

WILLIAN ZALEWSKI

**INFLUÊNCIA DA INCERTEZA DE
MEDIÇÃO NA PREVISÃO DE
DADOS EM SEGURANÇA DE
BARRAGENS**

FLORIANÓPOLIS

2010

**UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO
EM METROLOGIA CIENTÍFICA E INDUSTRIAL**

**INFLUÊNCIA DA INCERTEZA DE MEDIÇÃO NA
PREVISÃO DE DADOS EM SEGURANÇA DE
BARRAGENS**

Dissertação submetida à
Universidade Federal de Santa Catarina
como parte dos requisitos para a
obtenção do grau de Mestre em Metrologia

WILLIAN ZALEWSKI

Florianópolis, Outubro de 2010

Catálogo na fonte pela Biblioteca Universitária
da
Universidade Federal de Santa Catarina

Z22i Zalewski, Willian
Influência da incerteza de medição na previsão de dados em
segurança de barragens [dissertação] / Willian Zalewski ;
orientador, Gustavo Daniel Donatelli. - Florianópolis, SC,
2010.

117 p.: il., grafs., tabs.

Dissertação (mestrado) - Universidade Federal de Santa
Catarina, Centro Tecnológico. Programa de Pós-Graduação em
Metrologia Científica e Industrial.

Inclui referências

1. Metrologia científica. 2. Medição. 3. Máquinas.
4. Dados - Previsão. 5. Monte Carlo, simulação de. 6. Análise
de séries temporais. 7. Represas. I. Donatelli, Gustavo
Daniel. II. Universidade Federal de Santa Catarina. Programa
de Pós-Graduação em Metrologia Científica e Industrial. III.
Título.

CDU 537.7

INFLUÊNCIA DA INCERTEZA DE MEDIÇÃO NA PREVISÃO DE DADOS EM SEGURANÇA DE BARRAGENS

WILLIAN ZALEWSKI

Esta Dissertação foi julgada adequada para obtenção do Título de Mestre em Metrologia, Área de concentração Gestão da Medição, e aprovada em sua forma final pelo Programa de Pós Graduação em Metrologia Científica e Industrial da Universidade Federal de Santa Catarina.

Gustavo Daniel Donatelli, Doutor
Orientador

Armando Albertazzi Gonçalves Jr., Doutor
Coordenador do Programa de Pós-Graduação em Metrologia Científica e
Industrial

Banca Examinadora:

Armando Albertazzi Gonçalves Jr., Dr.
Presidente

Mauricio de Campos Porath, Dr. -Ing.

Marcelo Ricardo Stemmer, Dr.

“A felicidade, em suma, obtém-se por meio da vida contemplativa, uma vida intelectual sossegada, longe das perturbações do cotidiano.”

Aristóteles

*Aos meus pais,
Sandra e José.*

*Aos educadores e amigos,
Huei Diana Lee e Wu Feng Chung*

*Aos amigos,
Renato Bobsin Machado, André Gustavo Ma-
letzke e Carlos Andres Ferrero, Maksoel Agus-
tin Krauspenhar Niz e Eduardo Lucas Konrad
Burin*

À Dabna Hellen Tomim

AGRADECIMENTOS

O desenvolvimento deste trabalho de dissertação compreendeu um período de fundamental aprendizado para minha vida. Muito além do aprendizado técnico, foi uma fase de significativas mudanças, na qual tive a oportunidade de vivenciar e experimentar peculiares circunstâncias da vida humana. Sobretudo, em inestimável importância foi a oportunidade de conhecer e conviver com boas pessoas, para as quais agradeço pela contribuição neste trabalho de dissertação e em minha vida.

Ao professor Gustavo Daniel Donatelli pela dedicação com a qual conduziu a orientação deste trabalho, pela compreensão em todas as etapas do desenvolvimento, em especial no período em que estive em Foz do Iguaçu. Pessoa pela qual sou muito grato pela confiança depositada e pela oportunidade de ter conhecido e compartilhado tal experiência. Admiro pela flexibilidade de trabalho e dedicação ao ensino e à pesquisa.

Aos professores Huei Diana Lee e Wu Feng Chung pela amizade e sempre atenção ao trabalho. Pessoas as quais tenho o imenso privilégio de conviver todos os dias. Verdadeiros educadores que dedicam-se à formação de pessoas justas e leais. Admiro pela dedicação à educação e à pesquisa e pelo esforço em sempre proceder de modo correto mesmo em meios contrários. Sou extremamente grato por todos os ensinamentos, para os quais levo sempre comigo e assim farei para toda a minha vida.

Aos professores Carlos Alberto Schneider, Gustavo Daniel Donatelli, Armando Albertazzi Gonçalves Júnior, Carlos Alberto Flesch e Analucia Vieira Fantin Pezzotta pelos ensinamentos e pela dedicação ao Programa de Pós-Graduação em Metrologia Científica e Industrial (PósMCI).

Aos professores Renato Bobsin Machado, André Gustavo Maletzke e Carlos Andres Ferrero pela amizade, pelo apoio e sempre presença e auxílio em todas os aspectos que envolveram a realização deste trabalho.

Aos professores Cláudio Saddy Rodrigues Coy, Juvenal Ricardo Navarro Góes (*in memoriam*), João José Fagundes e Maria Carolina Monard pela motivação para a educação e a pesquisa.

Aos colegas da turma 2008 do PósMCI, Arthur Nunes Silveira, Edineí Ballmann, Eduardo Lucas Konrad Burin, Humberto Leitão Riella, Iara Simoni Pereira, João Gabriel Fadel da Costa, Lúnia Coelho de Almeida de Lima, Maksoel Agustin Krauspenhar Niz e Vladimir Rodrigues de Lima pelas experiências e conhecimentos compartilhados e pela união da turma. Ao professor Carlos Alberto Schneider pelo incentivo à integração.

Aos colegas das turmas 2007, 2006 e 2005, Alexandre Marcondes,

Benjamim M. de Souza Neto, Francisco Augusto Arenhart, Juliana Marques Schöntag, Alexandre Victor Casella, Igor Gaviola, Mauro Eduardo Benedit, Maria Regina Kunzler, Miguel Burg Demay, Rafael Moreira Miggorin, Sidnei Rodrigo Basei, Kleber Steinbach e Sérgio A. Bitencourt Petrovic.

À Rosana pela dedicação ao correto e harmônico funcionamento do Labmetro e do PósMCI e pela sempre atenciosidade.

Ao Laboratório de Metrologia e Automatização (Labmetro) e à Fundação Certi pela oportunidade de conviver em um ambiente de pesquisa dinâmico e por possibilitar o desenvolvimento deste trabalho.

Ao Laboratório de Bioinformática (LABI) pelas oportunidades oferecidas para meu crescimento técnico e humanístico desde o início da minha vida acadêmica. Aos membros da “labilândia” Richardson Floriani Voltolini, Daniel de Faveri Honorato, Joylan Nunes Maciel, Luiz Henrique Dutra da Costa, Newton Spolaôr, Neimar Neitzel, Rafaella Aline Lopes da Silva, Bianca Espindola, Adrieli Cristina da Silva, Dabna Hellen Tomim, Jefferson Tales de Oliva, Ricardo Gil Belther Nabo, Antônio Rafael Sabino Parmezan, Simone Pinto Romero.

Aos meus familiares pelo constante apoio. Em especial aos meus pais, Sandra Zalewski e José U. de Oliveira, pela educação que me proporcionaram e pelo incentivo às minhas escolhas em todas as etapas de minha vida. Aos meus avós Lúcia Zalewski e Geremias Zalewski.

À Dabna Hellen Tomim pelo apoio, amizade e carinho e por me tornar uma pessoa mais feliz. À José Carlos Tomim, Fátima Minto Tomim e Leandro Tomim pelo também incentivo.

À Maksoel Agustin Krauspenhar Niz e Eduardo Lucas Konrad Burin pelas experiências compartilhadas, amizade e auxílio durante todo esse período.

À Francieli Françosi, Bianca Espindola, Gisele Cavanha Tomim, Giselle Colpani, Chris Mayara Tibes, Cecília Noro Pfeifer, Sydnei Bruce Shiki e Everton Alvares Chermam pelos momentos compartilhados

À Ricardo Luís Schaefer, Eduardo Francisco Maran Bueno, André Marcelo Trentini pelas experiências compartilhadas.

À Universidade Estadual do Oeste do Paraná, minha casa, pela oportunidade de retornar e poder contribuir para a formação de novos profissionais.

Ao PTI Ciência & Tecnologia (PTI C&T) da Fundação Parque Tecnológico de Itaipu (FPTI/BR) pelo auxílio por meio da linha de financiamento de bolsas. Também, ao Centro de Estudos Avançados em Segurança de Baragens (CEASB), em especial aos fundadores Professor Dr. Wu Feng Chung, Professora Dr. Huei Diana Lee e Professor Msc. Renato Bobsin Machado.

RESUMO

INFLUÊNCIA DA INCERTEZA DE MEDIÇÃO NA PREVISÃO DE DADOS EM SEGURANÇA DE BARRAGENS

Willian Zalewski

Outubro / 2010

Orientador: Gustavo Daniel Donatelli, Doutor.

Área de Concentração: Gestão da Medição.

Palavras-chave: Incerteza de Medição, Aprendizado de Máquina, Previsão de Dados, Simulação de Monte Carlo, Séries Temporais, Segurança de Barragens.

Número de Páginas: 117

O avanço tecnológico tem possibilitado a aquisição e o armazenamento de uma grande quantidade de dados em diversas áreas, dentre os quais, dados oriundos de medições realizadas continuamente ao longo do tempo. A representação e a análise de eventos e comportamentos no tempo é uma tarefa complexa e dependente do domínio de aplicação. Em segurança de barragens, a previsão de valores consiste em uma importante ferramenta de monitoração e prevenção pela capacidade de antecipar informações do processo de medição sob análise e com base nessas, realizar ações de contenção do problema. No entanto, os métodos tradicionais por se basearem na aplicação de técnicas estatísticas e na amostragem de dados, possuem limitações para detectar mudanças no processo. Desse modo, abordagens baseadas em aprendizado de máquina têm sido propostas para auxiliar nessa tarefa, tal como o algoritmo *k-Nearest Neighbor - Time Series Prediction* (*k*NN - TSP). Na monitoração da segurança de barragens, os dados das variáveis de interesse são obtidos por meio de processos de medição. Sua confiabilidade está afetada por diversas causas, que resultam no afastamento do resultado da medição com referência ao valor verdadeiro da variável medida, denominado de erro de medição. As técnicas de aprendizado de máquina aplicadas no apoio à tomada de decisões em segurança de barragens utilizam assim uma informação de entrada distorcida e, conseqüentemente, pode se esperar que as previsões geradas pelos algoritmos sejam também distorcidas. Para um desempenho adequado das ferramentas de apoio à tomada de decisão, é necessário que os dados sejam gerados por um processo de medição sob controle e capaz. Assim, torna-se fundamental o estudo

da influência do uso de dados incertos sobre as ferramentas de auxílio à tomada de decisão, no contexto de controle de processos de medição. Desse modo, neste trabalho foi proposto um modelo de simulação para avaliar o efeito da incerteza de medição, em função de distintas composições de erros sistemáticos e aleatórios, e da frequência de amostragem dos dados, sobre a efetividade das previsões do algoritmo k NN - TSP. Inicialmente, o modelo de simulação foi validado por meio da aplicação do método sobre uma série temporal artificial com características de interesse para a segurança de barragens, tais como tendência e sazonalidade. Posteriormente, o modelo foi aplicado em séries provenientes de um processo de monitoração real, como da medição de deslocamentos da Usina de Itaipu. Com base nos resultados das simulações realizadas, algumas orientações com foco em uma relação custo-benefício entre a qualidade de ajuste do algoritmo aos dados da série e a incerteza de medição puderam ser posicionadas, com o intuito de auxiliar na seleção de ferramentas de previsão de dados para análise de riscos em segurança de barragens.

ABSTRACT

MEASUREMENT UNCERTAINTY EFFECT IN SAFE DAMS DATA PREDICTION

Willian Zalewski

October / 2010

Advisor: Gustavo Daniel Donatelli, Doutor.

Area of Concentration: Management of Measurement.

Keywords: Measurement Uncertainty, Data Mining, Data Prediction, Monte Carlo Simulation, Time Series, Dams Safety.

Number of pages: 117

Technological progress has enabled the acquisition and storage of a large amount of data in several areas, such as data from measurements continuously made in the course of time. The representation and analysis of events and behaviors in the course of time is a complex task and depend on the application domain. In dams safety, data prediction is an important tool for monitoring and preventing due to the ability to anticipate information from the measurement process under analysis and, based on the information, perform actions to refrain the problem. However, traditional methods are based on applying statistical techniques and data sampling, have limitations on detecting changes in the process. Approaches based on machine learning have been proposed to assist this task, one of them is the *k-Nearest Neighbor - Time Series Prediction (kNN - TSP)* algorithm. In the dams safety monitoring, the data of variables of interest are obtained by measurement processes. The reliability is affected by various sources which cause the measurement result deviation using the measured variable real value as a reference, called measurement error. The machine learning techniques applied to help decision making on dams safety uses skewed input data, and the forecasts generated by algorithms are also distorted. For a suitable development of tools to support decision it is necessary that data is generated by a measurement process capable and under control. Thus, it becomes important to study the influence of using uncertain data on tools to assist decision making in the context of measurement processes control. This work proposed a simulation model to evaluate the effect of uncertainty measurement on different compositions of systematic and random errors and the sampling frequency of data, regarding the effectiveness of the predictions of the *kNN - TSP* algorithm. The simulation model was validated through method application on an artificial time

series presenting interesting characteristics for the dams safety, such as trend and seasonality. Later on, the model was applied in series obtained from a real monitoring process, as the displacement measurement monitoring of Itaipu Power Plant. Based on the results from the simulations, some guidelines focused on cost-effectiveness relation between algorithm fitness quality on the data series and uncertainty measurement could be suggested to assist on selecting tools for data forecast to risk analysis on dams safety.

SUMÁRIO

1	INTRODUÇÃO	27
1.1	Objetivos	28
1.1.1	Objetivo Geral	28
1.1.2	Objetivos Específicos	29
1.2	Organização do Trabalho	29
2	PREVISÃO DE DADOS EM SÉRIES TEMPORAIS.....	31
2.1	Fundamentos de Séries Temporais	31
2.1.1	Tendência	32
2.1.2	Sazonalidade	33
2.1.3	Resíduo	34
2.2	Modelagem de Séries Temporais para Previsão de Dados ..	34
2.3	Aprendizado de Máquina	35
2.4	Algoritmo k NN para Previsão de Séries Temporais	37
2.4.1	Conjunto de Séries de Treinamento	39
2.4.2	Medida de Similaridade	39
2.4.3	Conjunto de Vizinhos mais Próximos	40
2.4.4	Função de Previsão	40
2.5	Considerações Finais	41
3	INCERTEZA DE MEDIÇÃO EM SEGURANÇA DE BARRAGENS	43
3.1	Instrumentação para Segurança de Barragens	43
3.1.1	Avaliação de Riscos	46
3.2	Incerteza de Medição	46
3.2.1	Método Clássico ou da Propagação das Incertezas ..	50
3.2.2	Método Numérico ou da Propagação das Distribuições	53
3.3	Influência da Incerteza de Medição em Ferramentas de Análise de Séries Temporais	56
3.4	Considerações Finais	59
4	MÉTODO DE AVALIAÇÃO PROPOSTO.....	61
4.1	Modelo de Simulação	61
4.2	Modelo Matemático de Medição	62
4.2.1	Modelo para Distribuição do Erro Aleatório	62
4.2.2	Modelo para Distribuição do Erro Sistemático	63

4.2.3	Modelo para a Incerteza Padrão Combinada e Incerteza Expandida	63
4.3	Etapas do Método de Avaliação	64
4.4	Avaliação do Ambiente de Simulação	70
4.4.1	Séries Temporais Artificias	71
4.4.2	Configurações dos Experimentos	71
4.4.3	Resultados	75
4.4.4	Discussão	83
4.5	Considerações Finais	86
5	ESTUDO DE CASO	87
5.1	Usina Hidrelétrica de Itaipu - Pêndulos	87
5.1.1	Procedimento de Medição	89
5.1.2	Sistemas de Medição	90
5.1.3	Série Histórica das Medições	91
5.2	Aplicação do Método de Avaliação	93
5.3	Resultados e Discussão	96
5.4	Considerações Finais	103
6	CONCLUSÕES E TRABALHOS FUTUROS	105

LISTA DE FIGURAS

1	Exemplo de extração de tendência de uma série temporal . . .	33
2	Aplicação do algoritmo kNN	38
3	Exemplo de aplicação do algoritmo kNNTToTSP	38
4	Normalização de séries temporais	40
5	Função de previsão Média de Valores e MVR	41
6	Instrumentação de bloco-chave da Usina de Itaipu	45
7	Comparação entre o Método Clássico e o Método Numérico	54
8	Fluxograma simplificado da avaliação de incerteza usando SMC	55
9	Esquema geral do método	61
10	Etapas do método de avaliação	65
11	Simulação de mudança do erro sistemático	66
12	Fluxograma de geração de erros	67
13	Série de referência e uma série obtida por simulação dos erros de medição	68
14	Fluxograma de previsão em séries temporais	68
15	Dispersão dos valores previstos	69
16	Procedimento de análise dos valores previstos	70
17	Série Temporal de Dependência Sazonal	72
18	Contribuições aleatória e sistemática à incerteza para diferentes cenários	73
19	Frequências de amostragem analisadas	74
20	Exemplo de séries com erros para o cenário 2	75
21	Exemplo de séries com erros para o cenário 3	76
22	Exemplo de séries com erros para o cenário 4	76
23	Conjunto de valores previstos para o cenário 3	78
24	Exemplo de séries previstas para o cenário 3	78
25	Exemplo de séries previstas para o cenário 2	79
26	Valores previstos para o cenário 2	80
27	Valores previstos para o cenário 3	81
28	Valores previstos para o cenário 4	82
29	Incerteza de previsão - Cenário x Frequência de amostragem	84
30	Valores de EMA para cada frequência de amostragem	84
31	Representação esquemática de: a) pêndulo direto e b) pêndulo invertido	88
32	Sistemas de coordenadas para a medição de deslocamentos com pêndulos.	89
33	Coordinômetro Óptico (a) e Coordinômetro Digital (b) . . .	90

34	Perfil de deslocamentos relativos do bloco-chave F-19 da Usina de Itaipu	92
35	Série histórica de deslocamentos absolutos do bloco F-19 . . .	93
36	Série de dados temporais do SAI (2000-2010)	94
37	Frequências de amostragem diária, semanal e mensal dos dados do ADAS	95
38	Valores previstos para os dados do SAI	98
39	Valores de incerteza de previsão para o sistema manual . . .	99
40	Valores de EMA para o sistema manual	99
41	Valores previstos para os dados do ADAS	100
42	Valores de incerteza de previsão para o sistema automatizado	101
43	Valores de EMA para o sistema automatizado	101
44	Valores de incerteza de previsão para os sistemas manual e automatizado	102
45	Valores de EMA para os sistemas manual e automatizado . .	102

LISTA DE TABELAS

1	Cenários para a composição da incerteza de medição.	64
2	Valores dos parâmetros para a composição da incerteza de medição.	73
3	Quantidade de pontos (P) x Frequência de amostragem (F). . .	74
4	Valores de EMA x Frequência de amostragem dos dados. . .	83
5	Avaliação da incerteza de previsão - Cenários de erros (C) x Frequência de amostragem dos dados (F).	83
6	Frequências de medição tipicamente empregadas para a monitoração de pêndulos diretos e invertidos	91
7	Contribuições de erro sistemático e de erro aleatório.	94
8	Quantidade de pontos (P) x Frequência de amostragem (F) para os dados provenientes do ADAS.	96

LISTA DE ABREVIATURAS E SIGLAS

*k*NN - TSP, *k*-Nearest Neighbor - Time Series Prediction

UFSC, Universidade Federal de Santa Catarina

PósMCI, Programa de Pós-Graduação em Metrologia Científica e Industrial

CEASB, Centro de Estudos Avançados em Segurança de Barragens

GUM, Guia para a Expressão da Incerteza de Medição

ISO, International Organization for Standardization

BIPM, Bureau International des Poids et Mesures

FDP, Função Densidade de Probabilidade

SMC, Simulação de Monte Carlo

CEP, Controle Estatístico de Processos

EMA, Erro Médio Absoluto

SAI, Sistema de Acompanhamento da Instrumentação

ADAS, Sistema de Aquisição Automática de Dados

LISTA DE SÍMBOLOS

ST , Conjunto ordenado de valores que descreve a série temporal
 ST_t , Valor de uma observação da série temporal no instante t
 T_t , Valor da tendência da série temporal no instante t
 S_t , Valor da sazonalidade da série temporal no instante t
 R_t , Valor do resíduo da série temporal no instante t
 O_{t+1} , Valor futuro
 O_t , Observação em um dado instante t
 $f_{MVR}(S)$, Função de previsão MVR para uma dada série temporal S
 ε , Erro
 Y_v , Valor do mensurando
 Y_k , Valor medido em uma medição k
 \bar{Y} , Média de k valores medidos
 S^2_Y , Variância de k valores medidos
 ν_{eff} , Graus de liberdade efetivos
 Y , Valor medido
 X_f , Valor do mensurando em uma determinada frequência de amostragem f
 E_{ale} , Erro aleatório
 E_{sis} , Erro sistemático
 u_{ale} , Incerteza padrão para o erro aleatório
 u_{sis} , Incerteza padrão para o erro sistemático
 u_Y , Incerteza padrão combinada
 $U_{95\%}$, Incerteza expandida para intervalo de abrangência de 95%
 X_t , Valor da série de referência em um dado instante t
 μ_{S_i} , Média das previsões realizadas para o ponto i
 σ_{S_i} , Desvio padrão das previsões realizadas para o ponto i
 C , Cenários para a composição da incerteza de medição
 σ_C , Média dos desvios padrões obtidos em cada cenário de simulação
 $U_{95\%}(\sigma_C)$, Incerteza expandida para cada cenário de simulação
 AS , Amplitude pico-a-pico do sinal
 M_a , Medição do deslocamento absoluto
 $M_r(D)$, Medição do deslocamento relativo para pêndulo direto
 $M_r(I)$, Medição do deslocamento relativo para pêndulo invertido

1 INTRODUÇÃO

Pesquisas na área segurança de barragens tem se tornado de grande importância nos últimos anos devido ao elevado número de aproveitamentos hidrelétricos em todo o mundo. Na construção de uma barragem o maior risco de desastres está no período correspondente ao início da fase de operação da usina. No entanto os riscos de ocorrência de anomalias continuam presentes durante a vida útil das barragens devido a processos de deterioração dos materiais e de possíveis falhas de projeto. Para isso, a fim de garantir a segurança das estruturas, é empregada uma numerosa instrumentação para monitorar diversos parâmetros (SILVEIRA et al., 1994; ROCHA et al., 1992; PORTELA, 2003).

Na Usina Hidrelétrica de Itaipu, vistorias periódicas e monitoramentos são realizados desde o início da operação da usina com o intuito de gerar dados de alta relevância. Atualmente, as leituras são realizadas automaticamente ou de modo manual por meio de mais de 2380 instrumentos e mais de 5200 drenos distribuídos pela barragem. Como resultado do monitoramento da barragem por meio dos sensores e sistemas computacionais, a Itaipu dispõe hoje de uma grande quantidade de dados armazenados relacionados ao histórico das medições realizadas (ITAIPU BINACIONAL, 2008).

A avaliação de riscos de uma barragem tem por objetivo identificar problemas e recomendar reparos corretivos, restrições operacionais e/ou modificações, visando evitar problemas relacionados ao correto funcionamento de uma barragem (MEDEIROS, 2003; CBGB, 1996). Frequentemente as avaliações de segurança de barragens são baseadas em dados adquiridos por meio de medições e análises mecânicas, que são confrontados com conhecimento de especialistas e de análises estatísticas. No entanto, com o avanço tecnológico, diversos processos têm sido propostos para auxiliar na análise de grandes repositórios de dados, entre eles o de mineração de dados apoiado por métodos de inteligência computacional (ALPAYDIN, 2004; MITCHELL, 1997; REZENDE, 2003).

O conhecimento extraído e os padrões encontrados podem ser utilizados para a realização de simulações de desempenho futuro, que podem confirmar condições normais de operação ou indicar riscos potenciais para a barragem. Esses estudos poderão contribuir também para o desenvolvimento e verificação de futuros empreendimentos auxiliando na análise do desempenho de uma barragem tanto na etapa de construção quanto na fase de operação (MALETZKE, 2009).

Os dados provenientes da instrumentação da barragem ao longo do tempo podem ser entendidos como uma série temporal, a qual pode ser utili-

zada para análise de padrões e previsão de valores (MORETTIN; TOLOI, 2006; EHLERS, 2005).

Séries com um comportamento predominantemente aleatório podem ser analisadas por métodos estatísticos, tais como o controle estatístico de processos (CEP). No entanto, esses métodos não são apropriados quando as séries são autocorrelacionadas (e.g. séries apresentando tendência e/ou sazonalidade). Desse modo, abordagens propostas na literatura baseiam-se na utilização de técnicas de aprendizado de máquina para auxiliar nessa tarefa, tal como o algoritmo *k-Nearest Neighbor*, redes neuronais artificiais e algoritmos genéticos (FERRERO, 2009; SILVA, 2005; MATTOS et al., 2005; SOLOMATINE; MASKEY; SHRESTHA, 2006; CASTRO, 2001). De especial importância para este trabalho é a pesquisa de Ferrero (2009), na qual foi proposta uma adaptação do algoritmo *k-Nearest Neighbor* para a previsão de valores em séries temporais, denominado algoritmo *k-Nearest Neighbor - Time Series Prediction* (*k*NN - TSP). Esse estudo foi aplicado em dados ambientais relacionados à segurança de barragens.

No entanto, um aspecto importante a ser considerado na aplicação dessas técnicas é a utilização de dados contaminados por erros de medição, que podem afetar a capacidade dos algoritmos para realizar previsões válidas. Embora o conhecimento da influência dos erros de medição na eficácia dos algoritmos de previsão seja de grande relevância na hora de tomar decisões sobre a instrumentação, não foram realizadas ainda pesquisas que elucidem essas relações entre a qualidade das medições e a qualidade das previsões.

1.1 Objetivos

O tema deste trabalho consiste em uma demanda existente da Usina Hidrelétrica de Itaipu para análise de riscos com base no estudo dos dados históricos de instrumentação de barragens. Nesse sentido, o escopo deste trabalho foi definido conjuntamente com a Universidade Federal de Santa Catarina (UFSC) por meio do Programa de Pós-Graduação em Metrologia Científica e Industrial (PósMCI), e com o Centro de Estudos Avançados em Segurança de Barragens (CEASB).

1.1.1 Objetivo Geral

O objetivo desse trabalho consiste em avaliar a influência da incerteza de medição na efetividade do algoritmo de previsão *k*NN - TSP em dados de segurança de barragens.

