{MV} e f_{MVR}. Os $p-valores < 0,05$ estão em negrito.

Tabela A.1: Comparação entre os critérios similaridade e similaridade e tempo

		f_{MV}			f_{MVR}		
i	**Série Temporal**	EMA	r	Melhor desempenho	EMA	r	Melhor desempenho
1	Dependência sazonal	$1,0000$	$1,0000$		$\mathbf{0,0079}$	$0,2827$	similaridade
2	Sazonalidade multi.	$0,9166$	$1,0000$		$1,0000$	$1,0000$	
3	Alta frequência	$0,2222$	$0,4206$		$0,2073$	$0,1412$	
4	Lorenz	$\mathbf{0,0079}$	$\mathbf{0,0079}$	similaridade	$\mathbf{0,0079}$	$\mathbf{0,0119}$	similaridade
5	Mackey-Glass	$\mathbf{0,0079}$	$\mathbf{0,0079}$	similaridade	$\mathbf{0,0079}$	$\mathbf{0,0079}$	similaridade

Tabela A.2: Comparação entre as funções f_{MV} e f_{MVR}

		Similaridade			**Similaridade e Tempo**		
i	**Série Temporal**	EMA	r	Melhor desempenho	EMA	r	Melhor desempenho
1	Dependência sazonal	$\mathbf{0,0079}$	$\mathbf{0,0119}$	f_{MVR}	$\mathbf{0,0079}$	$\mathbf{0,0109}$	f_{MVR}
2	Sazonalidade multi.	$\mathbf{0,0317}$	$1,0000$	f_{MVR}	$0,0952$	$1,0000$	
3	Alta frequência	$0,0556$	$1,0000$		$0,0952$	$\mathbf{0,0159}$	f_{MVR}
4	Lorenz	$\mathbf{0,0079}$	$\mathbf{0,0079}$	f_{MVR}	$\mathbf{0,0079}$	$\mathbf{0,0119}$	f_{MVR}
5	Mackey-Glass	$\mathbf{0,0079}$	$\mathbf{0,0079}$	f_{MVR}	$0,6905$	$\mathbf{0,0317}$	f_{MV}

APÊNDICE

B

Resultados do Estudo de Caso

As tabelas a seguir apresentam os $p-valores$ referentes ao estudo de caso — Capítulo 7. A Tabela B.1 apresenta os resultados da comparação entre os critérios de seleção de vizinhos próximos (1) de similaridade e (2) de similaridade e tempo. A Tabela B.2 apresenta os resultados da comparação entre as funções de previsão f_{MV} e f_{MVR}. Os $p - valores < 0,05$ estão em negrito.

Tabela B.1: Comparação entre os critérios similaridade e similaridade e tempo

	f_{MV}			f_{MVR}		
Série Temporal	EMA	r	Melhor desempenho	EMA	r	Melhor desempenho
Temperatura da água	$0,8413$	$0,7503$		$0,4206$	$0,5959$	
Temperatura do ar	$0,3095$	$0,4633$		$0,4633$	$0,2073$	
Oxigênio dissolvido	$\mathbf{0,0269}$	$\mathbf{0,0079}$	similaridade	$\mathbf{0,0079}$	$\mathbf{0,0079}$	similaridade

Tabela B.2: Comparação entre as funções f_{MV} e f_{MVR}

	Similaridade			**Similaridade e Tempo**		
Série Temporal	EMA	r	Melhor desempenho	EMA	r	Melhor desempenho
Temperatura da água	$0,1508$	$0,8320$		$0,9166$	$0,7533$	
Temperatura do ar	$\mathbf{0,0079}$	$\mathbf{0,0079}$	f_{MV}	$\mathbf{0,0079}$	$0,0952$	f_{MV}
Oxigênio dissolvido	$0,6723$	$\mathbf{0,0317}$	f_{MVR}	$0,0593$	$0,0952$	