1.1.2 *Objetivos Específicos*

- Elaborar um modelo de simulação que possibilite propagar a incerteza de medição através do algoritmo k NN - TSP;
- Desenvolver algoritmos para adaptar o k NN - TSP para a aplicação do método de Simulação de Monte Carlo;
- Propor recomendações de adequabilidade para o uso de técnicas de previsão de dados em avaliação de riscos em barragens considerando também a frequência de amostragem dos dados;
- Avaliar a viabilidade do uso do k NN - TSP como ferramenta de auxílio à tomada de decisão na monitoração de deslocamentos em barragens.

1.2 **Organização do Trabalho**

O restante deste trabalho está organizado do seguinte modo:

No capítulo 2 são apresentados os conceitos básicos e notações referentes à área de análise de séries temporais. Também são abordados temas como aprendizado de máquina para análise de séries temporais e o algoritmo k NN-TSP para a previsão de dados com dependência temporal.

No capítulo 3 são apresentados aspectos relevantes da instrumentação para segurança de barragens e o conceito de incerteza de medição, assim como os principais métodos para sua avaliação. Posteriormente são abordados estudos realizados na literatura sobre a influência do uso dados incertos no controle de processos e sobre a aplicação de técnicas de aprendizado de máquina para a previsão de dados.

No capítulo 4 é apresentada uma proposta de um modelo de simulação para reproduzir as influências provocadas pela incerteza de medição e pela frequência de amostragem dos dados na capacidade de predição do algoritmo k NN - TSP. Também é apresentada a aplicação do modelo proposto a uma série temporal artificial.

No capítulo 5 é apresentado um estudo de caso realizado em séries temporais provenientes de um processo de medição real. Nessa aplicação do modelo de simulação são utilizados os dados de monitoração de deslocamentos, provenientes dos instrumentos de pêndulo direto e invertido, empregados na Usina Hidrelétrica de Itaipu.

No capítulo 6 são apresentadas as conclusões deste trabalho e as suas principais contribuições e limitações, assim como propostas para trabalhos futuros.

2 PREVISÃO DE DADOS EM SÉRIES TEMPORAIS

Dados coletados ao longo do tempo podem ser representados por meio de uma série temporal (MORETTIN; TOLOI, 2006). Séries temporais podem ser provenientes de diversas áreas do conhecimento como economia (preços diários de ações, taxa mensal de desemprego, produção industrial), medicina (eletrocardiograma, eletroencefalograma), epidemiologia (número mensal de novos casos de meningite), meteorologia (precipitação pluviométrica, temperatura diária, velocidade do vento) (EHLERS, 2005). A utilização de métodos e técnicas de análise de séries temporais para o estudo de eventos e comportamentos contidos nas séries consiste em uma tarefa complexa e dependente do domínio de aplicação (FERRERO, 2009). Como a maior parte dos procedimentos estatísticos foi desenvolvida para analisar observações independentes, o estudo de séries temporais requer o uso de técnicas específicas (EHLERS, 2005).

Um dos principais objetivos do estudo de séries temporais consiste na previsão de valores futuros com base em valores passados. Nesse sentido, a tarefa de previsão torna-se uma importante ferramenta no auxílio a processos de tomada de decisões. Devido a grande quantidade de dados geralmente utilizada nessa tarefa e a complexa relação de dependência temporal dos dados, muitas abordagens tem tratado o problema de previsão por meio de técnicas de aprendizado de máquina (FERRERO, 2009; CASTRO, 2001; MATTOS et al., 2005).

Neste capítulo são apresentados os conceitos básicos referentes à área de análise de séries temporais bem como os métodos de previsão adotados na literatura. Também são abordados temas como aprendizado de máquina para análise de séries temporais e o algoritmo *k*NN adaptado para a previsão de dados com dependência temporal.

2.1 Fundamentos de Séries Temporais

Uma série temporal consiste em um conjunto de observações de um determinado fenômeno realizadas de modo sequencial ao longo do tempo, não necessariamente igualmente espaçadas. Esse tipo de dado tem como característica fundamental a dependência entre as observações adjacentes, ou seja, entre instantes de tempo (EHLERS, 2005). De modo geral, uma série temporal pode ser denotada como:

$$ST = (O_1, \dots, O_N) \quad (2.1)$$

onde N é o número de observações coletadas, O_t representa o dado observado no instante $t \in T = (0, \dots, N)$ e ST o conjunto ordenado de valores que descreve a série temporal.

Comumente na literatura, para uma melhor compreensão e análise dos eventos representados por uma série temporal, utiliza-se o conceito de decomposição da série em um conjunto finito de componentes independentes. As principais componentes abordadas são denominadas tendência, sazonalidade e resíduo (MORETTIN; TOLOI, 2006). Desse modo, cada uma das observações que compõem uma determinada série temporal pode estar influenciada por uma ou mais dessas componentes. Assim, pode-se definir também ST em termos dessas componentes como na Equação:

$$ST_t = T_t + S_t + R_t \quad (2.2)$$

onde $t \in T = (0, \dots, N)$ sendo N o número de observações coletadas. ST_t representa o valor de uma observação da série temporal no instante t e T_t , S_t , R_t representam a tendência, a sazonalidade e o resíduo no instante t da série.

No entanto, em grande parte dos problemas estudados, não é possível identificar diretamente a atuação dessas componentes na série, as quais podem ser extraídas e compreendidas por meio da aplicação de técnicas específicas propostas para a decomposição de uma série temporal (BROCKWELL; DAVIS, 1996).

Geralmente as componentes S_t e T_t possuem uma forte relação, de modo que a influência da tendência sobre a componente sazonal pode afetar significativamente os métodos de estimação de S_t . Nesse sentido, por meio da decomposição de uma série, pode-se isolar uma componente da outra para uma melhor análise dos dados (MORETTIN; TOLOI, 2006).

2.1.1 *Tendência*

A componente tendência corresponde ao movimento dominante em uma série temporal, o qual exerce influência sobre as observações por longos períodos de tempo, alterando o nível médio da série. Essa componente tem como característica o quase constante movimento crescente ou decrescente, que atua de modo suave ao longo da série (MALETZKE, 2009). Na Figura 1 a componente de tendência é representada por uma linha preta tracejada e a série temporal do fenômeno é observado em cinza.

As séries temporais podem possuir distintos tipos de comportamentos de tendência, nos quais os métodos de identificação dessa componente estão

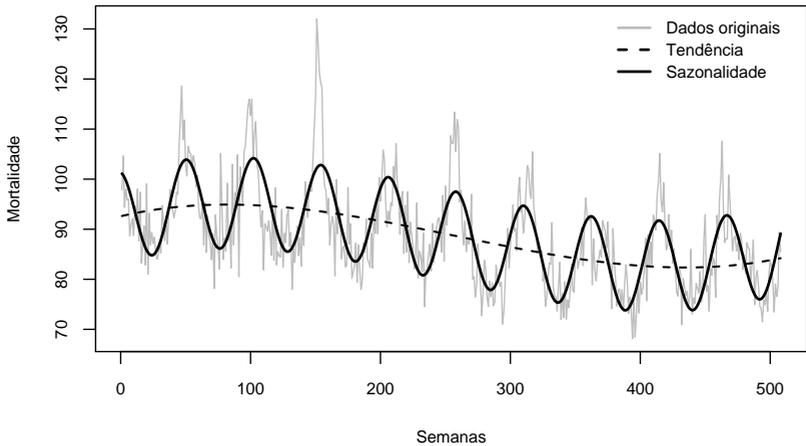


Figura 1: Série de dados referentes a mortalidade vascular (FERRERO, 2009).

baseados (EHLERS, 2005; SILVA, 2005). Os métodos mais utilizados nesse tipo de tarefa são (MORETTIN; TOLOI, 2006):

- Suavizar os valores da série próximos a um determinado ponto;
- Ajustar um função aos valores observados da série;
- Suavizar os valores da série por meio de sucessivos ajustes de retas de mínimos quadrados ponderados.

2.1.2 Sazonalidade

Um determinado comportamento que tende a se repetir em uma série temporal em diferentes períodos de tempo é denominado sazonalidade. Essa componente representa as oscilações ao longo da componente de tendência de acordo com uma determinada característica (FERRERO, 2009).

Dependendo do domínio de aplicação, a componente de sazonalidade pode ser de grande interesse pois pode permitir a identificação de comportamentos relevantes do fenômeno observado ou pode dificultar a percepção de outros eventos relevantes. Desse modo, a identificação dessa componente consiste em um procedimento importante no contexto de análise de séries temporais, a qual por sua existência pode revelar informações relevantes e sua

remoção pode ressaltar outras características da série temporal (MALETZKE, 2009). Na Figura 1, a componente sazonalidade está representada por uma linha preta contínua e a série temporal do fenômeno observado em cinza.

2.1.3 *Resíduo*

Geralmente os fenômenos caracterizados por meio de séries temporais podem apresentar eventos aleatórios característicos do próprio fenômeno ou gerados a partir dos procedimentos e equipamentos utilizados durante o processo de amostragem dos dados de interesse (MALETZKE, 2009).

No contexto de análise de séries temporais, esses comportamentos são representados pela componente de resíduo. A existência dessa componente não pode ser captada pelas componentes de tendência e sazonalidade e pode ainda dificultar a identificação das mesmas. Desse modo, a modelagem das demais componentes permite isolar e analisar o resíduo.

2.2 Modelagem de Séries Temporais para Previsão de Dados

A previsão de valores em séries temporais consiste em utilizar as N observações anteriores da série $ST = (O_1, O_2, \dots, O_N)$ para realizar a previsão do valor O_{t+1} por meio de um modelo de previsão m . Desse modo, a previsão do valor futuro O_{t+1} pode ser denotado pela Equação:

$$O_{t+1} = m(O_t, O_{t-1}, O_{t-2}, \dots, O_{t-N+1}) \quad (2.3)$$

Geralmente as técnicas de previsão são divididas em duas categorias. A primeira corresponde à utilização de modelos lineares, os quais podem ser entendidos como modelos para processos estacionários e modelos para processos não-estacionários (MORETTIN; TOLOI, 2006).

Nos processos estacionários assume-se que os valores das séries temporais oscilam em torno de uma média com variância constante ao longo do tempo. Alguns modelos com essa característica amplamente utilizados na literatura são: modelos auto-regressivos (AR), modelos de médias móveis simples (MA), modelos auto-regressivos de médias móveis (ARMA) e auto-regressivos de médias móveis integrados (ARIMA) (BOX; JENKINS; REINSEL, 1994; SILVA, 2005).

Séries temporais que apresentam algum tipo de tendência ou de sazonalidade são exemplos de séries com comportamento não-estacionário. Uma variação do modelo ARIMA que pode ser aplicada nesse caso é o modelo sazonal auto-regressivo de médias móveis integrado (SARIMA) (EHLERS,

2005).

A outra abordagem utilizada para a previsão em séries temporais baseia-se na construção de modelos não-lineares. Esses modelos são aplicados em séries cujo comportamento dificilmente poderia ser caracterizado por meio de modelos lineares (FERRERO, 2009; CAMILLERI, 2004).

Nesse sentido, vários métodos foram propostos na literatura para construir uma modelagem não-linear para séries temporais, tais como modelos bilinear, modelos auto-regressivos exponenciais, modelos de limiar auto-regressivo e modelos gerais de estados independentes.

No entanto, devido à complexidade matemática envolvida para a construção de modelos não-lineares estatísticos, abordagens baseadas em técnicas de aprendizado de máquina têm sido desenvolvidos na literatura para a previsão de séries temporais (MATTOS et al., 2005).

2.3 Aprendizado de Máquina

O aprendizado de máquina é uma importante subárea de pesquisa da inteligência artificial e tem como objetivos o desenvolvimento de métodos e técnicas computacionais que permitam adquirir de modo automático novos conhecimentos (ALPAYDIN, 2004).

Os algoritmos de aprendizado de máquina dependem fortemente de conhecimento para realizarem o processo de aprendizagem. Um dos métodos mais utilizados para extrair conhecimento e prever eventos futuros é a inferência indutiva, a qual faz uso de um conjunto de exemplos conhecidos para derivar novos conhecimentos. Esse modo de aprendizado apresenta um alto grau de complexidade e constitui uma das estratégias de maior interesse entre os pesquisadores da área (ALPAYDIN, 2004; MITCHELL, 1997).

As técnicas de aprendizado de máquina, segundo a literatura, podem ser divididas em paradigmas, os quais são brevemente apresentados a seguir (REZENDE, 2003):

- **Paradigma estatístico:** consiste em utilizar modelos estatísticos para encontrar uma aproximação do conceito induzido assumindo que os dados possuem distribuição normal;
- **Paradigma simbólico:** os sistemas baseados no paradigma simbólico realizam o processo de aprendizagem por meio da análise de exemplos e contra-exemplos, os quais geralmente estão representados na forma de expressão lógica, árvore de decisão e redes semânticas;

- **Paradigma baseado em exemplos:** nesse paradigma não é construído um modelo explícito com o conjunto de exemplos. Sistemas desenvolvidos com esse paradigma armazenam os exemplos e utilizam medidas de similaridade para identificar os casos mais similares ao exemplo a ser classificado;
- **Paradigma conexionista:** o conexionismo baseia-se em construções matemáticas inspiradas em conexões neurais do sistema nervoso humano. As redes neuronais artificiais são exemplos de técnicas baseadas nesse paradigma;
- **Paradigma genético:** as técnicas desenvolvidas nesse paradigma são inspiradas pela teoria de Darwin, na qual os indivíduos mais adaptados sobrevivem. Nesse sentido, um algoritmo genético consiste em um procedimento iterativo que mantém uma população de indivíduos, onde cada um é candidato à solução de algum problema específico.

No contexto de previsão de séries temporais por meio de técnicas de aprendizado de máquina, duas principais abordagens são utilizadas para a modelagem não-linear.

Uma abordagem para a previsão de dados por meio da modelagem não-linear consiste em realizar uma aproximação global para a previsão do valor futuro, para a qual são utilizadas todas as observações da série. A construção desses modelos é realizada por meio de um procedimento de treinamento no qual o modelo é um resultado da criação de relações entre os dados da série. As aproximações globais são geralmente realizadas por redes neuronais artificiais, funções polinomiais e funções racionais (CASTRO, 2001; MATTOS et al., 2005).

Outra estratégia utilizada para a construção de modelos não-lineares baseia-se na aproximação local, onde somente os valores mais próximos ou mais importantes em relação ao valor atual são usados para a previsão do valor futuro. Esses valores são selecionados na série por meio de uma medida de similaridade e utilizados como entradas para uma função de cálculo do valor futuro (FERRERO, 2009; SILVA, 2005). Geralmente, as funções mais aplicadas para essa tarefa são média local ponderada e não ponderada (KARUNASINGHE; LIONG, 2006; SOLOMATINE; MASKEY; SHRESTHA, 2006).

Como mencionado, a modelagem por meio de redes neuronais artificiais baseia-se na construção de modelos a partir de todos os valores da série. Nessa abordagem, o conhecimento do processo de previsão é encapsulado no modelo, não sendo transparentes as decisões do algoritmo (SOLOMATINE;

MASKEY; SHRESTHA, 2006). Desse modo, aplicações dessa técnica em áreas críticas, tais como medicina, hidrologia e segurança de barragens são pouco viáveis.

Nesse contexto, abordagens baseadas em aproximação local têm sido aplicadas com sucesso por meio de algoritmos como o k -Nearest Neighbor (FERRERO, 2009; SOLOMATINE; MASKEY; SHRESTHA, 2006).

2.4 Algoritmo k NN para Previsão de Séries Temporais

O algoritmo k NN baseia-se no aprendizado supervisionado¹ e pertence ao paradigma baseado em exemplos. A ideia geral desse algoritmo consiste em encontrar os k exemplos vizinhos mais próximos, contidos em um dado conjunto de treinamento, e classificar os novos exemplos como os exemplos mais similares já conhecidos (AHA; KIBLER; ALBERT, 1991). Se $k = 1$ um novo exemplo será classificado com a mesma classe de um único exemplo mais próximo determinado pela medida de similaridade. Se $k > 1$, a classe predominante dos k exemplos mais similares encontrados será atribuída ao novo exemplo.

Na Figura 2 é ilustrado o funcionamento do algoritmo k NN com $k = 1$ e $k = 4$ para a classificação de um novo exemplo E_i a partir de um conjunto de exemplos positivos (+) e negativos (-) disponíveis. Como pode ser observado, a determinação do valor de k dependendo do domínio de aplicação pode influenciar significativamente a atuação do algoritmo, pois para $k = 1$ o exemplo E_i foi classificado como positivo e para $k = 4$ a classificação de E_i foi como negativo.

O k NN demanda baixo esforço computacional durante a fase de treinamento, no entanto o custo para classificar novos exemplos pode ser alto, pois no pior caso o algoritmo utilizará para comparação todos os exemplos contidos no conjunto de treinamento.

No contexto de previsão de dados em séries temporais para a aplicação do algoritmo k NN é necessário adaptar a abordagem tradicional do algoritmo (KULESH; HOLSCHNEIDER; KURENNAYA, 2008). Desse modo, em (FERRERO, 2009) é proposta essa adaptação do algoritmo, denominado k -Nearest Neighbor - Time Series Prediction (k NN - TSP). A ideia consiste em considerar as w últimas observações ocorridas para encontrar sequências de tamanho w que apresentam comportamentos similares. E com base nessas subsequências é realizada a estimativa do valor futuro O_{t+1} .

¹No aprendizado supervisionado, o conjunto de exemplos que compõem o conjunto de treinamento são previamente classificados.

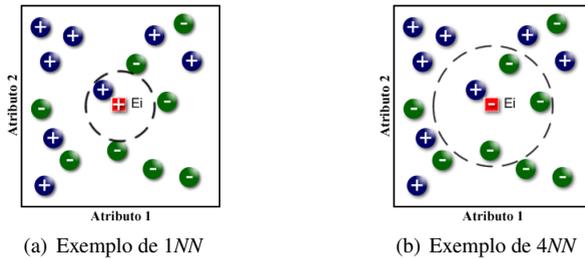


Figura 2: Aplicação do algoritmo k NN, com $k = 1$ e $k = 4$ (FERRERO, 2009).

Na Figura 3 é apresentado um exemplo de aplicação do algoritmo k NN-TSP para um valor de $k = 4$. O gráfico da figura apresenta uma série temporal da temperatura da água em um determinado local onde a linha em cinza representa as observações que compõem a série; a linha vermelha representa a janela de busca para extrair as sequências (quatro últimos valores registrados) e as linhas em verde representam as sequências similares encontradas pelo algoritmo.

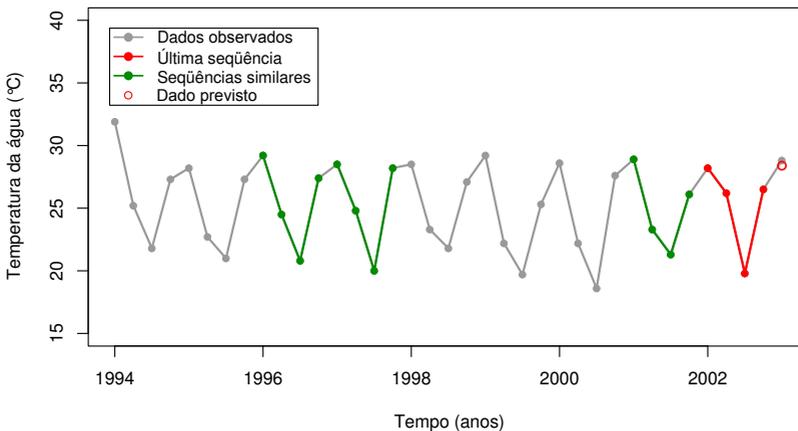


Figura 3: Exemplo de aplicação do algoritmo k NN-TSP (FERRERO, 2009).

Assim como para o algoritmo k NN, o desempenho efetivo do algoritmo k NN-TSP depende de alguns parâmetros importantes, tais como o conjunto de exemplos de treinamento, a medida de similaridade, cardinalidade do

conjunto de vizinhos mais próximos e a função de previsão (KULESH; HOLSCHNEIDER; KURENNAYA, 2008).

2.4.1 Conjunto de Séries de Treinamento

O conjunto de exemplos de treinamento consiste nas séries ou subsequências de uma mesma série que serão utilizadas para a busca das subsequências similares. Dependendo do domínio de aplicação o custo computacional de execução do algoritmo pode se tornar significativo, pois será necessário realizar comparações com todo o conjunto de treinamento (ALPAYDIN, 2004).

Para a seleção das séries ou subsequências de uma série é utilizada uma janela de tamanho w . O valor de w pode influenciar diretamente o desempenho do algoritmo pois afeta o formato dos padrões encontrados e consequentemente a representação do comportamento do fenômeno observado (FERRERO, 2009). Desse modo, foram desenvolvidas algumas abordagens na literatura que permitem determinar o valor de w , tais como: pela análise visual (KULESH; HOLSCHNEIDER; KURENNAYA, 2008), pelo cálculo da dimensão de correlação (GRASSBERGER; PROCACCIA, 1983), pelo método de Falsos Vizinhos Próximos (KENNEL; BROWN; ABARBANEL, 1992) e pelo método *Dynamic Time Wrapping* (MIZUHARA; HAYASHI; SUEMATSU, 2006).

2.4.2 Medida de Similaridade

A medida de similaridade é utilizada para quantificar o grau de similaridade entre os exemplos contidos no conjunto de treinamento e um novo exemplo. Quando o conjunto de dados é descrito por valores numéricos, as medidas de distância podem ser utilizadas (FERRERO, 2009). Existem diversos tipos de medidas de similaridade propostas na literatura (JAIN; DUBES, 1988), sendo a distância Euclidiana a mais utilizada nessas aplicações pois trata-se de uma medida intuitiva e de baixo custo computacional (KEOGH; KASSETTY, 2002).

No contexto de séries temporais, para a comparação entre duas séries ou entre duas subsequências de uma série, geralmente é necessária a normalização dos valores para nivelar ambas as séries.

Na Figura 4 está representado um exemplo de aplicação de normalização entre duas sequências E_i e E_j , as quais se encontram em níveis diferentes e apresentam o mesmo tamanho w . A normalização realizada nessa representação consiste em subtrair de cada série o valor médio da respectiva série (KEOGH; KASSETTY, 2002).

Muitas outras medidas de similaridade são propostas na literatura, tais como normalização de variâncias e combinação de distintas medidas de similaridade (KULESH; HOLSCHNEIDER; KURENNAYA, 2008; ILLA; ALONSO; MARRÉ, 2004; FABRIS; DRAGO; VAREJÃO, 2008).

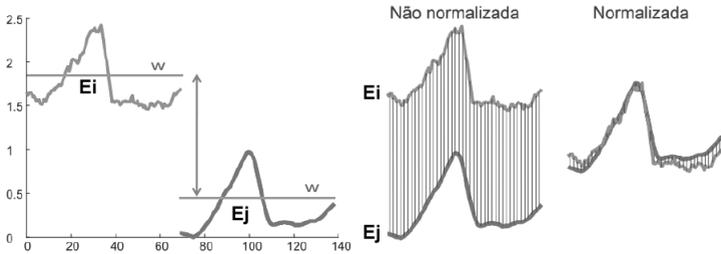


Figura 4: Normalização de séries temporais (KEOGH; KASSETY, 2002).

2.4.3 Conjunto de Vizinhos mais Próximos

Por meio da medida de similaridade são selecionadas pelo algoritmo $k \geq 1$ séries consideradas similares, as quais são utilizadas para a previsão do valor futuro. Frequentemente, para cada domínio de aplicação estudado, distintos valores de k são testados com o objetivo de encontrar o mais adequado ao problema (AHA; KIBLER; ALBERT, 1991).

No caso de simplesmente utilizar a classe predominante das k séries mais próximas para a previsão, valores ímpares de k são mais apropriados para evitar situações de empate. Algumas abordagens baseiam-se na atribuição de pesos para cada uma das k séries mais próximas de modo que as séries de maior similaridade terão maior peso do que as séries de menor similaridade (FERRERO, 2009).

2.4.4 Função de Previsão

Com base nas séries selecionadas $S = \{S_1, S_2, \dots, S_k\}$ para o conjunto de vizinhos mais próximos é realizado o cálculo do valor futuro O_{t+1} por meio da aplicação de uma função de previsão f .

Geralmente as funções de previsão utilizadas na literatura aproximam O_{t+1} pela média local ou pela média ponderada dos valores de $S_{i,w+1}$ de cada série $S_i \in S$, onde w representa o tamanho da janela utilizado (KARUNASINGHE; LIONG, 2006; SOLOMATINE; MASKEY; SHRESTHA, 2006).

Em Ferrero (2009) é proposta uma função de previsão denominada Média de Valores Relativos (MVR). Essa função aproxima o valor de O_{t+1} pelo valor de O_t mais a média local da diferença entre os valores de $S_{i,w+1}$ e $S_{i,w}$ de cada série $S_i \in \mathcal{S}$. Essa função é donatada por $f_{MVR}(S)$ definida pela Equação 2.4.

$$f_{MVR}(S) = O_t + \frac{\sum_{i=1}^k \Delta S}{k} \quad (2.4)$$

onde $\Delta S = (S_{i,w+1} - S_{i,w})$.

Uma das vantagens da função MVR em relação à função Média de Valores, geralmente utilizada, consiste no fato que permite prever valores futuros a partir de padrões encontrados em níveis diferentes da série.

Na Figura 5 é apresentado um exemplo de aplicação da função MVR e Média de Valores para o parâmetro $k = 2$.

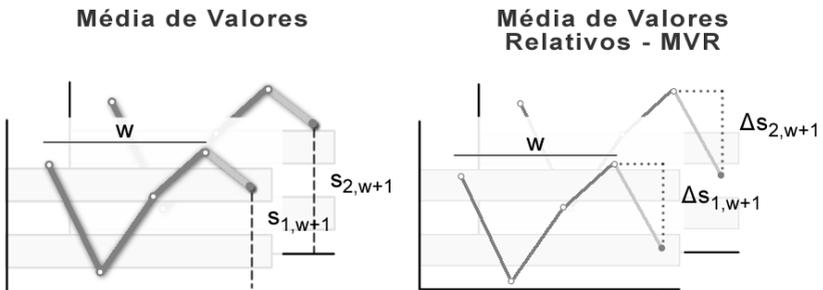


Figura 5: Função de previsão Média de Valores e MVR (FERRERO, 2009).

2.5 Considerações Finais

A previsão de dados em séries temporais pode ser realizada por meio da utilização de diferentes abordagens dependendo do comportamento das séries que representam o fenômeno observado. Nesse sentido, foram descritos nesse capítulo alguns fundamentos de séries temporais e duas principais abordagens propostas na literatura para a construção de modelos, os métodos lineares e métodos não-lineares. Posteriormente foram abordados conceitos básicos de aprendizado de máquina e também descrito o algoritmo k NN-TSP para previsão de dados em séries temporais.

No próximo capítulo são abordados aspectos de segurança de barra-

gens e a previsão de dados como ferramenta para a análise de riscos. Também é introduzido o conceito de incerteza de medição e os métodos propostos na literatura para a sua estimativa.

3 INCERTEZA DE MEDIÇÃO EM SEGURANÇA DE BARRAGENS

No monitoramento da segurança de barragens, os dados das variáveis de interesse são obtidos por meio de processos de medição. Sua confiabilidade está afetada por diversas causas, que resultam no afastamento do resultado da medição com referência ao valor verdadeiro da variável medida, denominado de erro de medição. As técnicas de aprendizado de máquina aplicadas no apoio à tomada de decisões sobre a segurança de barragens utilizam assim uma informação de entrada distorcida e, conseqüentemente, pode se esperar que as previsões geradas pelos algoritmos sejam também distorcidas. Se os erros de medição são pequenos, obter-se-ão previsões mais confiáveis, dentro da capacidade do próprio algoritmo de previsão utilizado. Se os erros são grandes, poder-se-ão esperar previsões pouco seguras.

De forma geral, pode se estabelecer que, para um desempenho adequado das ferramentas de apoio à tomada de decisão, é necessário que os dados sejam gerados por um processo de medição sob controle e capaz. O primeiro requisito significa que a distribuição do erro de medição deve ser consistente no tempo; o segundo visa a garantir que os erros de medição sejam pequenos em relação à tolerância da característica e à variação do processo (SILVA, 2003). Nesse sentido, torna-se fundamental o estudo da influência do uso de dados incertos sobre as ferramentas de auxílio à tomada de decisão, no contexto de controle de processos de medição (KONRATH, 2008; HAMBURG-PIEKAR; DONATELLI; SCHNEIDER, 2003; SILVA, 2003).

Neste capítulo são apresentados aspectos relacionados à instrumentação para segurança de barragens e aos conceitos de incerteza de medição, tais como os principais métodos de estimativa de incerteza. Posteriormente são abordados estudos realizados na literatura, sobre a influência do uso de dados incertos no controle de processos de medição e sobre a aplicação de técnicas de aprendizado de máquina para a previsão de dados.

3.1 Instrumentação para Segurança de Barragens

A segurança de barragens consiste em um conjunto de métodos para a observação do comportamento de distintas características que podem descrever a situação de uma barragem. A partir desses parâmetros é realizada uma avaliação das condições de segurança da barragem, por meio da verificação da validade das hipóteses e métodos de cálculo estabelecidos na fase de projeto. Desse modo, é possível constatar a necessidade da aplicação de procedimen-

tos corretivos e fornecer subsídios para a elaboração de novos critérios de avaliação (FEDERAL ENERGY REGULATORY COMMISSION - FERC, 2008; MEDEIROS, 2003; CBGB, 1996).

Nos últimos anos, a segurança de barragens tem se tornado um tema de grande interesse devido ao crescente número de barragens em exploração e ao processo de deterioração das estruturas com o passar do tempo, em uma quantidade significativa dessas obras (ROCHA et al., 1992). No Brasil, esse tema possui a maior relevância, pela existência de um elevado número de barragens, com importância estratégica do ponto de vista energético, abastecimento de água e irrigação, e também com grande potencial de perdas, em caso de acidentes, em função de suas localizações (PORTELA, 2003).

Durante o período construtivo e a fase de operação de uma usina hidrelétrica é necessário garantir a segurança da obra, visto que existe a possibilidade de que os maciços da obra apresentem comportamentos não previstos, devido às novas condições a que estão sendo submetidos. A instrumentação de barragens permite o monitoramento das estruturas por meio da coleta de grandezas relacionadas a situações que podem colocar em risco as estruturas de uma barragem. Desse modo, a instrumentação permite determinar quando uma segurança mínima aceitável é alcançada, garantindo que o dimensionamento de equipamentos, materiais e mão-de-obra serão adequados (SILVEIRA et al., 1994) e também possibilita detectar problemas com suficiente antecedência, permitindo a intervenção com medidas corretivas (ELETROBRÁS, 2003; SILVEIRA et al., 1994).

As grandezas usualmente monitoradas são: deslocamentos, pressões, vazão, topografia, temperatura e vibração (FEDERAL ENERGY REGULATORY COMMISSION - FERC, 2008). A modo de exemplo, na Figura 6 pode ser observada uma representação esquemática da distribuição dos instrumentos de auscultação em um bloco-chave da Usina Hidrelétrica de Itaipu, caracterizado pela densa malha de instrumentos que agrega em função do importante papel que desempenha para a integridade da estrutura. Esse bloco situa-se na região central da barragem, possuindo cerca de 185 metros de altura (ITAIPU BINACIONAL, 2008).

Grande parte da instrumentação de barragens é instalada durante o período construtivo, de modo que os instrumentos podem estar localizados dentro do concreto ou na fundação. Na maioria das barragens a instrumentação está dispersa nos mais diversos locais da estrutura. Grande parte dos instrumentos estão em lugares de difícil acesso para os operadores, em locais com escassa iluminação e em alguns casos em regiões de alta umidade e temperatura. Também, a instrumentação pode estar disposta em uma

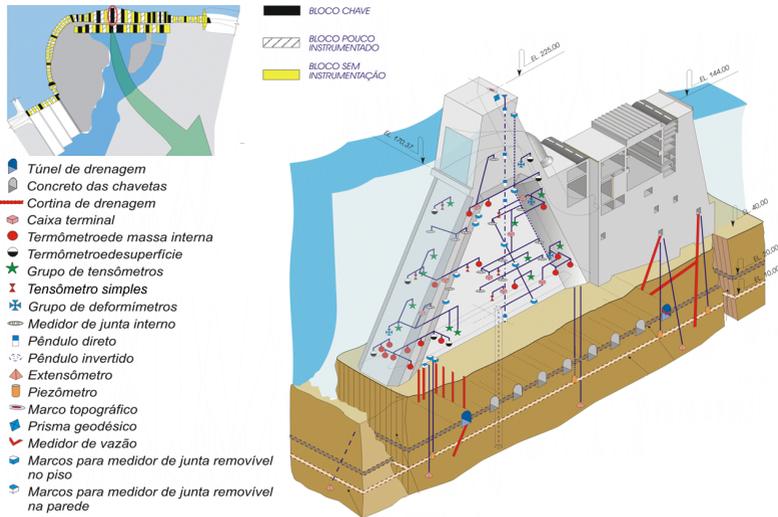


Figura 6: Instrumentação de bloco-chave da Usina de Itaipu (OSAKO, 2008).

longa faixa de extensão e de altura em uma barragem, como por exemplo, na barragem de Itaipu, os instrumentos estão dispostos em 7600 m de extensão e 196 m de altura (ITAIPU BINACIONAL, 2008). Desse modo, o ambiente físico em uma barragem traz algumas dificuldades para a realização das leituras, as quais por muitas vezes podem fazer com que os operadores façam o registro incorreto das medições realizadas (ELETROBRÁS, 2003).

As frequências de leitura da instrumentação devem ser adequadas para o acompanhamento do comportamento das estruturas de modo que seja possível observar a taxa de variação das grandezas monitoradas. Nesse sentido, em situações nas quais a frequência de coleta dos dados seja ajustada de forma inconsistente a interpretação física das grandezas pode trazer resultados não realistas (CBGB, 1996).

Em grandes aproveitamentos hidrelétricos são geralmente instalados algumas centenas de instrumentos. O trabalho necessário para a aquisição, processamento, armazenamento e análise dessa grande quantidade de leituras consiste em uma operação dispendiosa e requer um planejamento adequado.

Nesse contexto, a automatização da instrumentação de barragens possibilita reduzir parte dessas tarefas (SILVEIRA, 2006). No entanto, aspectos relacionados à transmissão dos dados coletados devem ser considerados no

processo de implantação da automação dos instrumentos, pois podem contribuir para o erro de medição dos instrumentos.

3.1.1 Avaliação de Riscos

A análise dos dados adquiridos pela instrumentação permite verificar se o comportamento da barragem está de acordo com o previsto, conforme indicado pelas ferramentas de análise, e avaliar o nível de segurança da barragem (PORTELA, 2003).

Nesse contexto, a avaliação de riscos em uma barragem deve ser capaz de identificar problemas e recomendar soluções, tais como estratégias corretivas e operacionais (PAN; 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 (FERRERO, 2009).

Pelas razões acima, a instrumentação de barragem representa um grande desafio para a metrologia, dadas as dificuldades existentes para assegurar a confiabilidade das medições ou mesmo calibrar os instrumentos. Na próxima seção são apresentados conceitos metrológicos chave, dando ênfase particular à definição de incerteza de medição e aos principais métodos de avaliação da incerteza propostos na literatura.

3.2 Incerteza de Medição

O erro ε é a diferença entre o resultado base da medição Y e o valor do mensurando (valor verdadeiro) Y_v (Equação 3.1).

$$\varepsilon = Y - Y_v \quad (3.1)$$

O valor do mensurando é uma quantidade que não pode ser conhecida com exatidão; conseqüentemente, o erro de medição também não pode ser conhecido com exatidão. Aceita-se geralmente que o erro de medição pode ser separado em duas componentes: o erro sistemático e o erro aleatório. O erro sistemático pode ser estimado pela diferença entre a média das medições repetidas e o valor de referência da característica medida. Essa estimativa é denominada de tendência. Os erros sistemáticos também não podem ser conhecidos com exatidão e não podem ser corrigidos completamente de modo que sempre existirá um resíduo de valor desconhecido. O comportamento do erro aleatório pode ser caracterizado pelas propriedades da distribuição experimental de medições repetidas. Em geral, assume-se que essa distribuição

é aproximadamente normal ou gaussiana. O desvio padrão experimental é geralmente utilizado para caracterizar a dispersão das medições.

A incerteza de medição é um conceito fundamental da metrologia atual e está na base do conceito de rastreabilidade¹ (BELANGER et al., 2000). Pode ser definida como "...um parâmetro associado ao resultado de uma medição que caracteriza a dispersão dos valores que podem ser fundamentalmente atribuídos ao mensurando" (ABNT; INMETRO, 2003). Nesse sentido, a incerteza expressa o efeito das contribuições aleatórias e da correção imperfeita dos resultados para os efeitos sistemáticos.

Os conceitos de erro de medição e incerteza de medição são distintos, no entanto estão relacionados, pois o valor do mensurando pode ser descrito pela existência dos erros de medição. Desse modo, uma região de dúvida é associada ao resultado da medição, a qual é representada pela incerteza de medição.

O conceito de incerteza de medição definido pelo Guia para a Expressão da Incerteza de Medição (GUM) (ABNT; INMETRO, 2003) possui a propriedade de ser transportável. Assim, as incertezas estimadas na calibração de um determinado instrumento podem formar parte da incerteza das medições realizadas com esse mesmo instrumento. Uma particularidade desse processo de propagação da incerteza é que, por exemplo, uma contribuição à incerteza considerada como aleatória em um nível da cadeia de rastreabilidade pode se propagar como sistemática para o nível inferior subsequente. Desse modo, a classificação das contribuições à incerteza não é realizada segundo o comportamento estocástico, mas baseado no procedimento utilizado para estimar as contribuições (SILVA, 2003).

Assim, as contribuições à incerteza podem ser classificadas em duas categorias, denominadas incertezas tipo *A* e incertezas tipo *B* (ABNT; INMETRO, 2003).

Incertezas tipo *A*

A incerteza tipo *A* é avaliada a partir da análise de uma série de observações, realizada conforme os métodos da estatística clássica (VUOLO, 1999).

Considerando que n observações Y_k realizadas de modo independente

¹Propriedade do resultado de uma medição ou do valor de um padrão estar relacionado a referências estabelecidas, geralmente a padrões nacionais ou internacionais, através de uma cadeia contínua de comparações, cada uma com incertezas estabelecidas. Os resultados de medição sem uma declaração de incerteza associada não podem ser comparados com outros valores ou com valores de referência (INMETRO, 2009).

em medições repetidas com a mesmas condições de medição de um determinado mensurando M , de valor essencialmente constante, a média \bar{Y} e a variância S^2_Y das observações são definidas pelas Equações 3.2 e 3.3, respectivamente.

$$\bar{Y} = \frac{1}{n} \sum_{k=1}^n Y_k \quad (3.2)$$

$$S^2_Y = \frac{1}{n-1} \sum_{k=1}^n (Y_k - \bar{Y})^2 \quad (3.3)$$

A raiz quadrada da variância é o desvio padrão experimental das observações e caracteriza a dispersão dos valores de Y_k em torno da média.

Quando a média de um conjunto de observações é utilizada para a estimativa do valor de um determinado mensurando e não há presença de outras fontes de incerteza, o desvio padrão das Y_k medições repetidas é uma estimativa não tendenciosa para a incerteza tipo A . Nessa condição, o desvio padrão define uma região, centrada no valor médio das observações, a qual contém os valores que podem ser fundamentalmente atribuídos ao mensurando.

Desse modo, o resultado de uma avaliação de incertezas do tipo A resulta em um valor de incerteza padrão da média denotado pela Equação 3.4.

$$u_A = \sqrt{\frac{S^2_Y}{n}} \quad (3.4)$$

Incertezas tipo B

Quando é pouco viável a realização de várias observações repetidas do mensurando, uma alternativa que pode ser aplicada é a avaliação de incertezas do tipo B . Nesse tipo de avaliação a estatística convencional não é aplicável e as contribuições à incerteza são estimadas por meio do julgamento científico baseado nas informações disponíveis relacionadas a variabilidade do mensurando. Essas informações podem incluir dados de experiências prévias, conhecimento geral do comportamento do fenômeno, propriedades dos materiais e instrumentos relevantes, especificações de fabricantes, dados fornecidos em certificados

de calibração e incertezas descritas em manuais.

Alguns exemplos que podem originar contribuições tipo *B* são o erro de arredondamento devido à resolução, as variações de temperatura ambiente não corrigidas, erros de linearidade não corrigidos, resíduos de zeragem.

Em avaliações de incertezas tipo *B* são utilizadas diferentes Funções Densidade de Probabilidade (FDP) para representar a incerteza de cada contribuição considerada. Desse modo, o procedimento para a determinação da incerteza tipo *B* consiste em admitir, para os possíveis valores do mensurando, uma distribuição de probabilidades que esteja de acordo com todo o conhecimento e informações disponíveis. As distribuições de probabilidade mais utilizadas são (ABNT; INMETRO, 2003; CASTRUP, 2001; SILVA, 2003):

- **Distribuição normal:** é utilizada para modelar o erro de repetitividade² e contribuições provenientes de processos de medições anteriores. Por exemplo, a incerteza associada ao valor de uma correção decorrente de um processo de calibração.
- **Distribuição retangular:** essa distribuição é utilizada quando é possível estimar os limites de variação do mensurando, porém não é conhecida a distribuição dos valores entre esses limites. Esse é o caso de informações oriundas de especificações do fabricante, tais como erros sistemáticos não corrigidos. Para uma contribuição à incerteza caracterizada por essa distribuição a relação entre a incerteza padrão e os limites de variação dessa distribuição $\pm a$ é dada pela Equação 3.5:

$$u_B = \frac{a}{\sqrt{3}} \quad (3.5)$$

- **Distribuição U:** essa distribuição é recomendada em casos semelhantes a amostragem de uma variação senoidal, nos quais existe um maior acúmulo de valores máximos e mínimos. Para a distribuição U a relação entre a incerteza padrão e os limites de variação $\pm a$ segue a Equação 3.6

$$u_B = \frac{a}{\sqrt{2}} \quad (3.6)$$

²Grau de concordância entre resultados de medições repetidas realizadas exatamente nas mesmas condições.

Um dos problemas das avaliações de incertezas tipo *B* está na subjetividade do procedimento, visto que depende em grande parte do conhecimento do avaliador em relação ao mensurando e o sistema de medição. Embora as avaliações tipo *B* sejam consideradas de menor qualidade, em algumas situações de restrições técnicas ou econômicas é pouco viável a aplicação de uma análise experimental. Adicionalmente, consiste na única alternativa de avaliação praticável durante o processo de seleção de sistemas de medição.

Em casos nos quais fontes significativas para a avaliação de incertezas são desconsideradas devido ao conhecimento limitado do avaliador, a amplitude da faixa de incerteza pode ser menor que os erros existentes. Em outras situações, suposições muito conservadoras em relação às contribuições para a amplitude da faixa da incerteza de medição podem ser muito maior que a magnitude dos erros existentes. Desse modo, embora os conceitos de erro de medição e incerteza de medição estejam relacionados ao mesmo fenômeno, pode ocorrer de não existir relação entre os erros e a incerteza expressa em uma determinada medição (SILVA, 2003).

O GUM é um documento publicado pela *International Organization for Standardization* (ISO) em conjunto com o *Bureau International des Poids et Mesures* (BIPM) e outras entidades internacionais da área científica para estabelecer métodos de cálculo de incerteza que possam ser universalmente aplicados. Desse modo, segundo o GUM, o procedimento de avaliação da incerteza de medição pode ser realizado por meio do Método da Propagação das Incertezas ou pelo Método da Propagação das Distribuições, os quais são descritos nas subseções seguintes.

3.2.1 *Método Clássico ou da Propagação das Incertezas*

O Método Clássico baseia-se na lei de propagação das incertezas, que estabelece que a incerteza de cada fator de influência na medição é propagada para a incerteza final da medição por meio de uma expansão em Série de Taylor. Assume-se, através da aplicabilidade do Teorema do Limite Central que o resultado da medição possuirá uma Função Densidade de Probabilidade (FDP) normal, ou t-Student quando o número de graus de liberdade efetivo for baixo ($v_{eff} \leq 30$).

A propagação das incertezas das grandezas de influência é realizada por meio de um modelo matemático que representa o mensurando.

$$M = f(Y_1, Y_2, \dots, Y_n) \quad (3.7)$$

onde M representa o valor do mensurando e Y_i as grandezas de influência.

O procedimento para a estimativa da incerteza por meio do Método Clássico proposto pelo GUM baseia-se na realização dos seguintes passos:

1. Definir o modelo matemático da medição;
2. Identificar e aplicar eventuais correções de erros sistemáticos;
3. Verificar todas as fontes de incerteza que podem ter influência no resultado do ensaio ou da calibração;
4. Calcular os coeficientes de sensibilidade de cada contribuição;
5. Avaliar as incertezas tipo A ;
6. Avaliar as incertezas tipo B ;
7. Avaliar as correlações entre as fontes de incerteza quando existirem e forem significantes;
8. Estimar a incerteza padrão de cada contribuição;
9. Obter a incerteza padrão combinada;
10. Obter a incerteza expandida.

Se o modelo matemático que representa a estimativa do valor mensurando é denotado por $m = f(y_1, y_2, \dots, y_n)$, a incerteza padrão combinada é calculada por meio da Equação 3.8:

$$u_c(m) = \sqrt{\sum_{i=1}^n \left[\frac{\partial m}{\partial y_i} \cdot u(y_i) \right]^2 + 2 \cdot \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{\partial m}{\partial y_i} \cdot \frac{\partial m}{\partial y_j} \cdot u(y_i) \cdot u(y_j) \cdot r(y_i, y_j)} \quad (3.8)$$

onde $u_c(m)$ representa a incerteza padrão combinada, $u(y_i)$ a incerteza padrão de uma grandeza de influência y_i , $\partial m / \partial y_i$ o coeficiente de sensibilidade da grandeza y_i e $r(y_i, y_j)$ o coeficiente de correlação entre duas grandezas de influência y_i e y_j .

Em aplicações práticas reais é necessária a utilização de uma medida de incerteza que defina um intervalo com maior probabilidade de abrangência. Essa medida é obtida por meio do cálculo da incerteza expandida dado pela Equação 3.9.

$$U = k \cdot u_c(m) \quad (3.9)$$

onde U representa a incerteza expandida e k um fator de abrangência calculado a partir dos graus de liberdade efetivos v_{eff} e da probabilidade de abrangência desejada. Por exemplo, as incertezas expandidas com valores de k de 2 a 3, definem intervalos com confiança de 95,5% a 99,7%, aproximadamente.

O cálculo dos graus de liberdade efetivos proposto no GUM baseia-se na utilização da fórmula de Welch-Satterthwaite, apresentado na Equação 3.10.

$$v_{eff} = \frac{u_c(m)^4}{\sum_{i=1}^n v_i^{-1} \cdot \left(\frac{\partial m}{\partial y_i} \cdot u(y_i) \right)^4} \quad (3.10)$$

onde v_i são os graus de liberdade de uma grandeza de influência y_i .

No entanto, para que a aplicação do Método Clássico apresente resultados aceitáveis é necessário considerar algumas premissas do método, tais como:

- *Linearidade do modelo matemático da medição*: quando um modelo matemático é utilizado para descrever um sistema de medição, pode ocorrer que esse modelo seja complexo demais, ou então, não permita uma solução analítica. Nesses casos, como a Equação 3.8 é obtida a partir de um desenvolvimento em série de Taylor, quando se considera insignificante a influência dos termos de ordem superior, e o modelo matemático apresenta não-linearidade a aproximação por meio dessa equação pode produzir resultados não representativos (COX; HARRIS, 2001; JORNADA; JORNADA, 2007);
- *Aplicabilidade do Teorema do Limite Central*: em determinadas situações, quando a incerteza combinada é dominada por uma fonte de incerteza com uma FDP significativamente diferente da distribuição normal, a FDP de saída pode não ser do tipo normal ou t-Student (COX; HARRIS, 2001; JORNADA; PIZZOLATTO, 2005).
- *Aplicabilidade da equação de Welch-Satterthwaite*: a equação de Welch-Satterthwaite para o cálculo dos graus de liberdade efetivos é adequada em situações nas quais as variáveis são independentes. Entretanto, quando efeitos de correlação são levados em consideração,

essa equação pode distorcer o valor dos graus de liberdade (AIBE, 2000; VUOLO, 1999; JORNADA; JORNADA, 2007).

Quando essas premissas não são satisfeitas, métodos alternativos para o cálculo da incerteza devem ser utilizados. Neste caso, a proposta elaborada no Suplemento 1 do GUM (GUM, 2005) consiste em utilizar um método numérico para a estimativa da incerteza.

3.2.2 *Método Numérico ou da Propagação das Distribuições*

A utilização de métodos numéricos para a solução de problemas físicos e matemáticos tem registros históricos em 1777 com Georges Louis Leclerc Comte de Buffon para a determinação do número π .

Durante a Segunda Guerra Mundial, no Projeto Manhattan, o uso de métodos numéricos passou a ter extensa aplicação após a utilização para a construção da bomba atômica conjuntamente com o desenvolvimento tecnológico computacional. Nesse projeto, foram realizadas simulações por meio de amostragens aleatórias com o objetivo de obter soluções aproximadas de problemas referentes à difusão aleatória de nêutrons do material nuclear (SOBOL, 1994; MOSCATTI; MEZZALIRA; SANTOS, 2004).

Os matemáticos John Von Neumann e Stanislaw Ulam são considerados os principais autores desse método numérico que ficou conhecido como Simulação de Monte Carlo (SMC). A SMC está inserida no ramo da estatística bayesiana, a qual faz uso de conhecimentos anteriores ao evento para a partir desses obter um conhecimento posterior. Neste sentido, a SMC utiliza informações *a priori* das grandezas de entrada que afetam a incerteza, por meio das suas distribuições de probabilidade para gerar um conjunto de valores aleatórios e então obter a distribuição de probabilidade da grandeza de saída.

Embora a SMC seja um método difundido para a análise de incertezas, recentemente foi incluído no GUM por meio da publicação do Suplemento 1 no qual foi chamado de Método Numérico para Propagação de Distribuições.

As principais diferenças entre o Método Clássico e o Método Numérico consiste no tipo de informação utilizada para caracterizar as grandezas de influência e no modo como essas informações são processadas para avaliar a incerteza de medição (KONRATH, 2008). Na Figura 7 é apresentada uma representação esquemática do funcionamento do Método Clássico (Figura 7(a)) e da SMC (Figura 7(b)).

No Método Clássico, as grandezas de influência devem ser caracterizadas por meio de uma FDP, da qual também são necessários para o método,

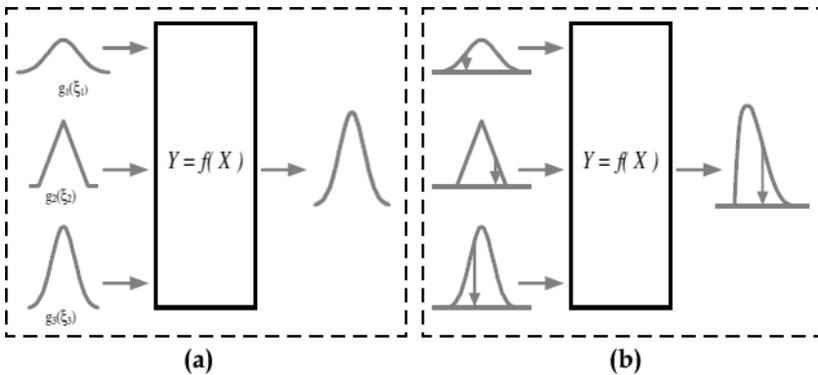


Figura 7: Comparação entre o Método Clássico (a) e o Método Numérico (b) (KONRATH, 2008) adaptado de GUM (2005).

a média, o desvio padrão e os graus de liberdade. A FDP da grandeza de saída é obtida a partir da avaliação do modelo matemático de medição por meio da combinação de amostras aleatórias das grandezas de influência, respeitando as características das respectivas FDPs. Pressupõe-se a premissa de que cada possível valor obtido aleatoriamente da distribuição, como grandeza de influência, é tão legítimo como qualquer outro valor (DONATELLI; KONRATH, 2005).

Desse modo, a partir da FDP resultante, as estatísticas de posição e dispersão da grandeza de saída podem ser calculadas. A incerteza de medição será calculada de acordo com um determinado nível de confiança requerido, após uma quantidade adequada de simulações executadas (GUM, 2005; BAZILIO et al., 2006; COUTO; DAMASCENO, 2006).

De modo geral, a avaliação da incerteza de medição por meio da aplicação da SMC é realizada em duas fases (COX; HARRIS, 2001). A primeira consiste em estabelecer o modelo matemático de medição e a segunda envolve a avaliação do modelo. A Figura 8 apresenta um fluxograma que descreve os passos para a aplicação da SMC, os quais podem ser estruturadas especificamente por meio dos seguintes passos (GUM, 2005):

1. Definir a grandeza de saída;
2. Decidir quais grandezas influenciam a grandeza de saída;
3. Desenvolver o modelo matemático que relaciona as grandezas de influência com a grandeza de saída;

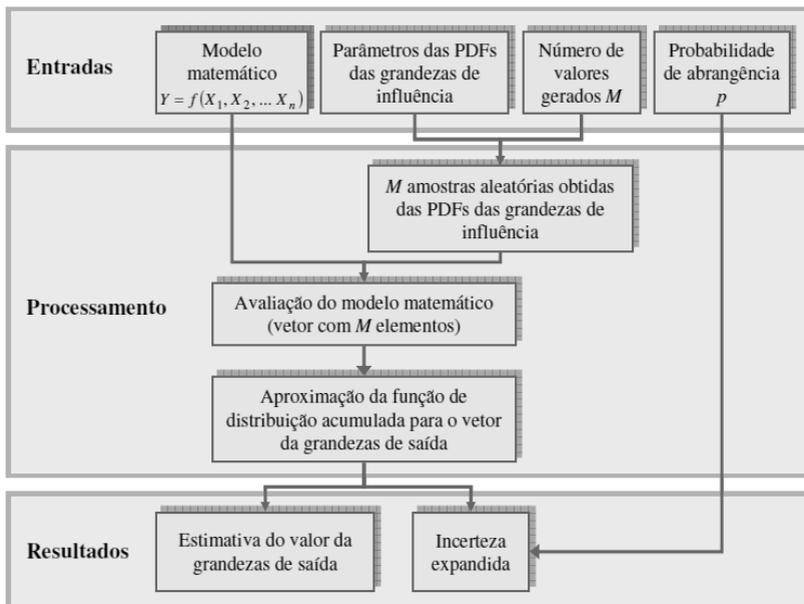


Figura 8: Fluxograma simplificado da avaliação de incerteza usando SMC (KONRATH, 2008) adaptado de GUM (2005).

4. Estimar os parâmetros apropriados das FDPs das grandezas de influência;
5. Estabelecer o número de simulações adequado para a incerteza do resultado que se deseja;
6. Realizar as simulações e propagar as FDPs por meio do modelo matemático de medição desenvolvido para obter a FDP da grandeza de saída.

A partir da FDP da grandeza de saída obtida por meio da SMC é possível extrair, dentre outros, os seguintes parâmetros:

- O valor médio da grandeza de saída;
- O desvio padrão, o qual se assume como incerteza padrão;

- O intervalo que contenha com determinada probabilidade de abrangência p , um valor desconhecido qualquer da grandeza de saída.

Modelos matemáticos não-lineares, distribuições assimétricas das grandezas de influência, contribuições não-normais dominantes, correlações entre as grandezas de influência são exemplos de restrições da aplicação do Método Clássico, que são automaticamente consideradas por meio da SMC (GUM, 2005; DONATELLI; KONRATH, 2005).

No entanto, para que a qualidade dos resultados provenientes da aplicação da SMC seja significativa, os seguintes fatores necessitam maior atenção:

- Representatividade do modelo matemático de medição;
- Qualidade da caracterização das grandezas de influência;
- Características do gerador de números pseudo-aleatórios utilizado;
- Número de simulações realizadas;
- Procedimento de definição do intervalo de abrangência.

3.3 Influência da Incerteza de Medição em Ferramentas de Análise de Séries Temporais

Processos de medição que operam em ambientes que se afastam das condições de referência e estão sujeitos às restrições econômicas e de tempo, podem ser significativamente afetados pelos erros de medição.

Poucos trabalhos na literatura avaliaram a influência da incerteza dos dados nas técnicas de aprendizado de máquina sobre os valores preditos. No estudo de Shrestha e Solomatine (2006) é proposto um método para avaliar a incerteza de predição de algumas técnicas de aprendizado de máquina por meio da determinação de um intervalo de confiança para os valores previstos. As técnicas de aprendizado de máquina utilizadas nesse estudo foram árvore de decisão, rede neuronal artificial e regressão localmente ponderada, que foram analisadas com diferentes configurações de estrutura e parâmetros. Para a construção dos modelos de predição foram utilizados dados reais e artificiais que representam comportamentos na área de hidrologia. Trabalhos posteriores propuseram refinamentos no método e o estudo de outras técnicas de aprendizado de máquina (SHRESTHA; SOLOMATINE,

2008; SHRESTHA; KAYASTHA; SOLOMATINE, 2009). Nesses estudos foi considerada a presença de uma variabilidade aleatória nos dados utilizados para a construção dos modelos.

No entanto, nenhuma literatura que relacionasse a efetividade de ferramentas de aprendizado de máquina para a previsão de dados e a influência de erros provenientes do processo de medição foi encontrada. Desse modo, o trabalho baseou-se no estudo de uma literatura, na qual é avaliado o efeito do uso de dados incertos nas ferramentas tradicionais de controle de processos.

No contexto de processos de medição, um dos métodos mais utilizados é Controle Estatístico de Processos (CEP) por meio da utilização de gráficos de controle (WOODAL, 2000). Esses gráficos são ferramentas visuais de acompanhamento das características da qualidade a serem controladas (ALENCAR et al., 2005; QS-9000, 1997).

Nesse sentido, os erros de medição podem gerar distorções nos gráficos de controle de modo que os resultados dessa ferramenta podem produzir bases pouco confiáveis para a tomada de decisão no processo (KONRATH, 2008).

Em estudos iniciais apresentados em Geus, Killmaier e Weckenmann (2001) o efeito da incerteza de medição foi avaliado no gráfico de controle de desvios padrões, no qual observou-se um aumento da média quando valores afetados pelo processo de medição são utilizados para a construção do gráfico. Outra importante constatação do estudo aponta que se a incerteza do processo de medição utilizado na aplicação do gráfico de controle é maior do que a incerteza do processo usado na fase construção do gráfico a probabilidade de ocorrência de alarmes falsos aumenta.

No trabalho de Hamburg-Piekar, Donatelli e Schneider (2003) foi analisada a influência da incerteza de medição na eficiência do posicionamento da média de processos normalmente distribuídos nos gráficos de controle de valores individuais por meio de simulação computacional. Para a detecção de sinais fora de controle foram utilizadas as regras de interpretação propostas por (WHEELER, 1991; WESTERN-ELECTRIC-COMPANY, 1958). No modelo de incerteza estabelecido foram considerados dois tipos de contribuições à incerteza: uma parcela associada ao erro de repetitividade e outra parcela associada aos erros sistemáticos desconhecidos e residuais, os quais poderiam variar linearmente com o valor do mensurando.

Por meio dos resultados apresentados nesse trabalho constatou-se que a incerteza de medição não tem efeito significativo em relação ao número esperado de unidades de produto para posicionar a média do processo e em relação ao número esperado de ajustes. No entanto, verificou-se que a incer-

teza de medição influencia de modo expressivo a amplitude da faixa dentro da qual pode estar a média verdadeira do processo.

Em Silva (2003) foi realizado um estudo sobre a influência da incerteza de medição na capacidade do gráfico de controle de valores individuais e do gráfico de amplitudes móveis na detecção de mudanças abruptas na média do processo. O modelo de incerteza considerado nesse trabalho consiste de contribuições de diferentes naturezas, tais como erros aleatórios, erros sistemáticos constantes e linearmente dependentes do mensurando e erros de arredondamento (resolução). Utilizando simulação computacional o modelo de incerteza proposto foi aplicado aos gráficos de controle com diversos conjuntos de regras de interpretação. Os parâmetros para os modelos de erro foram determinados com base em uma dada declaração de incerteza de medição e a análise foi restringida a processos de medição com propriedades estatísticas invariáveis no tempo.

Com base nos resultados desse estudo, verificou-se que os erros aleatórios e a resolução com que se registram os dados são as contribuições de maior influência sobre a sensibilidade dos gráficos de controle. Desse modo, foram propostos critérios para seleção e validação de processos de medição utilizados em aplicações de CEP, os quais foram comparados com os usados comumente na indústria. Com base nos resultados dessas comparações constatou-se que os critérios adotados na indústria permitem aceitar sistemas de medição com resolução inadequada.

No estudo realizado por Konrath (2008), a incerteza de medição foi analisada em relação ao desempenho do gráfico de controle da média. Considera-se que os estudos anteriormente mencionados não realizam uma completa modelagem do processo de medição, visto que comparam o desempenho do gráfico de controle construído com os valores do mensurando com o desempenho médio de um conjunto de gráficos construídos a partir desse. Desse modo, a contribuição dos erros de medição na etapa de construção do gráfico de controle não é considerada na avaliação da incerteza de medição na etapa de validação do desempenho do gráfico.

Nesse contexto, por meio da aplicação computacional do método de SMC, foi desenvolvido um modelo matemático para propagar a incerteza de medição através das fases do CEP. Desse modo, o efeito da incerteza de medição foi avaliado sobre a incerteza dos limites de controle e sobre a capacidade do gráfico de controle de médias em detectar mudanças no processo.

O estudo abordou também distintas configurações dos parâmetros de influência do CEP, tais como tamanhos de subgrupos e número de subgrupos usados para a estimativa dos limites de controle, tipos de regras de

interpretação, e equações de medição com quantidades de entrada normais e não-normais, correlacionadas e comportamentos sistemáticos.

Constatou-se que os limites de controle inferior e superior se afastam da média quando construídos e utilizados com dados incertos de modo que esse efeito pode ser reduzido com o aumento do tamanho dos subgrupos. Também verificou-se com base nos resultados que a capacidade do gráfico de controle de médias em detectar mudanças diminui quando a incerteza de medição aumenta, independentemente do tamanho dos subgrupos.

Outra importante observação dos resultados desse trabalho está relacionada à influência da correlação entre as quantidades de entrada, na qual foi possível identificar que a capacidade do gráfico em detectar mudanças na média do processo é sobreestimada quando não é considerado o efeito das correlações.

Como pode ser observado nas descrições dos estudos apresentados nesta seção, a incerteza de medição pode resultar em potenciais distorções dos gráficos de controle e conseqüentemente em um aumento na taxa de alarmes falsos e na redução da capacidade para detectar sinais de instabilidade ou anomalias.

Como mencionado, os modelos construídos para a previsão de valores futuros, por meio de técnicas de aprendizado de máquina, também baseiam-se nos dados coletados por meio de um processo sujeito a erros de medição. Desse modo, torna-se importante a avaliação da capacidade dessas ferramentas quando aplicadas em dados incertos.

3.4 Considerações Finais

Sistemas eficientes de monitoração são fundamentais no apoio a processos de medição que exijam alta qualidade. Nesse contexto, a previsão de valores futuros consiste em uma importante ferramenta de monitoração e prevenção pela capacidade de antecipar informações do processo de medição sob análise e com base nessas, realizar ações de contenção do problema.

No entanto, os métodos tradicionais, por se basearem na aplicação de técnicas estatísticas e na amostragem de dados, possuem limitações para detectar mudanças no processo. Desse modo, abordagens baseadas em aprendizado de máquina têm sido propostas para auxiliar nessa tarefa.

Analogamente aos métodos tradicionais, a efetividade dessas ferramentas está influenciada pela qualidade dos dados gerados pelo processo de medição.

Desse modo, neste capítulo foram apresentados os principais métodos

para a estimativa da incerteza de medição propostos pelo Guia para a Expressão da Incerteza de Medição e trabalhos propostos na literatura para a avaliação do efeito do uso de dados incertos para a construção de ferramentas de apoio ao controle de processos de medição.

No capítulo seguinte é apresentado o método proposto neste trabalho para a avaliação da influência da incerteza de medição na previsão de dados em séries temporais. Também é apresentada uma avaliação do modelo proposto sobre uma série temporal artificial com características de interesse para a segurança de barragens.

4 MÉTODO DE AVALIAÇÃO PROPOSTO

A utilização do algoritmo k NN - TSP para a predição de valores futuros de séries de dados temporais pode ser uma ferramenta de grande valia para a avaliação de riscos em barragens. No entanto, como mencionado, a existência de erros de medição no procedimento e na instrumentação utilizada pode distorcer o resultado da medição de tal modo que o valor do mensurando pode ser conhecido somente por meio de estimativas. Desse modo, a descaracterização dos valores verdadeiros pelo processo de medição pode interferir significativamente nas ferramentas de análises dos dados.

A simulação consiste em representar o comportamento de um processo, fenômeno ou sistema sem experimentação. Desse modo, neste capítulo é apresentada uma proposta para a construção de um modelo de simulação para reproduzir as influências provocadas pela incerteza de medição e pela frequência de amostragem dos dados na capacidade de predição do algoritmo k NN - TSP.

4.1 Modelo de Simulação

Neste trabalho é proposta a avaliação da capacidade de predição do algoritmo k NN - TSP sob diferentes condições de incerteza de medição e de frequência de amostragem dos dados. Desse modo, foi elaborado um modelo de simulação para possibilitar a propagação da incerteza de medição dos dados da série através do algoritmo de previsão k NN-TSP. Na Figura 9 é apresentada uma representação esquemática geral do método.

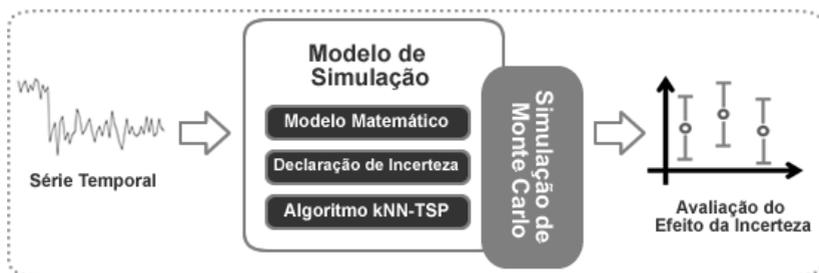


Figura 9: Esquema geral do método.

A utilização da SMC no modelo de simulação baseia-se na geração e na aplicação de erros de medição sobre uma série temporal em análise. Esses

erros de medição são simulados a partir de uma dada declaração da incerteza de medição do processo de medição que originou os dados da série temporal.

Como resultado desse processo, tem-se um conjunto de séries temporais contaminadas com erros de medição, todas elas consistentes com a série temporal original e com a declaração de incerteza. Para cada série desse conjunto é aplicado o algoritmo *k*NN-TSP para a previsão de valores futuros.

A avaliação do efeito da incerteza de medição sobre a capacidade do algoritmo *k*NN-TSP é realizada por meio da análise da dispersão dos valores provenientes das repetidas previsões realizadas.

4.2 Modelo Matemático de Medição

O erro de medição é composto por uma parcela aleatória e outra parcela sistemática, sendo que cada uma dessas componentes interfere de forma diferenciada no processo de medição. Desse modo, o modelo matemático de medição que será utilizado neste trabalho permite simular o comportamento metrológico de um sistema de medição por meio da geração de múltiplas características de erros sistemáticos e de erros aleatórios. Os erros gerados devem ser consistentes com uma determinada declaração de incerteza de medição.

Nesse modelo, o resultado da medição será dado pela adição do valor verdadeiro do mensurando com o valor do erro de medição como apresentado na Equação 4.1.

$$Y = X_f + (E_{ale} + E_{sis}) \quad (4.1)$$

Onde:

- Y é uma variável aleatória correspondente aos valores medidos registrados na série temporal;
- X_f é a variável que representa o valor do mensurando em uma determinada frequência de amostragem f ;
- E_{ale} é a parcela aleatória do erro;
- E_{sis} representa a parcela sistemática do erro.

4.2.1 Modelo para Distribuição do Erro Aleatório

Na modelagem realizada nesse estudo, o erro aleatório foi considerado estável em relação ao tempo e descrito por uma distribuição normal, como

apresentado na Equação 4.2. Por meio da função *Norm* são gerados valores aleatórios normalmente distribuídos com média $\mu = 0$ e desvio padrão σ .

$$E_{ale} \approx Norm(0, \sigma) \quad (4.2)$$

Como mencionado, o erro aleatório pode ser estimado como uma fonte de incerteza do tipo A ou também como uma fonte de incerteza do tipo B com distribuição normal em situações nas quais não seja possível realizar experimentos sobre o sistema de medição. Desse modo, a incerteza padrão para o erro aleatório u_{ale} é simulada a partir da declaração de incerteza segundo a igualdade:

$$u_{ale} = \sigma \quad (4.3)$$

4.2.2 Modelo para Distribuição do Erro Sistemático

No modelo de simulação elaborado neste trabalho, o erro sistemático é considerado desconhecido. Nesse caso é utilizada uma distribuição retangular para descrever esse comportamento Equação 4.4.

$$E_{sis} \approx Ret(-a, +a) \quad (4.4)$$

Onde, *Ret* é uma função geradora de valores aleatórios seguindo uma distribuição retangular, a qual apresenta como limite inferior $-a$ e como limite superior $+a$.

Desse modo, a incerteza padrão correspondente ao erro sistemático u_{sis} pode ser descrita pela Equação 4.5.

$$u_{sis} = \frac{a}{\sqrt{3}} \quad (4.5)$$

4.2.3 Modelo para a Incerteza Padrão Combinada e Incerteza Expandida

O modelo estabelecido para representar a incerteza padrão combinada e incerteza expandida do processo de medição simulado baseou-se no uso do método clássico. Nesse caso, considerou-se que as componentes de incerteza u_{ale} e u_{sis} são independentes de modo que a incerteza padrão combinada u_y é descrita pela Equação 4.6.

$$u_y = \sqrt{u_{ale}^2 + u_{sis}^2} \quad (4.6)$$

Para a incerteza expandida $U_{95\%}$ definida pela Equação 4.7 foi con-

siderado um intervalo de abrangência correspondente a 95,5% de modo que $k = 2$.

$$U_{95\%} = k \cdot u_Y \quad (4.7)$$

4.3 Etapas do Método de Avaliação

O método proposto neste trabalho para avaliar a influência da incerteza de medição na capacidade de predição do algoritmo k NN-TSP consiste em analisar quatro cenários com distintas contribuições do erro aleatório e do erro sistemático à incerteza. Os cenários são caracterizados conforme apresentado na Tabela 1.

Tabela 1: Cenários para a composição da incerteza de medição.

Cenário	a	σ
1	0%	0%
2	100%	0%
3	0%	100%
4	50%	50%

Nesse contexto, a avaliação da adequabilidade do algoritmo k NN-TSP para cada cenário descrito, consiste fundamentalmente na aplicação do método de Simulação de Monte Carlo sobre o modelo de simulação para uma determinada frequência de amostragem dos dados. De modo geral, pode-se estruturar a aplicação do método de avaliação em três principais etapas:

1. Geração de erros de medição;
2. Previsão de dados;
3. Avaliação do efeito da incerteza.

Na Figura 10 é apresentada uma representação esquemática das três etapas de aplicação do método.

Na Etapa 1, inicialmente é realizada uma avaliação de incertezas sobre o processo de medição utilizado para gerar a série temporal. Posteriormente, são gerados erros de medição para as componentes aleatória e sistemática de modo que esses erros sejam consistentes com a declaração de incerteza. Desse modo, para cada valor correspondente a um instante t na série temporal

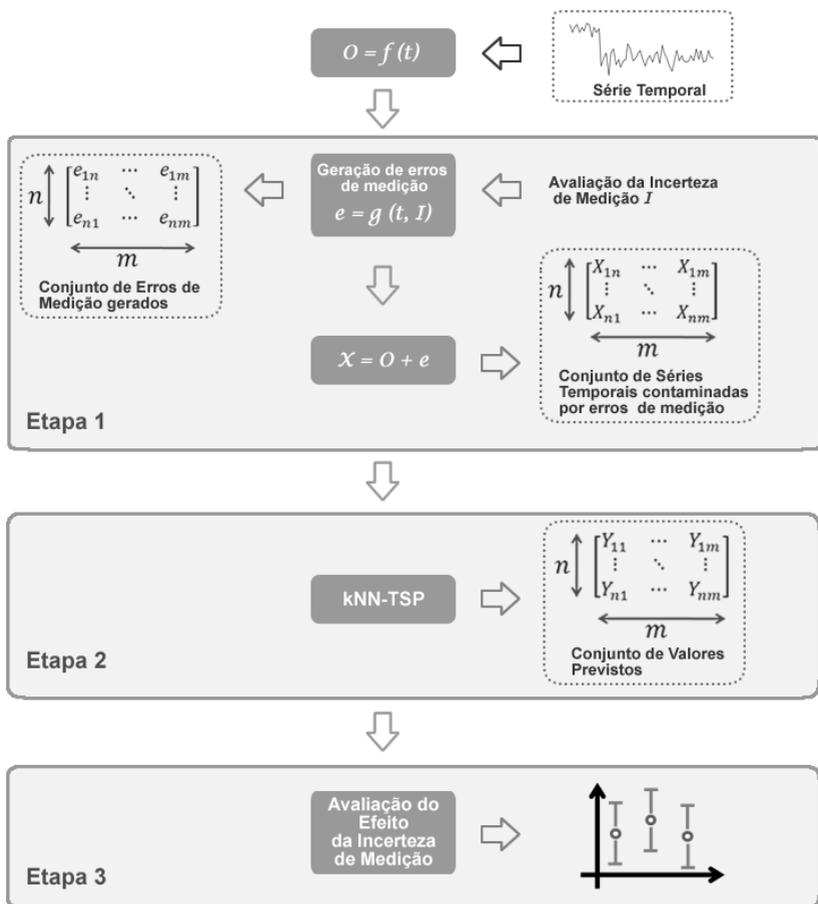


Figura 10: Etapas do método de avaliação.

avaliada, é realizada a contaminação deste valor pelo erro de medição gerado seguindo o modelo matemático de medição proposto na Equação 4.1.

Analisando o comportamento das séries temporais, pode ser inferido que erros sistemáticos cujo valor permanece constante desde o início da série até a realização da previsão, não configuram um caso de interesse para este trabalho, pois resultam apenas em um deslocamento da série, sem efeito sobre a operação do algoritmo *kNN-TSP*. Adicionalmente, esse modelo não corres-

ponde à realidade, pois sistemas de medição são recalibrados e ajustados ao longo de sua vida útil, de tal modo que o valor residual é alterado, mesmo que a incerteza que descreve tal contribuição permaneça constante. Nessa proposta de avaliação, o erro sistemático gerado permanece constante por intervalos expressivos na série temporal. Posteriormente, um novo valor para o erro sistemático é gerado com base no modelo estabelecido na Equação 4.4. O ponto de alteração do erro sistemático na série temporal é escolhido aleatoriamente uma única vez e é utilizado para o restante dos valores.

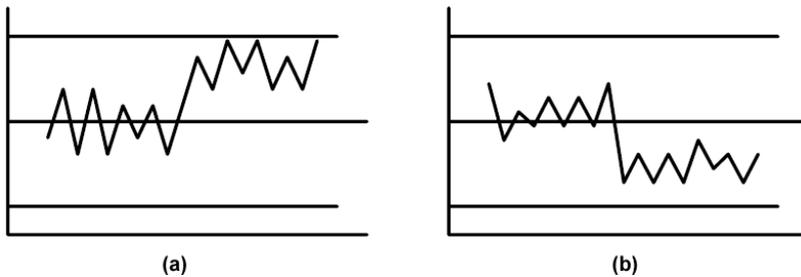


Figura 11: Simulação de mudança do erro sistemático.

Na Figura 11 é apresentada uma representação esquemática da mudança do erro sistemático em duas situações. Na situação da Figura 11(a) a tendência da série é deslocada para valores superiores em relação aos que estavam sendo medidos, e na situação representada pela Figura 11(b) a tendência é deslocada para valores inferiores.

Os valores para os erros aleatórios são gerados para cada instante t da série temporal com base no modelo estabelecido na Equação 4.2.

Uma representação esquemática da aplicação da SMC para a geração de erros sistemáticos e aleatórios em séries temporais é apresentada na Figura 12. Dada uma série temporal denotada por $X_t = (X_1, \dots, X_N)$, cada observação X_t da série será contaminada M vezes por uma quantidade de erro sistemático e de erro aleatório.

A quantidade correspondente à parcela do erro aleatório será determinada M vezes para cada série, ou seja, em cada observação X_t da série um valor distinto da componente aleatória do erro será gerada. Em relação à parcela do erro sistemático, esta será gerada no início da execução da simulação e posteriormente será alterada uma única vez ao longo das M execuções do algoritmo.

Nesse sentido, distintas configurações do erro de medição para a

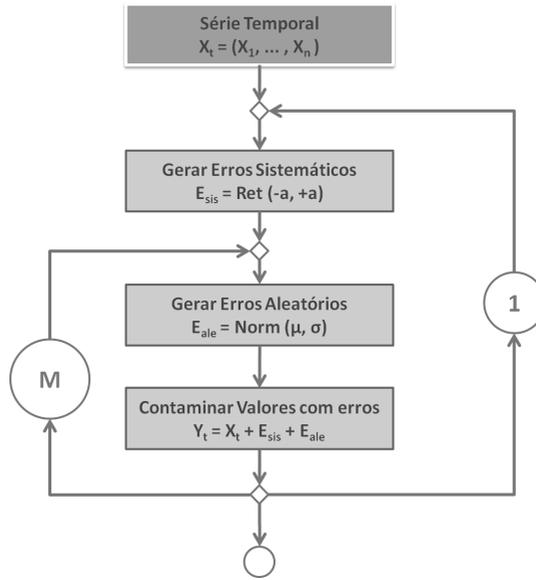


Figura 12: Fluxograma de geração de erros.

composição da incerteza de medição são analisadas. Como resultado dessa etapa, tem-se um conjunto de séries temporais afetadas com erros de medição.

Na Figura 13 é apresentada uma representação esquemática como exemplo de contaminação de uma série temporal por erros de medição. A série de dados descrita em azul representa a série de referência, que atua como pivô para a construção das séries subsequentes. O conjunto de pontos em vermelho representa uma série resultante da aplicação de erros sobre a série de referência.

Na Etapa 2, sobre as séries temporais contaminadas por erros de medição, é realizada a aplicação do algoritmo k NN - TSP para a previsão dos valores futuros.

O procedimento realizado nessa etapa é descrito no fluxograma representado na Figura 14. Para cada série temporal do conjunto de M séries produzidas na Etapa 1 é realizado um processo de amostragem dos valores dessas séries com base em F frequências de amostragem. Desse modo, para cada série amostrada é aplicado o algoritmo de previsão N vezes, as quais correspondem à quantidade de valores futuros a serem analisados.

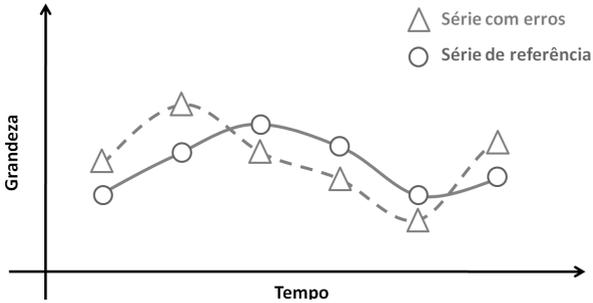


Figura 13: Série de referência e uma série obtida por simulação dos erros de medição.

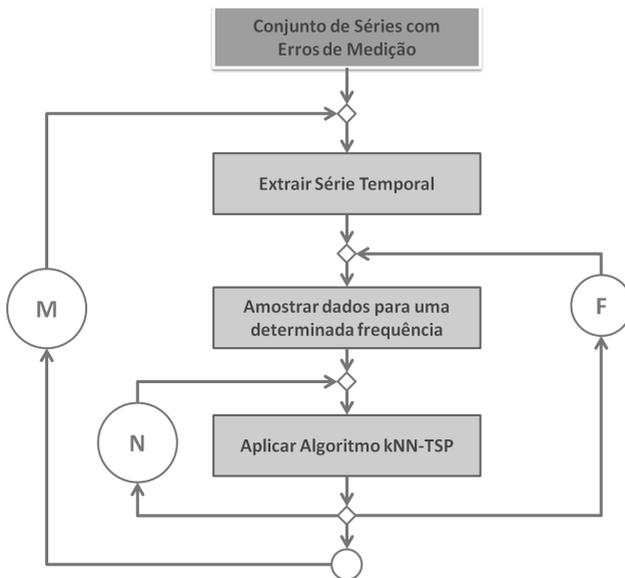


Figura 14: Fluxograma de previsão em séries temporais.

Como resultado dessa etapa, tem-se um conjunto de pontos previstos, cada qual com uma dispersão associada, devido aos diferentes valores obtidos para cada ponto. A Figura 15 contém uma representação esquemática, que permite visualizar a dispersão dos valores previstos para uma série temporal

exemplo em um instante futuro (nesse caso $N = 1$).

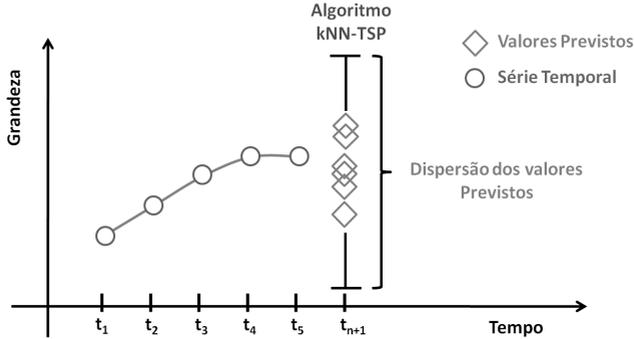


Figura 15: Dispersão dos valores previstos.

Na Etapa 3, o conjunto de N valores previstos pelo algoritmo é analisado por meio da dispersão produzida. Desse modo, segundo a SMC os parâmetros média e desvio padrão podem ser tratados como a estimativa do valor da grandeza de saída e a incerteza padrão combinada, respectivamente. Com base nesses parâmetros, os cenários definidos para a incerteza de medição são avaliados em função do impacto sobre a incerteza de medição resultante da utilização do algoritmo $kNN - TSP$.

O resultado produzido pelo algoritmo representado no fluxograma da Figura 14 pode ser entendido como o conjunto de matrizes descritas na Figura 16. Para cada frequência F de amostragem dos dados é gerado um conjunto de valores conforme descrito nessas matrizes. A partir das M simulações realizadas, para cada ponto previsto $i = (1, \dots, N)$ da série tem associado os parâmetros de média μ_{S_i} e desvio padrão σ_{S_i} , os quais são calculados conforme as Equações 4.8 e 4.9, respectivamente.

$$\mu_{S_i} = \frac{1}{M} \cdot \sum_{j=1}^M Y_{ij} \quad (4.8)$$

$$\sigma_{S_i} = \sqrt{\frac{1}{M-1} \cdot \sum_{j=1}^M (Y_{ij} - \mu_{S_i})^2} \quad (4.9)$$

Como indicadores para avaliar o efeito da incerteza de medição sobre os valores previstos para cada um dos distintos C cenários de erros

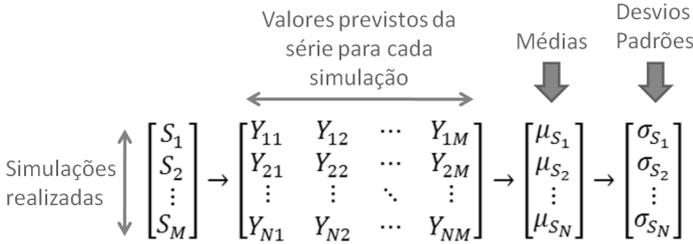


Figura 16: Procedimento de análise dos valores previstos.

estabelecidos são utilizados os descritores σ_C (Equação 4.10) e $U_{95\%}(\sigma_C)$ (Equação 4.11). O descritor σ_C representa a média dos desvios padrões obtidos em cada simulação. Esse parâmetro foi estimado em uma experimentação prévia, por meio da qual constatou-se que a distribuição de saída das previsões do algoritmo kNN - TSP seguia um comportamento normal.

$$\sigma_C = \frac{1}{N} \cdot \sum_{i=1}^N \sigma_{S_i} \quad (4.10)$$

$$U_{95\%}(\sigma_C) = 2 \cdot \sigma_C \quad (4.11)$$

Em relação ao cenário 1, no qual não é considerado o efeito da incerteza de medição, a avaliação realizada refere-se à análise da qualidade do ajuste dos modelos de previsão aos dados da série. Desse modo, é utilizada uma medida frequentemente utilizada na literatura para a estimativa da qualidade da previsão de valores, denominada Erro Médio Absoluto (EMA) (FERRERO, 2009; HYNDMAN; KOEHLER, 2006). Essa medida permite calcular o erro de previsão a partir das diferenças entre os valores previstos e observados. O cálculo do EMA é definido pela Equação 4.12.

$$EMA = \frac{\sum_{i=1}^N |Y_N - X_N|}{N} \quad (4.12)$$

4.4 Avaliação do Ambiente de Simulação

Com o objetivo de avaliar a viabilidade do método proposto, foi realizada uma aplicação do método utilizando uma série temporal artificial. Os modelos construídos com o algoritmo kNN - TSP foram analisados em comportamentos característicos de séries temporais, como tendência e sazo-

nalidade. Para isso, o algoritmo k NN - TSP proposto em Ferrero (2009) foi adaptado para a aplicação das etapas do método proposto.

O desenvolvimento do ambiente de simulação foi realizado por meio do uso da linguagem computacional R¹ (R Development Core Team, 2009). Essa linguagem de programação apresenta um grande conjunto de bibliotecas para a construção de sistemas de modelagem e simulação.

4.4.1 Séries Temporais Artificiais

Inicialmente foram utilizadas séries de dados temporais artificiais para simular comportamentos como tendência e sazonalidade. Nesse estudo, foi utilizada uma série proposta em Kulesh, Holschneider e Kurennaya (2008) denominada Série Temporal de Dependência Sazonal. Os valores que compõem essa série são gerados por meio de uma função definida pela Equação 4.13, a qual considera sazonalidade constante e tendência linear.

$$ST(t) = \cos\left(\frac{t}{25}\right) \times \sin\left(\frac{t}{100}\right) + \frac{t}{1000} + 1, t \in [0, N] \quad (4.13)$$

A série temporal gerada por essa equação pode ser observada na Figura 17, na qual é possível verificar a variação da tendência ao longo do tempo e também a sazonalidade constante.

Nessa abordagem, por meio da análise de séries temporais artificiais, o valor do mensurando é considerado conhecido. Desse modo, é possível avaliar de modo completo o efeito da incerteza de medição, pois também torna-se conhecido o erro sobre os valores previstos.

4.4.2 Configurações dos Experimentos

No contexto da aplicação do método proposto sobre uma série temporal artificial foi realizada uma avaliação de incertezas considerando um caso pessimista. Nesse caso, a incerteza máxima foi estabelecida em 1/3 de metade da amplitude pico-a-pico da série temporal. Desse modo, considerando as Equações 4.6 e 4.7 foi estabelecida uma relação entre a amplitude pico-a-pico do sinal (AS) e as contribuições aleatória u_{ale} e sistemática u_{sis} definidas nas Equações 4.3 e 4.5, respectivamente. Essa relação é expressa pela Equação 4.14.

¹R version 2.9.0 - Copyright (C) 2009 The R Foundation for Statistical Computing. This software is distributed under the terms of the GNU General Public License Version 2, June 1991. <http://www.r-project.org/>

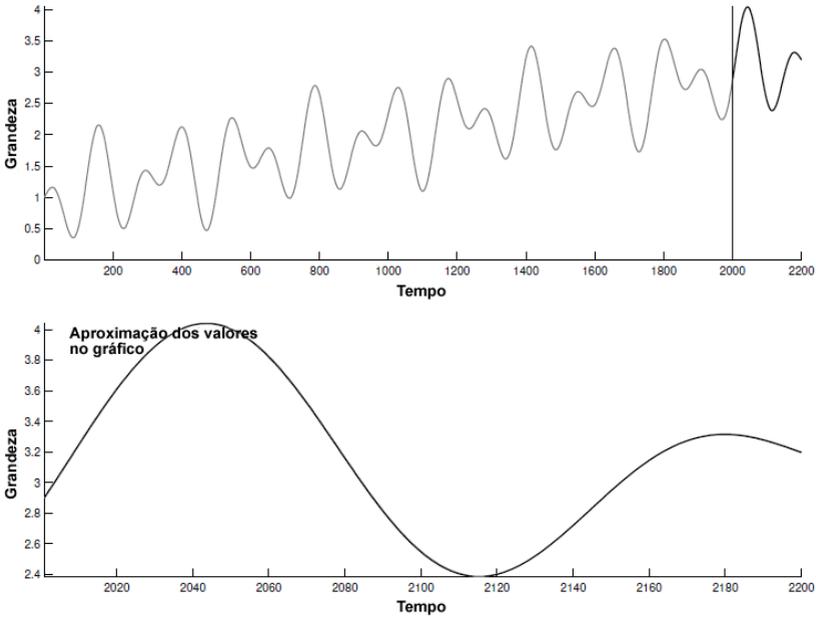


Figura 17: Série Temporal de Dependência Sazonal (KULESH; HOLSCHNEIDER; KURENNAYA, 2008).

$$U_{95\%} = \frac{1}{3} \cdot \frac{AS}{2} = 2 \cdot \sqrt{(E_{ale})^2 + \left(\frac{E_{sis}}{\sqrt{3}}\right)^2} \Leftrightarrow E_{ale}^2 + \frac{E_{sis}^2}{3} = \left(\frac{AS}{12}\right)^2 \quad (4.14)$$

Com base nessa equação e considerando os cenários de composição da incerteza de medição estabelecidos para avaliação, na Tabela 1 foram calculados os parâmetros de contribuição do erro sistemático e do erro aleatório, os quais são apresentados na Figura 18.

A amplitude da série de dependência sazonal foi estimada em $AS = 2$ e os parâmetros σ e a foram determinados para a aplicação dos modelos de distribuição do erro aleatório e do erro sistemático. Os valores correspondentes são apresentados na Tabela 2.

A série temporal artificial foi gerada como descrito na Equação 4.13 para um conjunto de $N = 2500$ pontos. O período da série reservado para a

realização das previsões foi constituído a partir dos 250 pontos finais da série, referente aos instantes $t = (2251, \dots, 2500)$. Desse modo, os 2250 valores iniciais da série foram utilizados para compor o conjunto de treinamento do algoritmo de previsão.

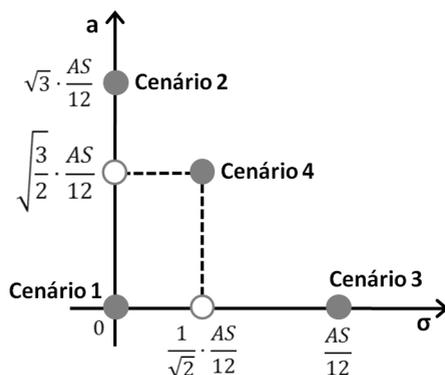


Figura 18: Contribuições aleatória e sistemática à incerteza para diferentes cenários.

Como mencionado no capítulo 2, o desempenho do algoritmo k NN - TSP depende de alguns parâmetros. A medida de similaridade utilizada para o algoritmo foi a distância Euclidiana. O conjunto de vizinhos próximos foi selecionado utilizando $k = 1$ e a função de previsão utilizada foi a Média de Valores Relativos (FERRERO, 2009). Conforme recomendado em Kulesh, Holschneider e Kurennaya (2008) para a Série Temporal de Dependência Sazonal foi adotada uma janela de tamanho $w = 100$ para a busca das subsequências na série.

Tabela 2: Valores dos parâmetros para a composição da incerteza de medição.

Cenário	a	σ
1	0,00	0,00
2	0,28	0,00
3	0,00	0,16
4	0,20	0,11

A aplicação do método de SMC foi realizada para um número de simulações $M = 10^4$ nas etapas de geração de erros de medição e de pre-

visão de dados. Essa mesma quantidade de simulações foi utilizada em cada um dos cenários para a composição da incerteza de medição e também para cada frequência de amostragem $F = (5\%, 10\%, 20\%, 50\%, 100\%)$ dos dados. Assim, a quantidade de valores a serem previstos e de pontos da série a serem utilizados como conjunto de treinamento, variam de acordo com a frequência de amostragem F . Esses valores estão descritos na Tabela 3.

Tabela 3: Quantidade de pontos (P) x Frequência de amostragem (F).

P/F	100%	50%	20%	10%	5%
Conjunto de treinamento	2250	1125	450	225	113
Período de previsão	250	125	50	25	12
Total de pontos	2500	1250	500	250	125

Na Figura 19 são apresentados os gráficos correspondentes às distintas frequências de amostragem dos dados analisadas. O trecho da série apresentado em cada um dos gráficos refere-se ao período utilizado para a previsão de valores.

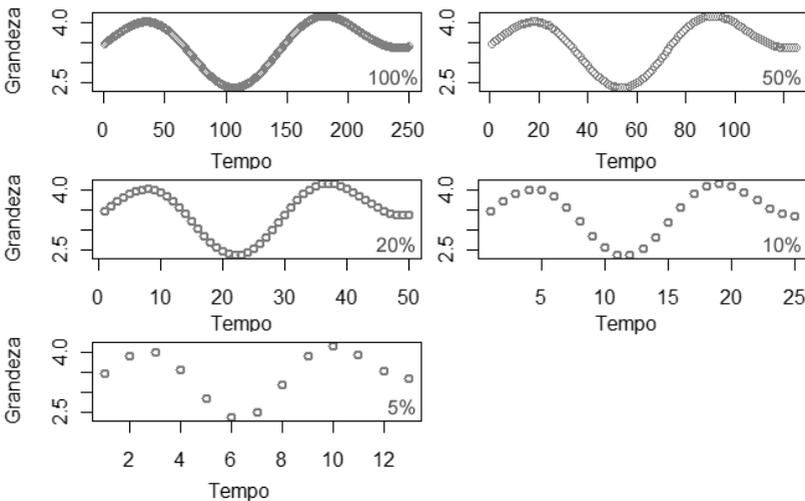


Figura 19: Frequências de amostragem analisadas.

4.4.3 Resultados

Nesta subseção são apresentados os resultados da aplicação do método proposto, para a avaliação do efeito da incerteza de medição sobre o algoritmo k NN - TSP em séries temporais artificiais. Os algoritmos desenvolvidos para cada uma das etapas do método de avaliação foram aplicados segundo as configurações descritas na subseção anterior.

Na Etapa 1 do método, os erros sobre a série artificial foram gerados conforme o modelo matemático estabelecido e os valores descritos na Tabela 2. Desse modo, distintas séries temporais contaminadas com erros de medição foram geradas a partir dessa série de referência.

Nas Figuras 20, 21 e 22 são apresentados gráficos exemplos do procedimento realizado para a avaliação do cenário 2, 3 e 4 do método proposto, respectivamente. As linhas coloridas na figura representam as séries geradas com erros de medição e a linha em cor preta representa a série de referência. Esses gráficos foram retirados do conjunto de experimentos de cada um dos cenários 2, 3 e 4, para os quais foi utilizada uma frequência de amostragem de 100% dos dados da série, totalizando 2500 pontos. A situação que pode ser observada nos gráficos refere-se ao instante dos experimentos no qual o número de simulações é $M = 100$.

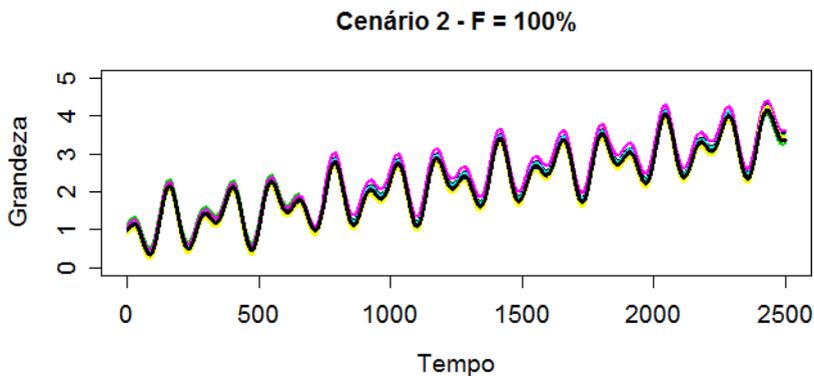


Figura 20: Exemplo de séries com erros para o cenário 2.

No cenário 2, como pode-se observar na Figura 20, a série temporal contaminada com erros, representada pela linha em cor rosa, apresenta um desvio sistemático que a mantém próxima aos valores da série de referência. No entanto, entre os instantes $t = 500$ e $t = 1000$ a componente de erro sis-

temático é atualizada de modo aleatório pelo algoritmo. Desse modo, o desvio entre a série com erros e a série de referência torna-se maior e é mantido até o final da série.

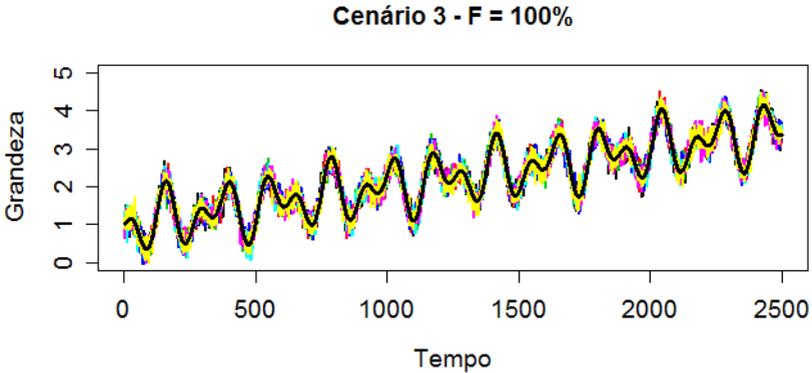


Figura 21: Exemplo de séries com erros para o cenário 3.

Na Figura 21, a qual descreve o comportamento das séries para o cenário 3, pode-se observar que as séries contaminadas com erros de medição variam de modo aleatório, mas sempre em torno da série de base. Esse comportamento é explicado devido ao fato de que no cenário 3, a incerteza de medição é composta em totalidade pelo erro aleatório.

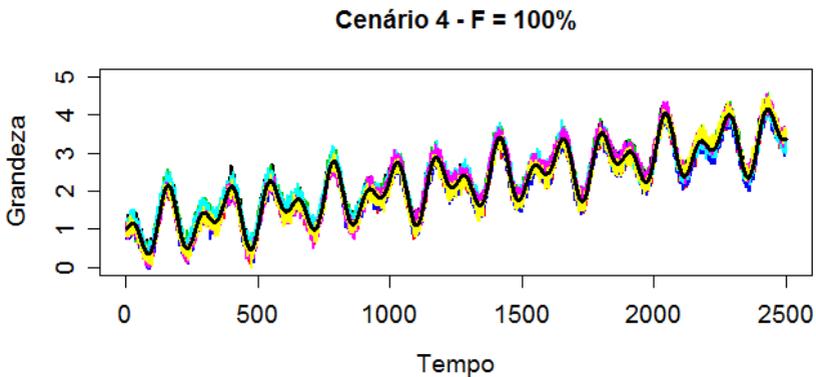


Figura 22: Exemplo de séries com erros para o cenário 4.

O comportamento das séries contaminadas por erros de medição conforme o cenário 4 é influenciado pela combinação das componentes de erro aleatório e de erro sistemático. Desse modo, pode-se observar na Figura 22, que a série com erros representada pela linha em cor rosa, possui um comportamento aleatório em torno da série base até pouco antes do instante $t = 1000$. Após esse instante, é aplicado um desvio sistemático que a distancia da série base e a mantém com esse desvio até o final da série. Outro exemplo é a série descrita no gráfico em cor amarela, a qual apresenta um desvio inferior dos seus valores, em relação à série base, até próximo do instante $t = 2000$. Posterior a esse instante, a série apresenta um desvio superior aos valores da série base.

Como mencionado, para o cenário 1, devido ao fato de que os parâmetros são $\sigma = 0$ e $a = 0$, não é realizada a geração de erros sobre a série base.

Na Etapa 2 do método, o algoritmo kNN - TSP, por meio dos parâmetros descritos na subseção anterior, foi utilizado para a previsão de valores futuros sobre as séries contaminadas com erros de medição. Nesse contexto, para uma melhor observação do comportamento dos valores previstos, foram extraídas do conjunto de experimentos algumas situações representativas dessa etapa.

Na Figura 23, é apresentada a série de base descrita em cor preta e o conjunto de valores previstos pelo algoritmo kNN - TSP, que é descrito pelas linhas em cor azul. A situação apresentada nessa figura corresponde às previsões realizadas para o cenário 3 com frequência de amostragem dos dados $F = 10\%$ utilizando um número de simulações $M = 100$. A linha em cor amarela representa a média dos M valores previstos para cada ponto. Pode-se observar nesse caso, que grande parte das previsões realizadas pelo algoritmo não correspondem aos valores da série base.

Outro comportamento da aplicação da Etapa 2 do método é apresentado na Figura 24. Nesse caso, as previsões foram realizadas também para o cenário 3, no entanto, utilizando uma frequência de amostragem dos dados $F = 20\%$ e um número de simulações $M = 5$. No gráfico apresentado na figura, a linha em cor azul representa os valores da série base para o período de previsão utilizado. A linha em cor amarela corresponde aos valores previstos pelo algoritmo kNN - TSP sobre os dados da série base, ou seja, sem influência de erros. Por meio da análise dessa linha no gráfico pode-se observar que as previsões, mesmo realizadas sobre uma série sem erros, se afastam da série base para todos os pontos.

A linha apresentada em cor vermelha no gráfico refere-se ao valor

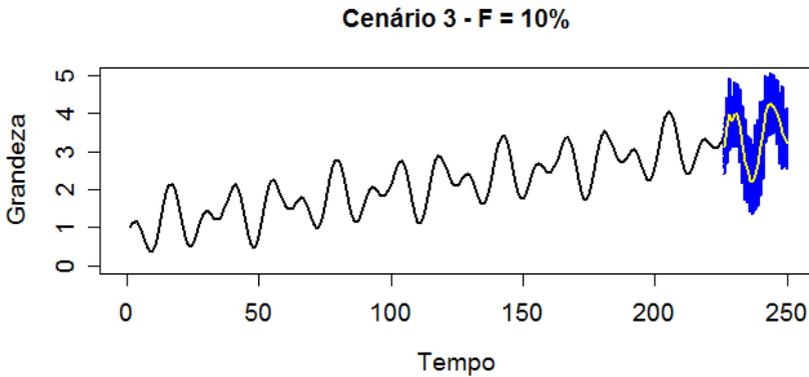


Figura 23: Conjunto de valores previstos para o cenário 3.

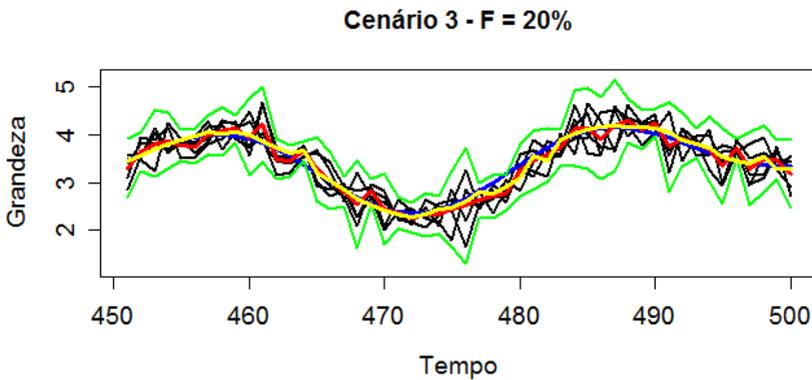


Figura 24: Exemplo de séries previstas para o cenário 3.

médio previsto a partir das cinco simulações realizadas para cada instante da série. As séries geradas por aplicação de erros de medição sobre a série base são representadas pelas linhas em cor preta. Desse modo, é possível observar para cada uma das cinco séries geradas com erros, os distintos valores previstos para cada instante do trecho analisado no gráfico. As linhas definidas em cor verde representam o intervalo de confiança (com nível de confiança de 95,5%) para cada ponto previsto.

Conforme pode ser observado na Figura 24, as previsões realizadas

pelo algoritmo kNN - TSP seguem o mesmo comportamento aleatório, em torno da série base, que as séries com erros utilizadas para o conjunto de treinamento. De modo análogo, esse comportamento pode ser observado por meio da análise do gráfico apresentado na Figura 25. Nesse caso é representado no gráfico uma situação do cenário 2, no qual a única fonte de incertezas é proveniente de erros sistemáticos. Desse modo, observa-se que os valores previstos apresentam um desvio sistemático em relação à série base, assim como as séries utilizadas para as previsões.

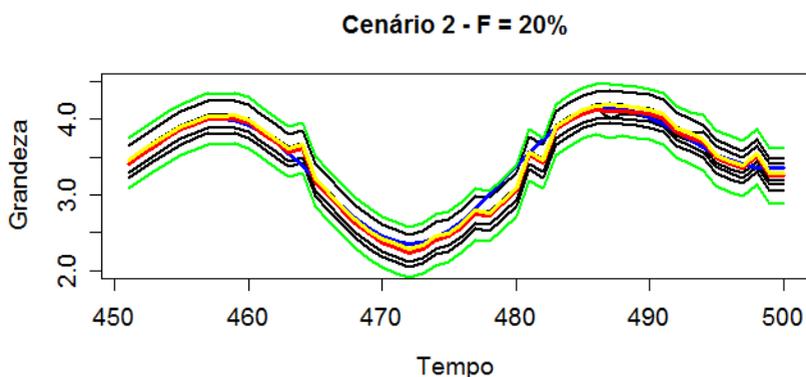


Figura 25: Exemplo de séries previstas para o cenário 2.

Na Etapa 3, a partir do conjunto de séries geradas pelas previsões, resultantes das M simulações realizadas em cada um dos cenários estudados e para cada frequência de amostragem dos dados, são calculados os parâmetros de avaliação propostos nesse trabalho.

O parâmetro de avaliação utilizado para as análises do cenário 1 é o EMA descrito na Equação 4.12. Como mencionado, a avaliação da incerteza dos valores previstos para os cenários 2, 3 e 4 é definida pela Equação 4.11.

Nas Figuras 26, 27 e 28 são apresentados os resultados das $M = 10^4$ simulações realizadas para os cenários 2, 3 e 4, respectivamente. Em cada figura são descritos cinco gráficos, os quais correspondem às frequências de amostragem utilizadas. Nos gráficos, são representados pela linha em cor preta, os valores da série base; os valores previstos sobre a série base, pela linha em cor amarela; o valor médio dos valores previstos a partir das M simulações realizadas, pela linha em cor vermelha e o intervalo de confiança para cada valor previsto, pelas linhas em cor azul.

Os resultados do cálculo desses parâmetros sobre os valores das séries

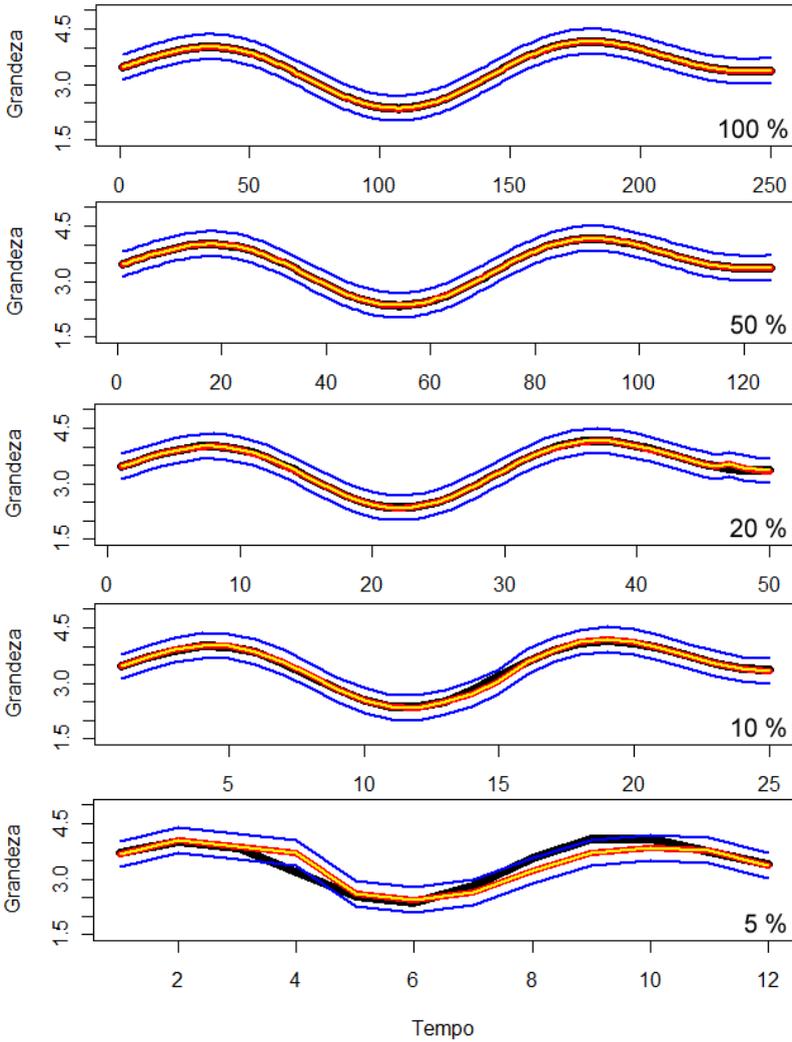


Figura 26: Valores previstos pelo algoritmo k NN - TSP considerando os parâmetros definidos para o cenário 2 e as frequências de amostragem dos dados $F = (5\%, 10\%, 20\%, 50\%, 100\%)$.

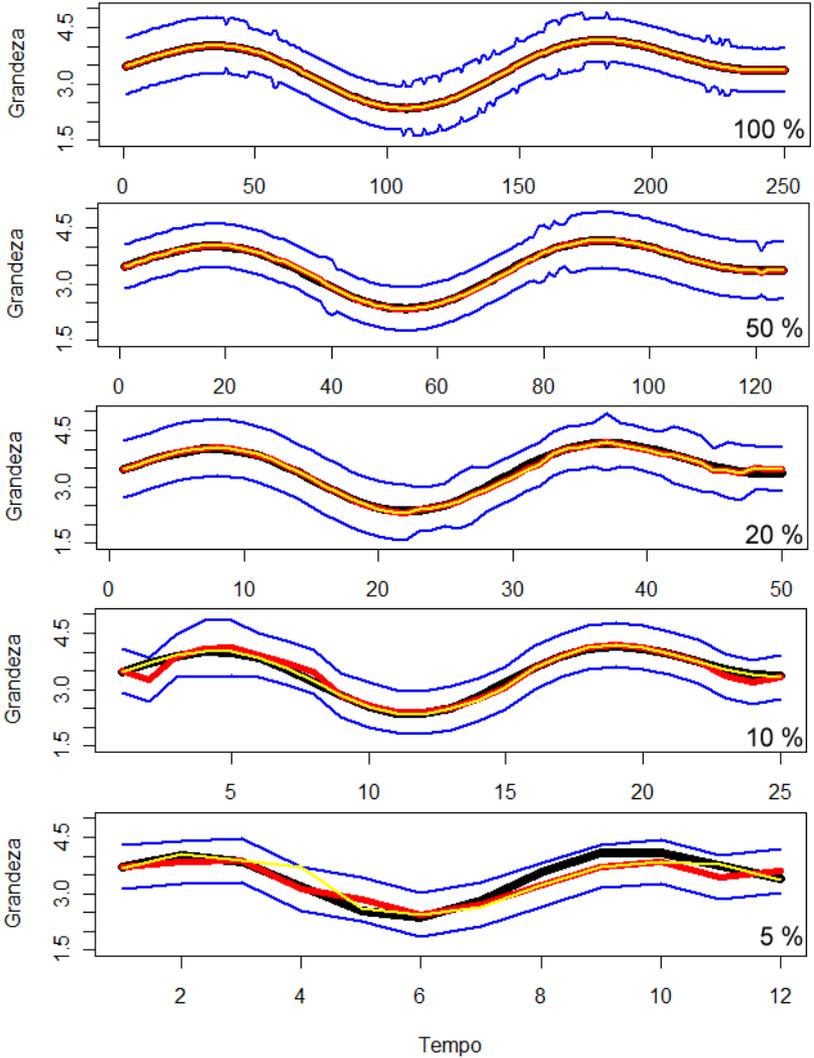


Figura 27: Valores previstos pelo algoritmo k NN - TSP considerando os parâmetros definidos para o cenário 3 e as frequências de amostragem dos dados $F = (5\%, 10\%, 20\%, 50\%, 100\%)$.

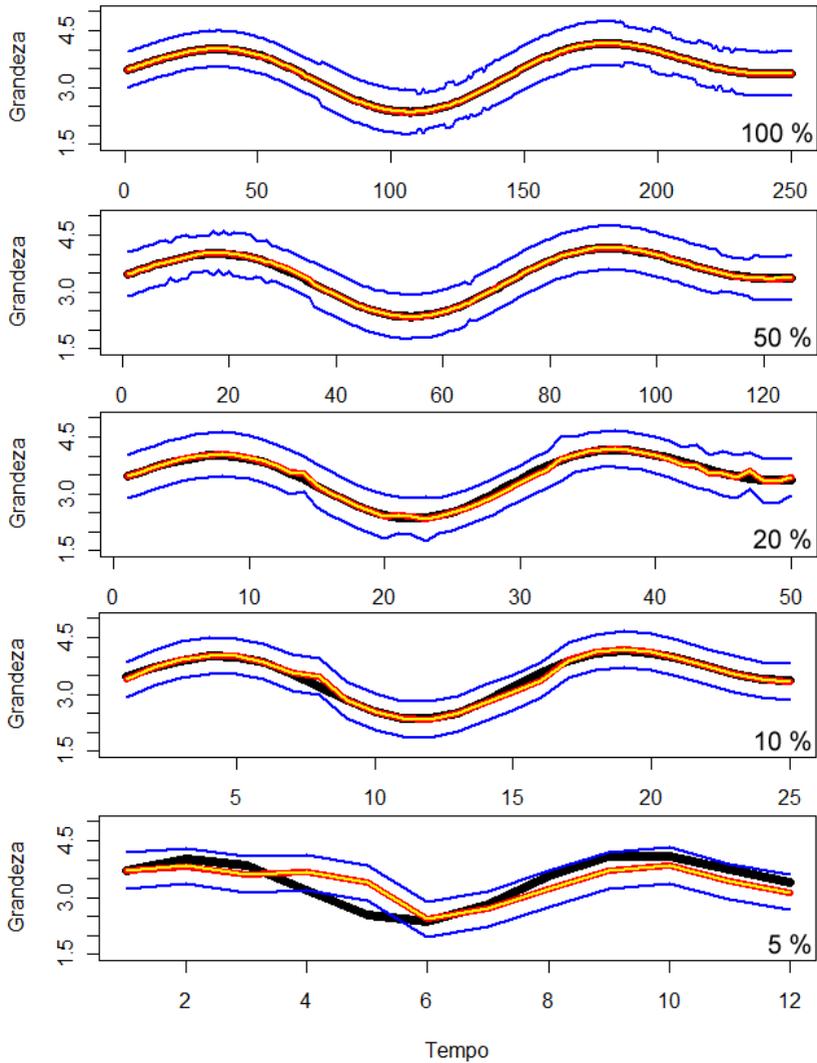


Figura 28: Valores previstos pelo algoritmo k NN - TSP considerando os parâmetros definidos para o cenário 4 e as frequências de amostragem dos dados $F = (5\%, 10\%, 20\%, 50\%, 100\%)$.

previstas em todos os cenários são apresentados nas Tabelas 4 e 5 para os valores de EMA no cenário 1 e para os valores previstos em cada frequência de amostragem, respectivamente.

Tabela 4: Valores de EMA x Frequência de amostragem dos dados.

	100%	50%	20%	10%	5%
EMA	0,001	0,013	0,036	0.081	0.186

Tabela 5: Avaliação da incerteza de previsão - Cenários de erros (C) x Frequência de amostragem dos dados (F).

C/F	100%	50%	20%	10%	5%
2	0,33	0,33	0,33	0,33	0,33
3	0,64	0,63	0,67	0,59	0,57
4	0,52	0,55	0,52	0,47	0,47

Os resultados descritos nas Tabelas 4 e 5 são também apresentados em gráficos nas Figuras 29 e 30. Na Figura 29 é descrito um gráfico com os valores de incerteza das previsões para cada uma das frequências analisadas. O gráfico descrito na Figura 30, refere-se aos valores de EMA calculados para cada uma das frequências de amostragem dos dados.

4.4.4 *Discussão*

O método de avaliação proposto neste trabalho baseia-se em um modelo de simulação para analisar a influência da incerteza de medição no desempenho do algoritmo *k*NN - TSP. Esse modelo permite a geração de erros de medição com diversas configurações de componentes aleatórias e sistemáticas do erro, compatíveis com uma dada declaração da incerteza de medição.

Nesse contexto, em alguns casos, a distribuição de um determinado parâmetro de entrada pode ser interpretada como representativa de um comportamento aleatório em medições sucessivas. Nessas situações, pode-se esperar que a cada nova medição realizada ocorra uma mudança aleatória do valor desse parâmetro. No entanto, em outros casos, a distribuição associada a um parâmetro de entrada não representa um comportamento aleatório em uma situação real (KONRATH, 2008).

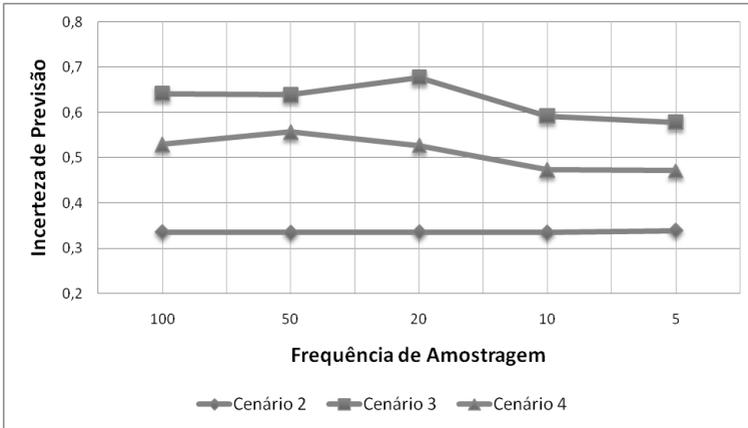


Figura 29: Incerteza de previsão - Cenário x Frequência de amostragem.

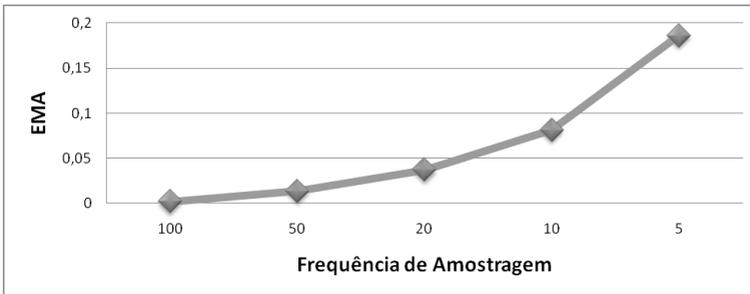


Figura 30: Valores de EMA para cada frequência de amostragem.

Desse modo, distintos cenários foram definidos nesse estudo com o objetivo de se avaliar diferentes combinações das contribuições à incerteza de medição. Assim, uma série temporal é afetada por um conjunto de erros que corresponderiam a um dos infinitos processos de medição, que são consistentes com a incerteza de medição declarada. Desse modo, a repetição desse procedimento utilizando outros valores de erros sistemáticos e de erros aleatórios permite obter valores distintos de média e de dispersão da grandeza de saída, de modo que sempre seja mantida a consistência com a declaração de incerteza estabelecida.

O cenário 1 corresponde a uma situação de medição perfeita, onde não

há a influência de erros sobre o resultado. Desse modo, a única fonte de erro em relação aos valores previstos está associada à qualidade do ajuste dos modelos construídos com o algoritmo k NN - TSP. No cenário 2, as contribuições à incerteza de medição são consideradas provenientes somente de componentes de erro sistemático de modo que fosse possível avaliar o impacto dessa componente sobre a incerteza dos valores previstos. Situação análoga foi analisada no cenário 3, no qual foi considerado somente o efeito das contribuições aleatórias. No cenário 4, as contribuições aleatória e sistemática foram analisadas de modo conjunto para avaliar o efeito da combinação de ambas sobre a incerteza de previsão.

Em relação à simulação da contribuição do erro sistemático para a incerteza de medição, foi elaborado um modelo para representar erros sistemáticos de valor desconhecidos que permanecem constantes durante certos períodos de tempo. Esse comportamento refere-se ao procedimento de ajuste periódico dos sistemas de medição durante as rotinas de produção ou monitoração. Desse modo, ajuste de um sistema de medição, pode resultar em um erro de *offset* residual de valor desconhecido devido à compensação incompleta dos efeitos sistemáticos.

Nesse estudo, a SMC foi aplicada sobre o modelo de simulação proposto para uma quantidade $M = 10^4$ de simulações. Esse número de simulações foi adotado nesse trabalho em virtude de aspectos relacionados à capacidade de processamento disponível e à complexidade do algoritmo k NN - TSP. Essas características devem-se ao custo de tempo para a geração de números aleatórios em cada ponto da série e ao processo de busca por sequências similares, em todo o conjunto de treinamento, para a determinação do valor futuro.

Por meio da análise dos resultados apresentados na subseção anterior é possível observar que o efeito da incerteza de medição com componente de erro puramente sistemática, representa o menor impacto na capacidade de predição do algoritmo k NN - TSP em relação aos cenários 3 e 4. Nesse caso, o nível de incerteza dos valores previstos é mantido para as diferentes frequências de amostragem dos dados utilizadas. Pode-se constatar também que a incerteza dos valores previstos no cenário 2 é muito próximo à declaração de incerteza utilizada para esta série artificial.

Em relação aos resultados encontrados para o cenário 3, observa-se que a incerteza de medição com componente de erro inteiramente aleatória foi responsável pelo maior nível de incerteza na previsão dos valores. Desse modo, a incerteza resultante dos valores previstos corresponde à quase o dobro da incerteza de medição declarada para a série artificial.

Pode-se notar também, que para os cenários 3 e 4, a incerteza de previsão dos valores apresentou significativa melhora nos casos de baixa frequência de amostragem dos dados. Esse comportamento deve-se ao fato de que, com uma frequência de amostragem menor dos dados, a variabilidade dos valores para a escolha do valor futuro pelo algoritmo k NN - TSP será reduzida. Desse modo, o resultado das previsões irá provocar uma menor dispersão dos valores e consequentemente uma redução da incerteza.

No entanto, mesmo que esses níveis de incerteza sejam menores para as frequências de amostragem reduzidas, é necessário observar que a qualidade de ajuste dos modelos do algoritmo k NN - TSP aos dados da série são os mais baixos para esses níveis de frequência de amostragem (Figura 30).

A partir da avaliação experimental realizada sobre a série artificial, pode-se concluir que o modelo de simulação proposto possibilita a propagação da incerteza de medição pelo algoritmo k NN - TSP e a avaliação do efeito de distintas contribuições à incerteza na capacidade de previsão do algoritmo.

4.5 Considerações Finais

Como mencionado, a avaliação da influência dos erros de medição sobre o desempenho das ferramentas de análise dos dados não é viável em situações reais, pois o valor do erro é desconhecido. Contudo, é possível afirmar que a incerteza de medição pode afetar a capacidade de ferramentas de apoio na monitoração de processos (KONRATH, 2008; SILVA, 2003).

Nesse contexto, por meio da utilização de diferentes níveis de incerteza de medição é possível simular o uso de diferentes sistemas de medição. Desse modo, torna-se possível um estudo comparativo para encontrar uma relação custo benefício de processos de medição e ferramentas de análise desses processos.

No próximo capítulo é apresentado um estudo de caso do método de avaliação proposto em séries temporais provenientes de processos de medição reais.

5 ESTUDO DE CASO

A medição dos deslocamentos horizontais da crista de uma barragem na supervisão do comportamento das estruturas consiste em um importante parâmetro para a avaliação de riscos. Essa grandeza é geralmente afetada pelos seguintes comportamentos:

- Deflexão da estrutura de concreto;
- Rotação da base da estrutura, devido à deformabilidade da fundação;
- Influências térmicas ambientais.

Desse modo, os deslocamentos são afetados pelas características do concreto e pelas propriedades do maciço rochoso de fundação e consistem em relevantes informações para a auscultação do comportamento da barragem e de sua fundação (SILVEIRA, 2003).

Em barragens de concreto, a monitoração de deslocamentos é tida como uma tarefa que exige elevada sensibilidade e exatidão, sendo normalmente realizada por meio da análise dos dados gerados pelos instrumentos como inclinômetros, extensômetros, medidores de juntas ou fissuras, estações geodésicas, pêndulos diretos e invertidos (BROWNJOHN, 2007).

A aplicação de pêndulos diretos e invertidos é realizada, pois consistem em instrumentos simples e confiáveis e que permitem monitorar conjuntamente deslocamentos relativos e absolutos de uma barragem (FEDERAL ENERGY REGULATORY COMMISSION - FERC, 2008).

Neste capítulo é apresentado o estudo de caso do método proposto no capítulo anterior, para aplicação na série histórica de medições da instrumentação de pêndulos, da barragem da Usina de Itaipu.

5.1 Usina Hidrelétrica de Itaipu - Pêndulos

A barragem de Itaipu, em virtude do porte das estruturas de concreto com blocos de até 196 m de altura e das descontinuidades geológicas que ocorriam na fundação, a instrumentação começou a ser instalada desde o início da concretagem (SILVEIRA, 2003).

No plano de instrumentação do maciço basáltico de fundação das estruturas foram selecionados os blocos F-13, F-14, F-19 F-20 e U9-A como blocos-chave, nos quais foi instalada a mais completa instrumentação da barragem.

Os deslocamentos cisalhantes da base da barragem são medidos por pêndulos invertidos, os quais são instalados aproveitando-se a disposição de furos de sondagem realizados na fase de projeto. O instrumento consiste na fixação de um fio de aço na parte mais profunda da sondagem, o qual é mantido tensionado verticalmente por meio de um flutuador imerso em um tanque de água, posicionado na interface fundação-estrutura (SILVEIRA, 2003; ITAIPIU BINACIONAL, 1992b). Na Figura 31(b) pode ser observada a representação esquemática e os elementos constituintes de um pêndulo invertido.

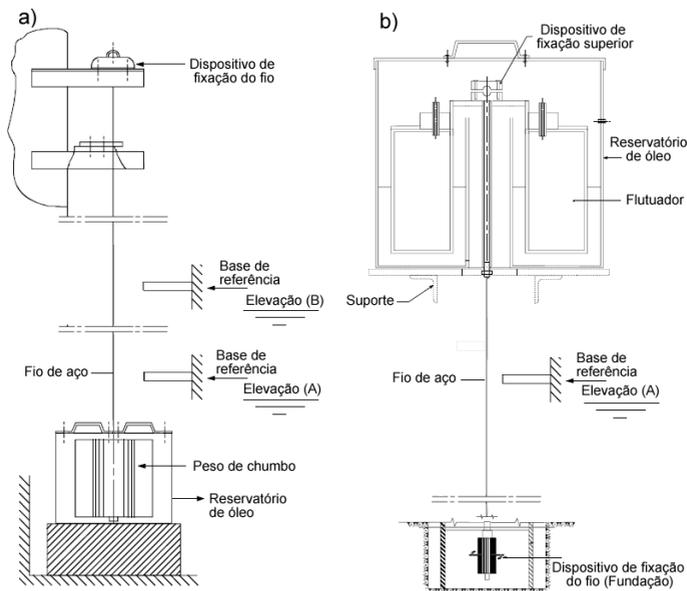


Figura 31: Representação esquemática de: a) pêndulo direto e b) pêndulo invertido (ITAIPIU BINACIONAL, 1992b).

Os deslocamentos horizontais da crista são medidos por meio de pêndulos diretos, os quais são instalados durante o período construtivo, a partir de tubulações posicionadas verticalmente nas estruturas da barragem. Esse instrumento consiste em um fio de aço inox, de 1 mm fixado na crista da barragem e mantido tensionado por um peso de 30 kg a 40 kg imerso em óleo, posicionado na interface fundação-estrutura. Desse modo, é realizada a monitoração do deslocamento desse fio de aço em bases de referência instaladas em distintas elevações da barragem (SILVEIRA, 2003; ITAIPIU BINACIONAL, 1992b). Na Figura 31(a) pode ser observada a representação esquemática dos

elementos constituintes de um pêndulo direto.

5.1.1 Procedimento de Medição

A medição de deslocamento por meio do uso de pêndulos diretos e invertidos baseia-se na criação de um sistema de coordenadas para cada base de referência. Nesse sistema, a posição do fio do pêndulo é dada pelo ponto $P = (X, Z)$, conforme representado na Figura 32. O eixo Y refere-se à origem do sistema de coordenadas e está alinhado com a orientação do fio do pêndulo. Os eixos X e Z estão relacionadas aos deslocamentos montante-jusante e lateral da estrutura.

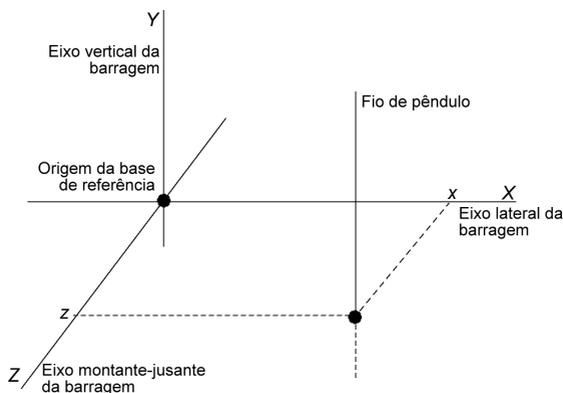


Figura 32: Sistemas de coordenadas para a medição de deslocamentos com pêndulos.

O deslocamento do fio do pêndulo é determinado por meio de uma medição $M_f = (X_f, Z_f)$ subtraída da medição inicial $M_i = (X_i, Z_i)$ realizada com o instrumento. O resultado dessa operação é denominado deslocamento relativo, definido por $M_r = (X_f - X_i, Z_f - Z_i)$.

A medição do deslocamento da crista da barragem em relação à fundação, denominado deslocamento absoluto, é realizado por meio da combinação da medição dos deslocamentos relativos do pêndulo direto e do pêndulo invertido, presentes em um mesmo bloco da barragem (ITAIPU BINACIONAL, 1992b; FEDERAL ENERGY REGULATORY COMMISSION - FERC, 2008). Essa medição é caracterizada pela Equação 5.1.

$$M_a = M_r(D) + M_r(I) \quad (5.1)$$

onde M_a representa a medição do deslocamento absoluto e $M_r(D)$ e $M_r(I)$, a medição do deslocamento relativo do pêndulo direto e do pêndulo invertido, respectivamente.

5.1.2 Sistemas de Medição

A medição dos deslocamentos do fio do pêndulo de acordo com as definições apresentadas na subseção anterior é realizada por meio de um instrumento denominado coordenômetro. Atualmente, na Usina de Itaipu são utilizados duas classes desse instrumento, coordenômetro óptico (Figura 33(a)) e coordenômetro digital (Figura 33(b)).

O coordenômetro óptico consiste em um sistema acoplável a uma base de fixação, o qual baseia-se no alinhamento de um par de escalas, dispostas de modo ortogonal, em relação à uma referência cônica por meio de uma lente objetiva (ITAIPU BINACIONAL, 1992a).

O coordenômetro digital baseia-se na utilização de sensores foto-sensíveis, posicionados ortogonalmente, para capturar imagens provenientes do reflexo de luz infra-vermelha sobre o fio do pêndulo. Nesse sistema, os sensores são alinhados em relação ao mensurando pela movimentação dos mesmos ao longo dos eixos de medição, por meio de motores de passo (SISGEO, 2007; ITAIPU BINACIONAL, 1992a).

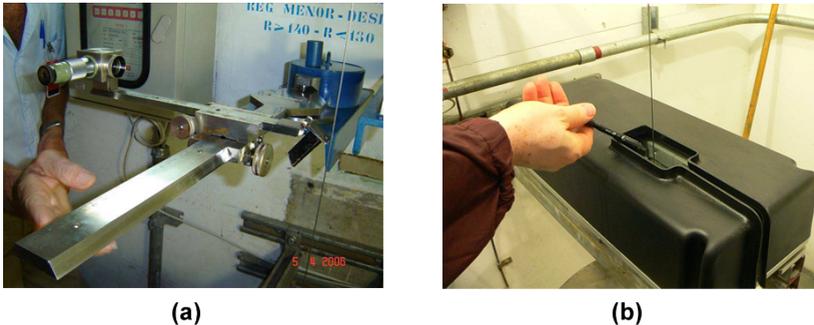


Figura 33: Coordenômetro Óptico (a) e Coordenômetro Digital (b).

Em relação às características metroológicas desses sistemas de medição, a incerteza de medição típica declarada pelos fabricantes do coordenômetro óptico é de $\pm 0,2 \text{ mm}$. No entanto, em estudo realizado por Silveira (2003) foi avaliado o procedimento de medição adotado pelos distin-

tos operadores, que fazem a medição dos coordenômetros, e constatou-se um nível de incerteza de medição de até $\pm 0,5 \text{ mm}$.

O coordenômetro digital é geralmente empregado por meio de um procedimento de medição automatizado. Nesses sistemas, o nível de incerteza típico declarado pelos fabricantes é de $\pm 0,05 \text{ mm}$.

5.1.3 *Série Histórica das Medições*

Como mencionado no capítulo 3, a frequência da realização de leituras da instrumentação de barragens deve ser adequada para permitir identificar o comportamento da grandeza monitorada. Desse modo, distintas frequências de medição são geralmente utilizadas para a monitoração de pêndulos diretos e invertidos dependendo da fase do ciclo de vida de uma barragem. Na Tabela 6 são apresentados os valores mínimos geralmente aplicados para a frequência de monitoração dos pêndulos (FEDERAL ENERGY REGULATORY COMMISSION - FERC, 2008).

Tabela 6: Frequências de medição tipicamente empregadas para a monitoração de pêndulos diretos e invertidos (FEDERAL ENERGY REGULATORY COMMISSION - FERC, 2008).

Período	Frequência
Enchimento	Semanal à mensal
Primeiro ano após o enchimento	Mensal à trimestral
Segundo ano após o enchimento	Mensal à semestral
Operação normal	Semestral à anual

Na barragem da Usina de Itaipu são adotadas duas frequências de leitura das medições dos pêndulos. No sistema de medição com coordenômetros ópticos, as leituras são realizadas mensalmente e no sistema com os coordenômetros digitais o registro das informações é realizado a cada 30 minutos.

Os dados das medições realizadas por meio da leitura manual dos coordenômetros ópticos são registrados no Sistema de Acompanhamento da Instrumentação (SAI) e das medições realizadas de modo automatizado, a partir dos coordenômetros digitais, são armazenadas em um Sistema de Aquisição Automática de Dados (ADAS).

Neste trabalho, os dados utilizados para análise correspondem às medições da instrumentação de pêndulos, registradas no SAI e no ADAS, para os deslocamentos absolutos no sentido montante-jusante.

Na Figura 34 são apresentados os deslocamentos relativos no sentido montante-jusante em diferentes níveis de instalação dos pêndulos diretos no bloco-chave F-19 da Usina de Itaipu. Os deslocamentos registrados nessa representação são provenientes do SAI e correspondem aos 16 primeiros anos de operação da usina.

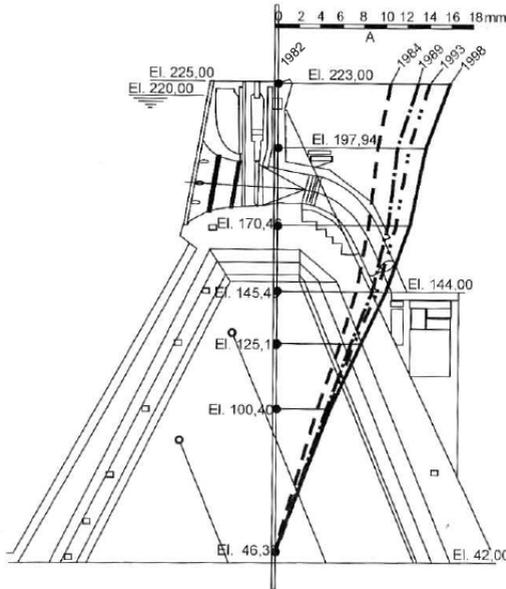


Figura 34: Perfil de deslocamentos relativos de um bloco-chave F-19 da Usina de Itaipu (RIBEIRO et al., 2008).

Como pode ser observada, a estrutura de concreto sofreu contínua deformação ao longo do tempo em relação ao seu eixo vertical. Esse comportamento deve-se à maior solicitação do reservatório nos níveis superiores.

Em relação às medições de deslocamento absoluto, na Figura 35 é apresentada a série histórica das medições registradas no SAI desde o início de operação da usina.

Como pode ser observado na Figura 35, o período de maior deslocamento compreende a fase de enchimento do reservatório. Características da série, como tendência e sazonalidade dos deslocamentos, também podem ser observados (ITAIPIU BINACIONAL, 2008). Nessas observações, a amplitude sazonal é estimada em aproximadamente 5 mm.

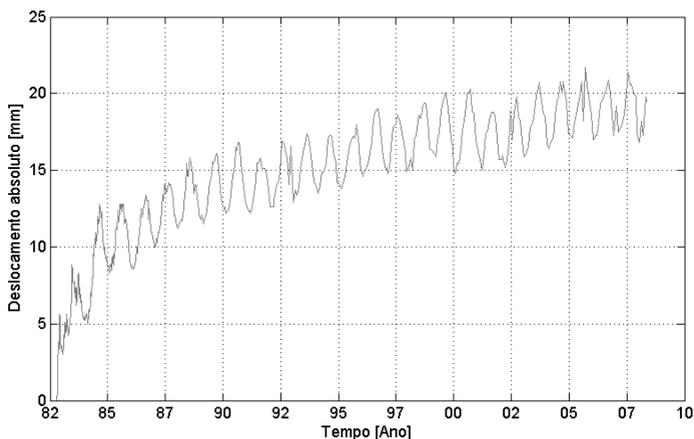


Figura 35: Série histórica de deslocamentos absolutos do bloco F-19.

5.2 Aplicação do Método de Avaliação

Para a aplicação do método de avaliação proposto neste trabalho, foram utilizados dados de deslocamento absoluto do bloco F-19, provenientes do sistema manual, registrados no SAI no período de 2000 até 2010. Na Figura 36 é apresentada a série histórica dos dados correspondente à esse período. Os pontos em cor azul correspondem ao período de medições realizadas entre o ano de 2000 e 2005, e os pontos em cor vermelha refere-se às observações feitas do ano de 2005 até o ano de 2010.

Neste trabalho, também é realizada a análise das medições efetuadas com o sistema automatizado e registradas no ADAS. Nesse caso, a série de dados utilizada compreende o período 2005 até 2010, a qual é apresentada na Figura 37.

Nesse contexto, como os dados utilizados para este estudo referem-se à medição de deslocamento absoluto, foi realizada a combinação das medições de deslocamento relativo dos pêndulos direto e invertido, conforme definido na Equação 5.1. Nesse caso, com base nas declarações de incerteza de $\pm 0,5 \text{ mm}$ e $\pm 0,05 \text{ mm}$ para os sistemas de medição manual e automatizado, respectivamente, foi realizada uma avaliação de incertezas considerando o Método Clássico.

Desse modo, a estimativa resultante foi de $\pm 0,82 \text{ mm}$ para o sistema

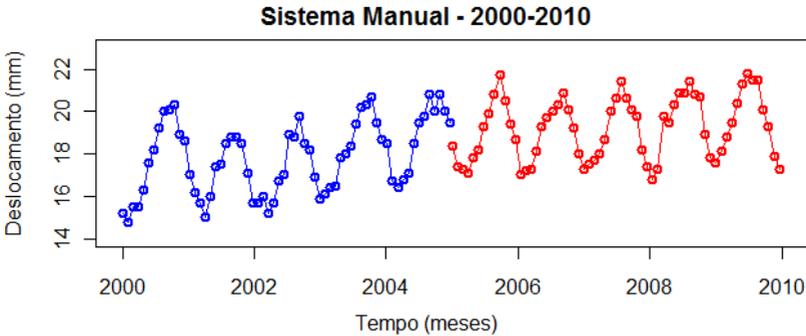


Figura 36: Série de dados temporais do SAI (2000-2010).

manual e de $\pm 0,082 \text{ mm}$ para o sistema automatizado, desconsiderando a existência de correlação.

Com base nessas estimativas de declaração de incerteza, na aplicação da Etapa 1 do método, foram calculadas as contribuições do erro aleatório e do erro sistemático para cada cenário proposto. Na Tabela 7 são apresentados esses valores para os sistemas manual e automatizado.

Tabela 7: Contribuições de erro sistemático e de erro aleatório.

Cenário	a (SAI / ADAS)	σ (SAI / ADAS)
Cenário 2	$\pm 0,71 / \pm 0,071 \text{ mm}$	$0,00 / 0,000 \text{ mm}$
Cenário 3	$0,00 / 0,000 \text{ mm}$	$\pm 0,41 / \pm 0,041 \text{ mm}$
Cenário 4	$\pm 0,29 / \pm 0,029 \text{ mm}$	$\pm 0,50 / \pm 0,050 \text{ mm}$

Os dados provenientes do SAI (Figura 36) foram divididos em duas séries temporais: uma compreendendo todas as observações no período de 2000 até 2010 (127 medições) e a outra correspondente aos registros de 2005 até 2010. Os dados provenientes do ADAS (71951 medições) foram organizados em três séries temporais com diferentes frequências de amostragem das observações: diária, semanal e mensal (Figura 37).

A partir da declaração de incerteza determinada para ambos os sistemas de medição e das séries temporais definidas extraídas foram geradas $M = 10000$ séries contaminadas com erros de medição.

Para a aplicação da Etapa 2, os parâmetros do algoritmo $k\text{NN}$ - TSP foram configurados conforme a avaliação experimental realizada no capítulo

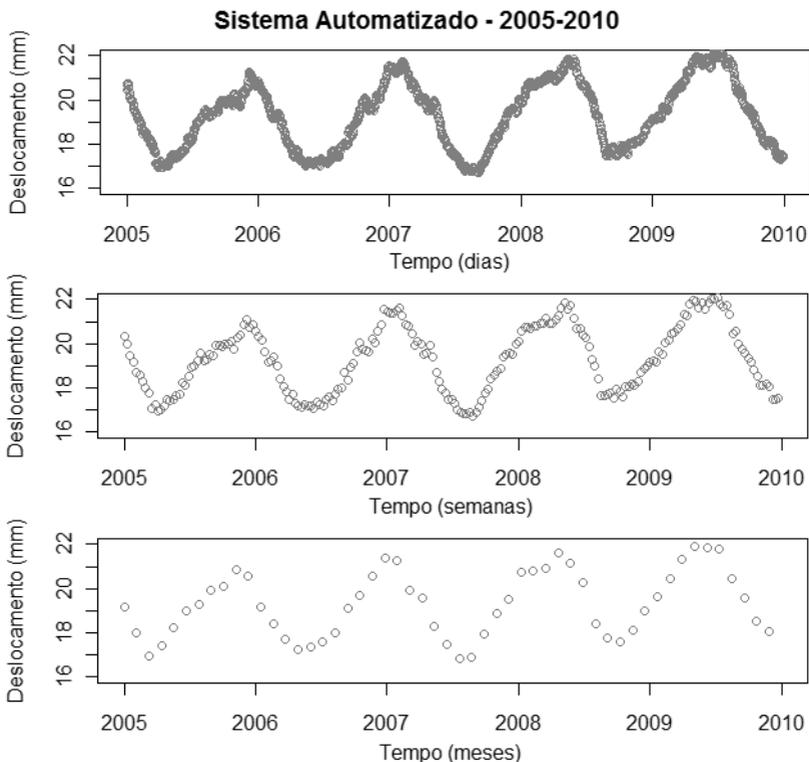


Figura 37: Frequências de amostragem diária, semanal e mensal dos dados do ADAS (2005-2010).

4. Com exceção dos atributos tamanho da janela de busca w e quantidade de valores a serem previstos N .

Para a previsão dos valores futuros sobre as séries de dados extraídas do SAI foi utilizada uma janela de busca de doze meses correspondente a um valor de $w = 12$ para realizar a previsão dados em seis meses, o que refere-se à uma quantidade $N = 6$ pontos.

Os valores de w e de N para as séries de dados extraídas do ADAS são descritos na Tabela 8 para cada frequência de amostragem dos dados adotada. Nesses casos, também foram utilizados 12 meses e 6 meses para tamanho da janela de busca e quantidade de valores a serem previstos, res-

pectivamente. No entanto, como a frequência de amostragem é diferente, o número de observações varia em cada situação.

Tabela 8: Quantidade de pontos (P) x Frequência de amostragem (F) para os dados provenientes do ADAS.

Pontos/Frequência	Diária	Semanal	Mensal
Janela de busca	336	48	12
Período de previsão	168	24	6
Total de pontos	1540	220	56

Posteriormente, na realização da Etapa 3, a partir das M previsões dos N valores para cada série foram calculados os parâmetros EMA e incerteza de previsão conforme o método proposto no capítulo anterior. Com base nesses valores foram realizadas as seguintes avaliações em relação à efetividade do algoritmo kNN - TSP:

- Sistema manual: conjunto de treinamento com dados dos últimos 10 anos (2000-2010) e dos últimos 5 anos (2005-2010);
- Sistema automatizado: frequência de amostragem dos dados (2005-2010) diária, semanal e mensal;
- Sistema manual e automatizado no período de 2005 até 2010.

5.3 Resultados e Discussão

Os dados utilizados nesse trabalho correspondem às medições de deslocamento realizadas no sentido montante-jusante em virtude de estas apresentarem maiores níveis de variação ao longo do tempo em relação ao deslocamento lateral. O bloco-chave F-19 foi adotado para estudo devido ao fato de que tem sido utilizado em outros estudos na área de segurança de barragens.

Para a combinação das incertezas dos pêndulos diretos e invertidos foi considerado um caso conservador no qual não se conhece a distribuição desses parâmetros. Desse modo, a distribuição da incerteza declarada pelo fabricante para os coordenômetros óptico e digital foi considerada como retangular.

As observações de deslocamentos extraídas do ADAS compreendem o período a partir do ano de 2005, para o qual o sistema com coordenômetro digital foi considerado estável pelos engenheiros de Itaipu. Os dados adquiridos do SAI correspondem aos registros das medições realizadas na mesma época dos registros do ADAS, para que fosse possível realizar uma comparação de ambos os sistemas de medição.

Outra característica avaliada em relação aos dados do sistema manual refere-se à utilização de uma série histórica maior. Essa avaliação foi realizada para verificar se existe uma melhora no desempenho do algoritmo k NN - TSP. Desse modo, foram extraídas do SAI medições entre o ano 2000 até 2010, e foram comparadas com as observações realizadas no período de 2005 até 2010. Essas duas séries temporais do SAI estão identificadas na Figura 36.

Na Figura 38 são apresentados os resultados da aplicação do método de avaliação para ambos os conjuntos de dados do SAI em cada um dos cenários de erro propostos. Os pontos em cor preta representam os valores reais da série para o período de 6 meses. Em cor amarela são representados os valores previstos pelo algoritmo k NN - TSP sobre a série de dados real. A linha representada em cor vermelha nos gráficos refere-se ao valor médio previsto a partir das 10000 simulações realizadas para cada instante da série. As linhas definidas em cor azul representam o intervalo de confiança para cada ponto previsto, conforme definido na Equação 4.10.

Os valores de incerteza de previsão para a série de dados 2000-2010 e 2005-2010, calculados segundo a Equação 4.12, para os cenários 2, 3 e 4 são apresentados no gráfico da Figura 39. Para as mesmas séries temporais são apresentados, no gráfico da Figura 40, os valores de EMA calculados segundo a Equação 4.13.

Por meio da análise dos resultados apresentados nesses gráficos, pode-se observar que a incerteza de previsão dos valores para os cenários 2, 3 e 4 permaneceu praticamente a mesma para ambas as séries. Observa-se também que a contribuição à incerteza de erro puramente sistemático não afetou significativamente a incerteza de previsão dos valores. Para os cenários 3 e 4, constata-se uma maior influência na incerteza dos valores previstos em virtude da presença da componente de erro aleatório, a qual é evidenciada principalmente no cenário 3.

Em relação à qualidade do ajuste do algoritmo k NN - TSP observa-se que a diferença entre os valores de EMA das duas séries foi baixa, em torno de 0,03 *mm*. No entanto, os valores de EMA foram significativos e próximos à incerteza declarada do instrumento.

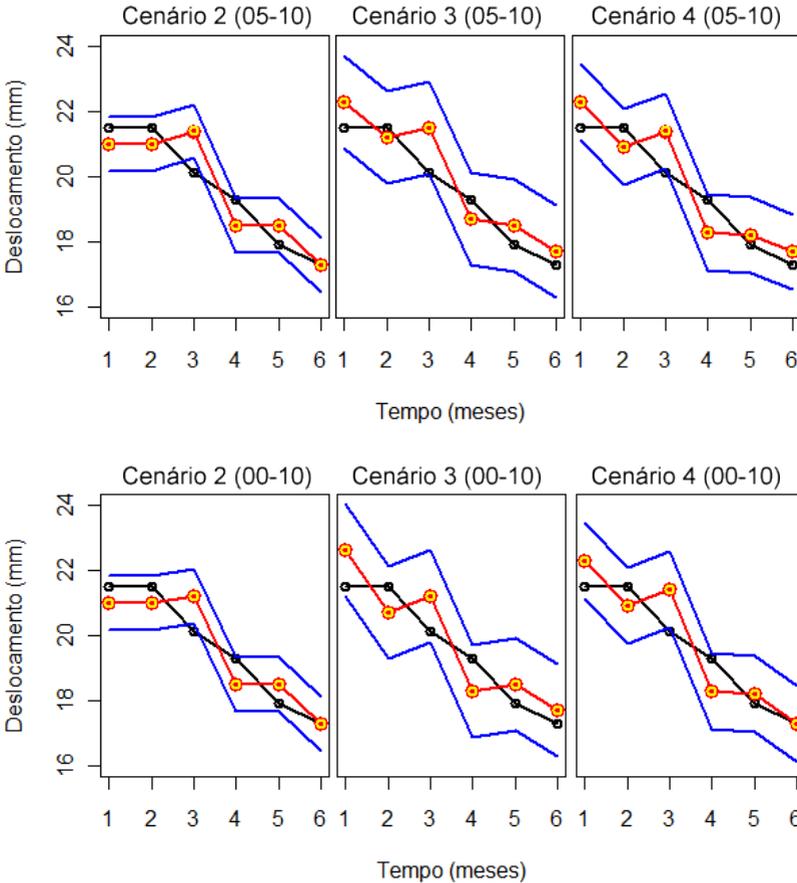


Figura 38: Valores previstos para os dados do SAI.

Nessa avaliação, pode-se verificar que um aumento no conjunto de treinamento não afeta a incerteza de previsão dos valores e que a qualidade das previsões do algoritmo kNN - TSP é melhorada.

Como mencionado, a frequência de amostragem dos dados da instrumentação de barragens, como o pêndulo, são variáveis ao longo do tempo. Conforme a estabilidade da grandeza monitorada, a frequência de amostragem dos dados é reduzida ou então elevada de acordo com a necessidade de monitoração. Desse modo, neste trabalho foi proposto o estudo

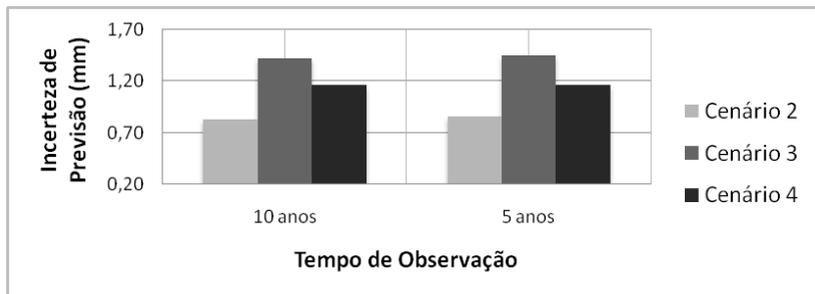


Figura 39: Valores de incerteza de previsão para o sistema manual.

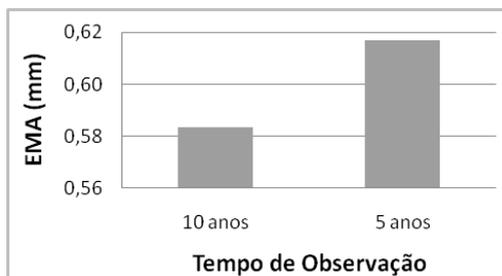


Figura 40: Valores de EMA para o sistema manual.

do comportamento do algoritmo k NN - TSP sob diferentes frequências de amostragem dos dados.

Nesse caso a série de dados extraída do ADAS foi organizada em séries com distintas frequências de amostragem dos dados, conforme a representado na Figura 37.

Os resultados da aplicação do método de avaliação para cada uma dessas séries, em cada cenário proposto, são apresentados nos gráficos contidos na Figura 41. O significado da coloração dos pontos e das linhas nesses gráficos baseia-se na mesma representação utilizada na Figura 38.

Os valores calculados para a incerteza de previsão e EMA em cada série com distinta frequência de amostragem, são apresentados nos gráficos das Figuras 42 e 43, respectivamente.

Por meio da análise dos resultados apresentados nesses gráficos, pode-se observar que o cenário de contribuição à incerteza composto somente de

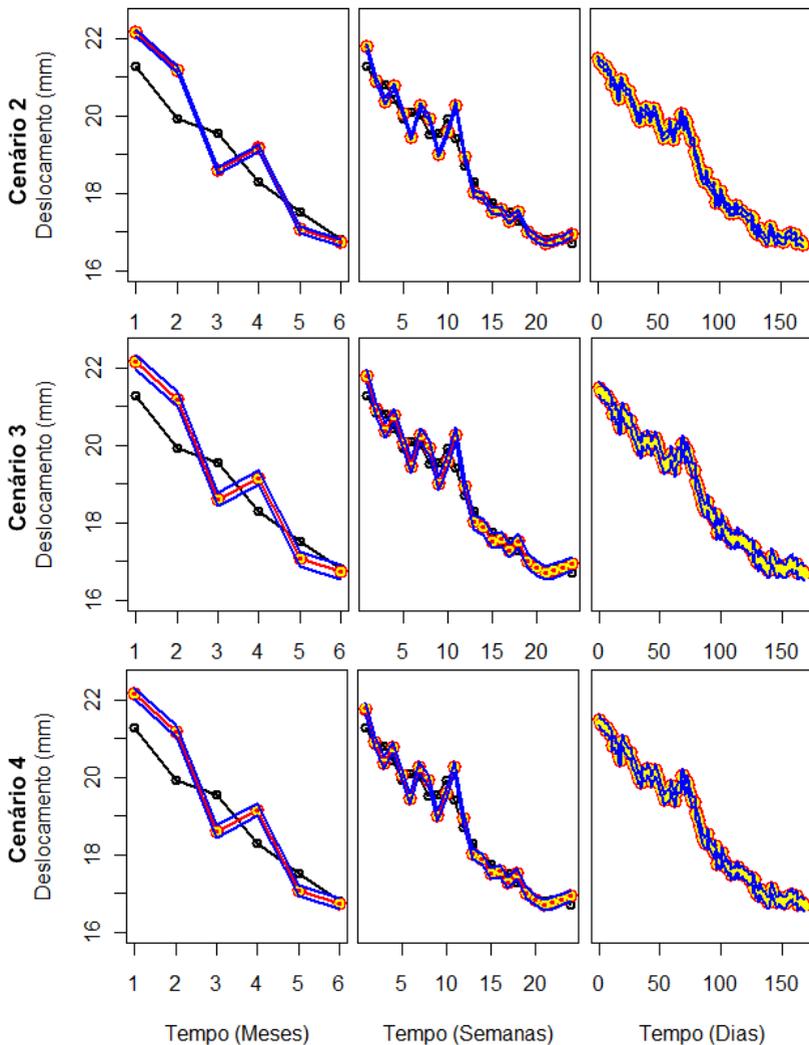


Figura 41: Valores previstos para os dados do ADAS.

erro sistemático não afetou a incerteza de previsão dos valores, mantendo-se constante nas distintas frequências de amostragem dos dados utilizadas. Além dessa característica, foi preservada a incerteza declarada do instrumento.

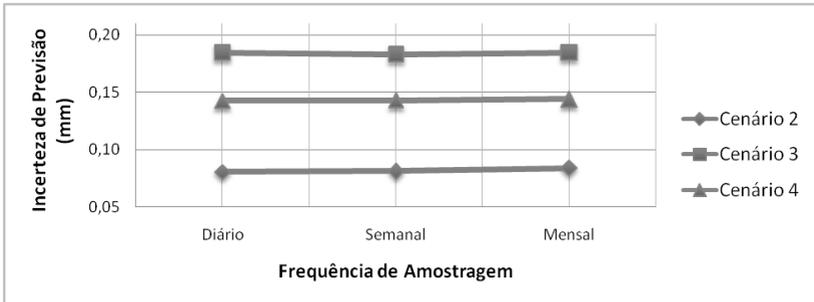


Figura 42: Valores de incerteza de previsão para o sistema automatizado.

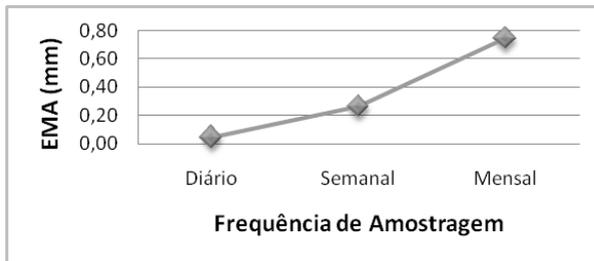


Figura 43: Valores de EMA para o sistema automatizado.

Em relação aos demais cenários, observa-se que também mantiveram a incerteza de previsão constante nas diferentes frequências de amostragem analisadas. No entanto, apresentaram um significativo aumento na incerteza de previsão, em especial para o cenário 3, o qual promoveu uma propagação de incertezas resultante de $\pm 0,18 \text{ mm}$, o que corresponde a mais da metade da incerteza declarada do instrumento.

Considerando os valores de EMA calculados para cada frequência de amostragem, verifica-se que o aumento na frequência de amostragem dos dados promove melhor qualidade de ajuste do algoritmo *k*NN - TSP na modelagem da série.

Como mencionado, os sistemas de medição manual e automatizado foram comparados no período do ano de 2005 até o ano de 2010. Os valores de incerteza de previsão para ambos os sistemas, em cada cenário de avaliação, são apresentados na Figura 44.

Pela análise das informações apresentadas nesse gráfico, pode-se ob-

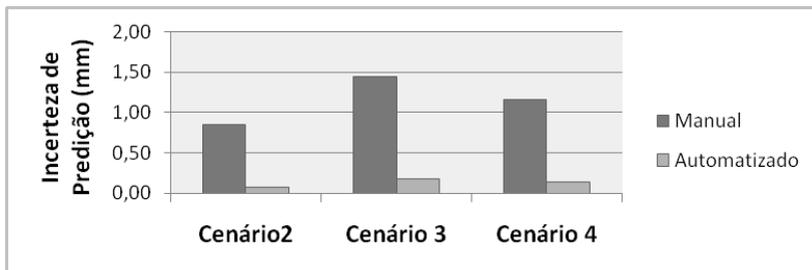


Figura 44: Valores de incerteza de previsão para os sistemas manual e automatizado.

servar que para ambos os sistemas de medição, o comportamento da incerteza de previsão propagada é mantido em todos os cenários. O cenário 2, apresenta o menor nível de incerteza, de modo que não revelou aumento em relação à incerteza declarada do instrumento. Os resultados do cenário 3 demonstram os maiores níveis de incerteza de previsão, com valores de mais que o dobro da incerteza declarada do instrumento.

Os níveis de incerteza de previsão do sistema de medição manual foram cerca de 10 vezes maiores em relação ao sistema de medição automatizado. Nesse caso, verifica-se que a relação de incerteza declarada dos instrumentos também foi mantida.

Na Figura 45 são apresentados os valores de EMA em um mesmo gráfico para ambos os sistemas. Considerando que a frequência de amostragem dos dados para as duas séries é mensal, observa-se que a qualidade de ajuste do algoritmo foi prejudicada em ambas as séries temporais.

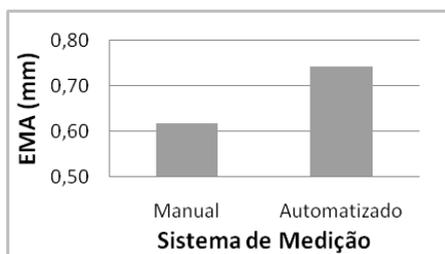


Figura 45: Valores de EMA para os sistemas manual e automatizado.

Nessa avaliação, verificou-se que mesmo com uma redução da incer-

teza da instrumentação, o fator frequência de amostragem afetou negativamente a capacidade de previsão do algoritmo k NN - TSP.

Em relação aos parâmetros do algoritmo k NN - TSP, estes foram ajustados segundo as características do comportamento apresentado pelas séries temporais das medições de deslocamentos e pelas informações fornecidas pelos especialistas da área.

Nesse contexto, sabe-se que a influência das variações ambientais de temperatura afetam diretamente as medições de deslocamentos. No inverno, a crista da barragem se desloca para jusante, como consequência da queda da temperatura ambiente, enquanto que a montante a água provoca certo isolamento térmico. (SILVEIRA, 2003).

Essas variações térmicas entre as estações de verão e inverno promovem um comportamento sazonal dos deslocamentos da barragem, conforme pode ser observado nas Figuras 35, 36 e 37. Desse modo, para representar esse comportamento sazonal, o parâmetro tamanho da janela de busca foi estabelecido para $w = 12$ meses.

Os pêndulos invertidos, por estarem instalados no interior do maciço rochoso de fundação, não são afetados pelas variações térmicas anuais.

Outro comportamento característico de séries temporais, presente nas medições de deslocamentos da Usina de Itaipu, refere-se à componente de tendência. Como pode ser observado na Figura 35, no período inicial de construção e enchimento, ocorreram as maiores variações na tendência da série dos dados de modo que nos anos posteriores a tendência tem apresentado um comportamento de estabilização, no entanto ainda em elevação¹.

Comportamentos de tendência e sazonalidade, como os presentes na série histórica de dados dos pêndulos, foram avaliados no capítulo anterior por meio de séries temporais artificiais, nas quais o algoritmo k NN - TSP foi capaz de representar esses comportamentos.

5.4 Considerações Finais

Deslocamentos são observados em uma barragem durante todas as fases de seu ciclo de vida, mais intensamente no período de enchimento. No entanto, são observados deslocamentos com comportamento do tipo sazonal durante todas as fases do ciclo de vida de uma barragem. Essa característica decorre da variação do nível do reservatório e da temperatura do ambiente que causa dilatação na estrutura.

¹Todos os valores encontram-se dentro da normalidade e de acordo com as condições de projeto (ITAIPU BINACIONAL, 2008)

A amplitude dos deslocamentos relativos e absolutos depende das características de cada barragem em particular. Desse modo, problemas com relação à integridade de uma estrutura podem ser observados por meio do surgimento de tendências anômalas nas curvas de controle (GOLZE, 1977; FEDERAL ENERGY REGULATORY COMMISSION - FERC, 2008).

Nesse contexto, a utilização de ferramentas de previsão de dados em segurança de barragens, tais como o algoritmo k NN - TSP aplicado à previsão de deslocamentos, torna-se uma importante ferramenta. Entretanto, observa-se a necessidade de se realizar uma avaliação da qualidade dos dados e da qualidade de ajuste do algoritmo aos dados, para que se tenha efetividade no uso dessas ferramentas.

Desse modo, no capítulo 6 são apresentadas as conclusões desse trabalho, bem como as limitações do método e as principais contribuições.

6 CONCLUSÕES E TRABALHOS FUTUROS

O avanço tecnológico ocorrido nos últimos anos tem possibilitado a aquisição de dados nas mais variadas áreas, dentre os quais, dados oriundos de medições realizadas continuamente ao longo do tempo. Por exemplo, dados são coletados por meio de processos de monitoração de sensores de temperatura e tráfego de rede, análises periódicas de materiais, flutuações do mercado financeiro e instrumentação de barragens. Dados coletados ao longo do tempo podem ser denominados de série temporal.

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, requerem que as séries temporais sejam compreendidas em termos das componentes que as constituem de modo a utilizar esse entendimento na realização dessas tarefas (MALETZKE, 2009).

Em sistemas de medição nos quais os dados podem ser entendidos como uma série temporal, as técnicas tradicionais podem apresentar algumas limitações. Desse modo, abordagens propostas na literatura baseiam-se na utilização de técnicas de aprendizado de máquina para auxiliar nessa tarefa (SOLOMATINE; MASKEY; SHRESTHA, 2006; FERRERO, 2009).

Nesse contexto, algoritmos de previsão de dados têm sido propostos na literatura para auxiliar na tomada de decisão em processos de medição. No trabalho de Ferrero (2009) foi proposto o algoritmo k NN-TSP, o qual foi aplicado na previsão de dados provenientes do monitoramento de grandezas ambientais na área de segurança de barragens.

Na aplicação desses algoritmos, deve-se considerar que a modelagem do processo com dados afetados por erros de medição pode produzir resultados que não caracterizam a realidade. Nesse contexto, torna-se importante avaliar a adequabilidade dessas ferramentas para a monitoração de processos de medição.

Um processo de medição, em condição de operação normal, apresenta um comportamento estocástico consistente. Nesse caso os erros de medição podem ser classificados como erros sistemáticos e aleatórios. Os erros sistemáticos apresentam valores aproximadamente constantes para repetidas medições e os erros aleatórios seguem uma determinada distribuição com média igual a zero (KONRATH, 2008).

Desse modo, neste trabalho foi proposto um modelo de simulação para

avaliar o efeito da incerteza de medição, em função de distintas composições de erros sistemáticos e aleatórios, e da frequência de amostragem dos dados, sobre a efetividade das previsões do algoritmo k NN-TSP. O modelo foi aplicado em séries temporais artificiais e em séries provenientes de um processo de monitoração real, nesse caso, dados da medição de deslocamentos da Usina Hidrelétrica de Itaipu.

A Simulação de Monte Carlo constitui uma alternativa ao método clássico para a avaliação de incertezas, aplicável em qualquer situação prática. No entanto, especificamente indicada quando modelos matemáticos complexos estão envolvidos, nos quais a grandeza medida não pode ser explicitamente expressa em razão das grandezas de influência e em situações nas quais o modelo matemático de medição apresenta acentuada não linearidade (DONATELLI; KONRATH, 2005).

Como mencionado, um dos métodos de aprendizado de máquina mais utilizados é o *k-Nearest Neighbor*. A adaptação desse método para a previsão de dados temporais como o k NN - TSP tem como vantagens, a simplicidade e aplicabilidade na análise de séries temporais não lineares e na previsão de comportamentos sazonais. Uma das principais características desse método em relação aos demais não lineares consiste no fato de que os modelos construídos com o algoritmo k NN são aproximações locais, enquanto outras técnicas tal como redes neuronais artificiais (FERRERO, 2009; SOLOMATINE; MASKEY; SHRESTHA, 2006) realizam aproximações globais dos modelos.

Os parâmetros utilizados nesse trabalho para a aplicação do algoritmo k NN - TSP basearam-se nos resultados apresentados no trabalho de Ferrero (2009), no qual foi realizado uma avaliação dos parâmetros que influenciam o desempenho do k NN - TSP.

A Linguagem Computacional R foi utilizada para o desenvolvimento dos algoritmos do ambiente de simulação utilizados nesse trabalho e também para o desenvolvimento do algoritmo k NN - TSP (FERRERO, 2009). Essa ferramenta possui um grande conjunto de bibliotecas para operações de cálculo numérico, de análise estatística e de desenvolvimento de algoritmos. Em especial para esse trabalho, a utilização de rotinas para a geração de valores aleatórios considerando distintas distribuições, para a construção de gráficos e para a manipulação de séries temporais.

Por meio da análise dos resultados experimentais apresentados neste trabalho, foi possível constatar alguns aspectos relacionados à aplicação do algoritmo k NN-TSP em função da incerteza de medição e da frequência de amostragem dos dados, bem como permitiu definir algumas recomendações em relação à previsão de dados no domínio de segurança de barragens.

No contexto deste trabalho, verificou-se que foi possível propagar a incerteza de medição através do modelo de simulação proposto, de modo que a incerteza resultante dos valores previstos pode ser afetada significativamente.

Os resultados das avaliações demonstraram que as contribuições à incerteza de medição estudadas no modelo de simulação proposto neste trabalho, interferem de modo diferenciado na incerteza de previsão resultante. Desse modo, observou-se que as contribuições devido aos erros sistemáticos não apresentam significativa influência na incerteza resultante da previsão dos valores.

Em relação à componente de erro aleatório, foi constatado que esta consiste na contribuição à incerteza que apresenta a maior influência na efetividade da aplicação do algoritmo k NN-TSP para a previsão de valores.

Desse modo, constatou-se que processos de medição que apresentam a mesma declaração de incerteza, podem representar diferentes composições das contribuições à incerteza de medição. Nesse caso, as previsões de valores de deslocamentos, da instrumentação de pêndulos, apresentarão diferentes desempenhos.

O aumento do conjunto de dados não implica em uma significativa melhoria na qualidade das previsões realizadas e não afeta a incerteza de previsão resultante. Desse modo, a aplicação do algoritmo k NN-TSP não apresentará melhor desempenho significativo com o aumento do conjunto de dados da série. Também, deve-se observar que um conjunto de treinamento reduzido resulta em um melhor desempenho do algoritmo em termos de custos computacionais. No entanto, é importante ressaltar que o conjunto de dados da série, mesmo que reduzido, deve representar os comportamentos atuantes da grandeza estudada.

Constatou-se que a frequência de amostragem dos dados afeta significativamente a qualidade das previsões realizadas pelo algoritmo k NN-TSP. Nesse caso, foi observado que o aumento da frequência de amostragem dos dados da série implica em um melhor desempenho do algoritmo de previsão. Na aplicação sobre os dados de deslocamentos dos pêndulos verificou-se que a utilização de uma amostragem mais densa das leituras refletiu em uma boa qualidade do ajuste do algoritmo aos dados, de modo que a qualidade das previsões foi afetada somente pela incerteza de medição. Outra característica observada refere-se ao fato de que uma redução da frequência de amostragem dos dados não tem efeito sobre a incerteza de previsão resultante, desde que o comportamento da grandeza monitorada não seja descaracterizado.

Na avaliação realizada por meio da análise de séries temporais artificiais verificou-se que a incerteza dos valores previstos, devido às contribuições

aleatórias, atingiu níveis de até 35% da amplitude da série. Em relação a esse comportamento pode-se induzir que em sistemas de medição com maior atuação de erros aleatórios, as previsões realizadas devem comprometer significativamente a identificação do comportamento sazonal da grandeza monitorada. Como mencionado, constatou-se que a incerteza devido às contribuições sistemáticas não representam significativa influência na incerteza resultante dos valores previstos, no entanto, é importante ressaltar que em situações de forte atuação dos erros sistemáticos as previsões realizadas irão indicar um posicionamento da grandeza afastado da realidade.

Por meio da comparação entre os sistemas de medição manual e automatizado, constatou-se que em níveis de incerteza maiores, como de 16% da amplitude da série para o sistema manual, a qualidade das previsões é principalmente afetada pela incerteza de medição. Em níveis de incertezas mais baixos, como 1,6% da amplitude da série para o sistema automatizado, a qualidade das previsões foi afetada, em quase sua totalidade, devido ao ajuste do algoritmo aos dados da série. No entanto, deve-se considerar que, em ambos os casos, a frequência de amostragem mensal prejudicou a qualidade de ajuste do algoritmo aos dados da série.

Com base nas observações mencionadas, algumas orientações com foco em uma relação custo-benefício entre a qualidade de ajuste do algoritmo aos dados da série e a incerteza de medição podem ser posicionadas com o intuito de auxiliar na escolha de ferramentas de análise de riscos em segurança de barragens.

Nesse contexto, conclui-se que a qualidade de ajuste do algoritmo k NN-TSP aos dados da série é afetada diretamente pela frequência de amostragem dos dados e que a incerteza dos valores previstos é influenciada pela incerteza do sistema de medição. Desse modo, em processos de medição com inadequada frequência de amostragem dos dados, deve-se buscar um melhor ajuste do algoritmo aos dados ou buscar outras alternativas ao invés de investir na melhoria da qualidade dos dados por meio de sistemas de medição com menor incerteza. Essa recomendação está relacionada ao fato de que os custos despendidos com uma instrumentação de maior qualidade não irão refletir em um melhor desempenho do algoritmo k NN-TSP.

Como mencionado, sabe-se que o aumento da frequência de amostragem pode melhorar o ajuste do algoritmo, no entanto devem-se considerar os custos associados à implantação de uma amostragem mais densa dos dados. Por outro lado, em situações nas quais é possível obter um adequado ajuste do algoritmo, a efetividade do controle de riscos por meio de previsões de valores futuros irá depender da qualidade das medições realizadas. Nesse caso, a

aplicação do método de avaliação proposto neste trabalho poderá ser utilizada como um guia na determinação do grau de influência da incerteza de medição na previsão dos dados.

É importante ressaltar que as orientações apresentadas a partir das conclusões deste trabalho referem-se ao caso específico de séries temporais com comportamentos característicos da área de segurança de barragens, tais como tendência e sazonalidade dos deslocamentos de pêndulos. Desse modo, a aplicação em outros domínios deve ser precedida de uma análise dos comportamentos da grandeza a ser monitorar.

Durante o desenvolvimento deste trabalho foram identificados alguns aspectos importantes que podem ser considerados em trabalhos futuros. Tais estudos podem contribuir para a solução das limitações deste trabalho, bem como permitir outras análises no contexto de previsão de dados e incerteza de medição. Esses trabalhos futuros incluem:

- Aplicar o método de avaliação em outras técnicas de aprendizado de máquina para a previsão de valores;
- Elaborar um modelo matemático de medição que possa compreender características de deriva temporal e resolução do instrumento de medição;
- Atualmente o algoritmo k NN-TSP utiliza somente o conjunto de treinamento para realizar a previsão de valores. Nesse caso, uma abordagem a ser avaliada nesses termos, refere-se à inclusão dos valores previstos no conjunto de treinamento. Desse modo, os valores previstos inicialmente terão impacto sobre a previsão dos dados posteriores;
- Aplicar o método de avaliação sobre dados provenientes de outras áreas que utilizam monitoração de grandezas ao longo do tempo;
- Avaliar o algoritmo k NN-TSP, por meio do método proposto, em séries temporais que apresentam comportamentos, além de tendência e sazonalidade, tais como variação da amplitude ao longo do tempo e variação da frequência ao longo do tempo;
- Analisar o efeito do ajuste do algoritmo k NN-TSP em relação aos demais parâmetros do mesmo.

REFERÊNCIAS

- ABNT; INMETRO. *Guia para a Expressão da Incerteza de Medição (GUM 2003)*. Terceira edição brasileira. Rio de Janeiro, 2003.
- AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. *Machine Learning* 6, p. 37–66, 1991.
- AIBE, V. Efeito da correlação no cálculo do número de graus de liberdade efetivo. In: *IMEKO World Congress*. [S.l.: s.n.], 2000.
- ALENCAR, J. R. B. et al. Uso de controle estatístico de processo para avaliação da estabilidade e validação da fase de compressão e formas farmacêuticas sólidas. *Acta Farm. Bonaerense*, v. 24, n. 3, p. 426–435, 2005.
- ALPAYDIN, E. *Introduction to Machine Learning*. Cambridge — MA, England: MIT Press, 2004.
- BAZILIO, F. S. et al. Evaluation of measurement uncertainty in analytical inorganic assays: A study of case. In: *XVIII IMEKO WORLD CONGRESS*. [S.l.: s.n.], 2006.
- BELANGER, B. C. et al. Traceability: an evolving concept - a chronicle on selected nbs/nist publications 1901-2000. In: *A Century of Excellence in Measurements, Standards and Technology*. Gaithersburg, NIST SP 958: D.R. Lide, 2000. p. 167–171.
- BOX, G. E. P.; JENKINS, C. M.; REINSEL, G. C. *Time Series Analysis: Forecasting and Control*. 3. ed. Englewood — NJ —, USA: Prentice hall, 1994.
- BROCKWELL, P. J.; DAVIS, R. A. *Introducing to Time Series and Forecasting*. Second. New York, NY-USA: Springer, 1996.
- BROWNJOHN, J. M. W. Structural health monitoring of civil infrastructure. *Philosophical Transactions of the Royal Society*, v. 365, p. 589–622, 2007.
- CAMILLERI, M. *Forecasting Using Non-Linear Techniques In Time Series Analysis: An Overview of Techniques and Main Issues*. Malta, 2004. Disponível em: <http://www.cs.um.edu.mt/~csaw/CSAW04/Proceedings/02.pdf>.

- CASTRO, M. C. F. de. *Predição Não-Linear de Séries Temporais Usando Redes Neurais RBF por Decomposição em Componentes Principais*. Tese (Doutorado) — Universidade Estadual de Campinas - UNICAMP, Campinas, SP, Março 2001.
- CASTRUP, H. Distributions for uncertainty analysis. In: *IDW Conference*. Knoxville, TN: [s.n.], 2001.
- CBGB. *Auscultação e Instrumentação de Barragens no Brasil*. Belo Horizonte, MG, Agosto 1996.
- COUTO, P. R. G.; DAMASCENO, J. C. Uncertainty estimation of mechanical assays by iso-gum 95 and monte carlo simulation - case study: Tensile strength, torque and brinell hardness measurements. In: *XVIII IMEKO World Congress*. Rio de Janeiro, Brasil: [s.n.], 2006.
- COX, M. G.; HARRIS, P. M. Method for analysis of uncertainties in optical radiation measurement. In: *GUM Supplements. CIE Expert Symposium on Uncertainty Evaluation*. [S.l.: s.n.], 2001.
- DONATELLI, G. D.; KONRATH, A. C. Simulação de monte carlo na avaliação de incertezas de medição. *Revista de Ciência de Tecnologia*, v. 13, n. 25/26, p. pp. 5–15, jan./dez. 2005.
- EHLERS, R. S. *Análise de Séries Temporais*. Curitiba — PR, Brasil, 2005. Disponível em: <http://leg.est.ufpr.br/~ehlers/notas/stemp.pdf>.
- ELETOBRÁS. *Critérios de Projeto Civil de Usinas Hidrelétricas*. [S.l.: s.n.], 2003.
- FABRIS, F.; DRAGO, I.; VAREJÃO, F. M. A multi-measure nearest neighbor algorithm for time series classification. In: *IBERAMIA '08: Proceedings of the 11th Ibero-American conference on AI*. Berlin, Heidelberg: Springer-Verlag, 2008. p. 153–162. ISBN 978-3-540-88308-1.
- FEDERAL ENERGY REGULATORY COMMISSION - FERC. *Engineering Guidelines for the Evaluation of Hydropower Projects*. [S.l.], 2008.
- FERRERO, C. A. *Algoritmo kNN para previsão de dados: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia*. Dissertação (Mestrado) — Universidade de São Paulo - USP, 2009.

- GEUS, D.; KILLMAIER, A.; WECKENMANN, A. *Measurement Uncertainty and Statistical Process Control in Manufacturing Processes*. [S.l.], 2001.
- GOLZE, A. R. *Handbook of Dam Engineering*. [S.l.: s.n.], 1977.
- GRASSBERGER, P.; PROCACCIA, I. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, v. 9, n. 1–2, p. 189–208, 1983. ISSN 0167-2789. Disponível em: <<http://www.sciencedirect.com/science/article/B6TVK-46JYPW4-50/2-1621ece0782ffb2c77e4bf4180d76a459>>.
- GUM. *Guide to the Expression of Uncertainty in Measurement GUM Supplement 1. Numerical Methods for the Propagation of Distributions*. [S.l.], 2005.
- HAMBURG-PIEKAR, D.; DONATELLI, G. D.; SCHNEIDER, C. A. Setting the process aim: The effect of measurement uncertainty. In: *XVII IMEKO World Congress*. Croácia: [s.n.], 2003.
- HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting*, v. 22, n. 4, p. 679–688, 00 2006. Disponível em: <<http://www.sciencedirect.com/science/article/B6V92-4K18VK5-1/1/227cef013b86100791c53712fae7677c>>.
- ILLA, J. M. G.; ALONSO, J. B.; MARRÉ, M. S. Nearest-neighbours for time series. *Applied Intelligence*, Kluwer Academic Publishers, Hingham, MA, USA, v. 20, n. 1, p. 21–35, 2004. ISSN 0924-669X.
- INMETRO. *Vocabulário Internacional de Metrologia: Conceitos Fundamentais e Gerais e Termos Associados (VIM 2008)*. Primeira edição brasileira do vim 2008. Rio de Janeiro, 2009.
- ITAIPU BINACIONAL. *Descrição de funcionamento - Coordenômetro portátil*. [S.l.], 1992.
- ITAIPU BINACIONAL. *Descrição de Funcionamento - Pêndulo direto e invertido*. [S.l.], 1992.
- ITAIPU BINACIONAL. *Usina Hidrelétrica de Itaipu: aspectos técnicos das estruturas civis*. [S.l.], 2008.
- JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988. ISBN 0-13-022278-X.

- JORNADA, D.; PIZZOLATTO, M. Uso de planilhas eletrônicas para implementação do método de monte carlo para a estimativa de incerteza de medição. In: *ENQUALAB-2005 - Encontro para a Qualidade de Laboratórios*. São Paulo, Brasil: [s.n.], 2005.
- JORNADA, D. H. da; JORNADA, F. H. da. Cálculo de incerteza da medição: Estudo comparativo entre métodos do gum, de monte carlo e de integração numérica. In: REDE METROLÓGICA DO ESTADO DE SÃO PAULO - REMESP. *ENQUALAB-2007 - Congresso da Qualidade em Metrologia*. São Paulo, Brasil, 2007.
- KARUNASINGHE, D. S. K.; LIONG, S.-Y. Chaotic time series prediction with a global model: Artificial neural network. *Journal of Hydrology*, v. 323, n. 1-4, p. 92–105, 2006.
- KENNEL, M. B.; BROWN, R.; ABARBANEL, H. D. I. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A*, v. 45, p. 3403–3411, mar. 1992.
- KEOGH, E.; KASETTY, S. 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*. New York, USA: [s.n.], 2002. p. 102–110. Disponível em: <citeseer.ist.psu.edu/article/keogh02need.html>.
- KONRATH, A. C. *Influência do Processo de Medição no Controle Estatístico de Processos*. Tese (Doutorado) — Universidade Federal de Santa Catarina - UFSC, 2008.
- KULESH, M.; HOLSCHNEIDER, M.; KURENNAYA, K. Adaptive metrics in the nearest neighbours method. *Physica D: Nonlinear Phenomena*, v. 237, n. 3, p. 283–291, 2008. ISSN 0167-2789. Disponível em: <<http://www.sciencedirect.com/science/article/B6TVK-4PN05G2-2/2/2c7e5eeea9f277bb7fc80af5b0911ddd>>.
- MALETZKE, A. G. *Uma metodologia para a extração de conhecimento em séries temporais por meio da identificação de motivos e da extração de características*. Dissertação (Mestrado) — Universidade de São Paulo - USP, 2009.
- MATTOS, P. S. G. de et al. Combinação de redes neurais artificiais com algoritmo genético modificado para a previsão de séries temporais. In:

- UNISINOS. *XXV Congresso da Sociedade Brasileira de Computação*. São Leopoldo, RS, 2005. p. 952–961.
- MEDEIROS, C. H. Segurança e auscultação de barragens. In: *XXV Seminário Nacional de Grandes Barragens*. [S.l.: s.n.], 2003. p. 13–50.
- MITCHELL, T. M. *Machine learning*. Boston, USA: McGraw-Hill, 1997.
- MIZUHARA, Y.; HAYASHI, A.; SUEMATSU, N. Embedding of time series data by using dynamic time warping distances. *Syst. Comput. Japan*, Wiley-Interscience, New York, NY, USA, v. 37, n. 3, p. 1–9, 2006.
- MORETTIN, P. A.; TOLOI, C. M. C. *Análise de Séries Temporais*. [S.l.]: Edgard Blucher, 2006.
- MOSCATTI, G.; MEZZALIRA, L. G.; SANTOS, F. D. dos. Incerteza de medição pelo método de monte carlo no contexto do "suplemento 1" do gum. In: *ENQUALAB-2004 - Encontro para a Qualidade de Laboratórios*. São Paulo, Brasil: [s.n.], 2004.
- OSAKO, C. I. Segurança de barragens na itaipu binacional. 2008.
- PAN, J.; HE, J. *Large Dams in China: a fifty-year review*. Beijing, China: China WaterPower Press, 2000.
- PORTELA, E. A. Sistemas de apoio à decisão: aplicação ao controle de segurança de grandes barragens. In: *XXV Seminário Nacional de Grandes Barragens*. [S.l.: s.n.], 2003. p. 287–302.
- QS-9000. *Fundamentos de Controle Estatístico de Processo - CEP*. Tradução da terceira edição americana. São Paulo: IQA - Instituto de Qualidade Automotiva, 1997.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2009. ISBN 3-900051-07-0. Disponível em: <<http://www.R-project.org>>.
- REZENDE, S. O. *Sistemas Inteligentes: Fundamentos e Aplicações*. Barueri — SP, Brasil: Editora Manole, 2003.
- RIBEIRO, F. C. D. et al. Comparison between geodetic technology and plum lines in monitoring of displacements on itaipu dam. In: *13th Symposium on Deformation Measurement and Analysis*. [S.l.: s.n.], 2008.

- ROCHA, R. et al. Aplicação de um sistema computacional para tratamento e análise quantitativa de resultados de instrumentação de barragens. In: *XX Seminário Nacional de Grandes Barragens*. [S.l.: s.n.], 1992. p. 71–78.
- SHRESTHA, D. L.; KAYASTHA, N.; SOLOMATINE, D. Encapsulation of monte-carlo uncertainty analysis results in a predictive machine learning model. In: *7 ISE & 8 HIG*. Chile: [s.n.], 2009.
- SHRESTHA, D. L.; SOLOMATINE, D. Comparing machine learning methods in estimation of model uncertainty. In: *International Joint Conference on Neural Networks*. [S.l.: s.n.], 2008. p. 1411–1417.
- SHRESTHA, D. L.; SOLOMATINE, D. P. Machine learning approaches for estimation of prediction interval for the model output. *Elsevier, Neural Networks*, n. 19, p. 225–235, 2006.
- SILVA, J. H. C. da. *A Influência da Incerteza de Medição na Carta de Controle de Valores Individuais*. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, 2003.
- SILVA, L. M. O. da. *Uma Aplicação de Árvores de Decisão, Redes Neurais e KNN para a Identificação de Modelos ARMA Não-Sazonais e Sazonais*. Tese (Doutorado) — Pontificadora Universidade Católica do Rio de Janeiro - PUC-RJ, Rio de Janeiro, RJ, Setembro 2005.
- SILVEIRA, J. F. A. *Instrumentação e Comportamento de Fundações de Barragens de Concreto*. [S.l.: s.n.], 2003.
- SILVEIRA, J. F. A. *Instrumentação e Segurança de Barragens de Terra e Enrocamento*. 1. ed. [S.l.]: Oficina de Textos, 2006.
- SILVEIRA, J. F. A. et al. A automação da instrumentação da uhe de xingó e os custos envolvidos. In: *XXI Seminário Nacional de Grandes Barragens*. [S.l.: s.n.], 1994. p. 111–119.
- SISGEO. *Pendulums system*. [S.l.], 2007.
- SOBOL, I. M. *A Primer for the Monte Carlo Method*. Flórida: CRC, 1994.
- SOLOMATINE, D.; MASKEY, M.; SHRESTHA, D. Eager and lazy learning methods in the context of hydrologic forecasting. *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, v. 1, n. 1, p. 4847–4853, 2006.

- VUOLO, J. H. Avaliação e expressão de incerteza em medição. *Revista Brasileira de Ensino de Física*, v. 21, n. 3, p. 350–358, 1999.
- WESTERN-ELECTRIC-COMPANY. *Statistical Quality Control Handbook*. New York, 1958.
- WHEELER, D. *Short Run SPC*. Knoxville: SPC Press, 1991.
- WOODAL, W. H. Controversies and contradictions in statistical process control. *Journal of Quality Technology*, v. 32, n. 4, 2000.