

UNIVERSIDADE FEDERAL DE PERNAMBUCO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

UMA METODOLOGIA PARA ANÁLISE DE
DISPONIBILIDADE DE SISTEMAS COMPLEXOS
VIA HIBRIDISMO DE REDES BAYESIANAS E
PROCESSOS MARKOVIANOS

DISSERTAÇÃO SUBMETIDA À UFPE
PARA OBTENÇÃO DE GRAU DE MESTRE
POR

PAULO FERNANDO DO RÊGO BARROS JÚNIOR

Orientador: Prof. Enrique López Droguett, Ph.D.

RECIFE, AGOSTO/2006

Livros Grátis

<http://www.livrosgratis.com.br>

Milhares de livros grátis para download.

**UNIVERSIDADE FEDERAL DE PERNAMBUCO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO**

**PARECER DA COMISSÃO EXAMINADORA
DE DEFESA DE DISSERTAÇÃO DE
MESTRADO ACADÊMICO DE**

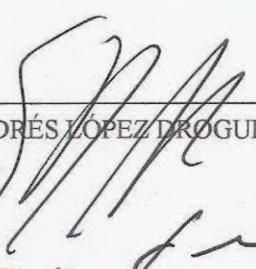
PAULO FERNANDO DO RÊGO BARROS JÚNIOR

***“UMA METODOLOGIA PARA ANÁLISE DE DISPONIBILIDADE DE SISTEMAS
COMPLEXOS VIA HIBRIDISMO DE REDES BAYESIANAS
E PROCESSOS MARKOVIANOS”.***

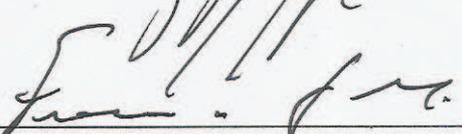
ÁREA DE CONCENTRAÇÃO: PESQUISA OPERACIONAL

A * comissão examinadora, composta pelos professores abaixo, sob a presidência do(a) primeiro(a), considera o candidato PAULO FERNANDO DO RÊGO BARROS JÚNIOR **APROVADO**.

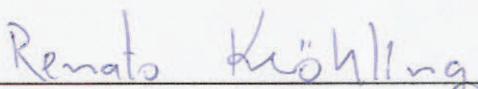
Recife, 24 de agosto de 2006.



Prof. ENRIQUE ANDRÉS LOPEZ DROGUETT, PhD (UFPE)



Prof. FRANCISCO DE SOUSA RAMOS, Docteur (UFPE)



Prof. RENATO ANTONIO KROHLING, PhD (UFES)

B277m **Barros Júnior, Paulo Fernando do Rêgo**

Uma metodologia para análise de disponibilidade de sistemas complexos via hibridismo de redes bayesianas e processos markovianos / Paulo Fernando do Rêgo Barros Júnior. – Recife: O Autor, 2006.

x, 92 f., il., gráfs., tabs

Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia de Produção, 2006.

Inclui referências bibliográficas.

1. Engenharia de Produção. 2. Redes bayesianas. 3. Processos markovianos. 4. Confiabilidade de sistemas. I. Título.

658.5 CDD (22 ed.)

UFPE/BCTG/2006-115

Esta dissertação é dedicada aos meus pais.

AGRADECIMENTOS

Como reconhecimento ao apoio nesse momento de muita reflexão, agradeço:

- Ao Meu Bom Deus que sempre Tem me mostrado os caminhos corretos a trilhar, muito obrigado;
- Aos meus pais, pelo o apoio irrestrito e amor transmitido que me possibilitaram a continuar na caminhada árdua do aprendizado, muito obrigado;
- Ao meu irmão, pelo companheirismo e apoio ao meu desenvolvimento profissional, muito obrigado;
- A minha noiva que tem estado comigo desde a época de adolescência e que mais uma vez se fez presente com muito amor nos momentos de maiores necessidades e isolamento, muito obrigado;
- Ao meu orientador, pela oportunidade oferecida e pelos imprescindíveis ensinamentos transmitidos na condição de mestre, muito obrigado;
- Aos membros da banca de avaliação, pelas sugestões gratuitamente oferecidas na melhoria deste trabalho, muito obrigado;
- Aos meus amigos do grupo de pesquisa RISCTEC, pela atenção e força despendida, especialmente a Heitor Duarte e a Márcio Moura pelos auxílios nos momentos de definição do trabalho, muito obrigado;
- Aos meus amigos de graduação Sérgio Parente, Rafael Assunção, Vândson Monteiro e Edgar Braga que continuaram comigo nessa jornada de aprofundamento do conhecimento, especialmente a Vândson pelo suporte no entendimento da língua francesa, muito obrigado;

Agradeço a todos que de alguma forma estiveram comigo ao longo desses quase dois anos de luta. Permanece em meu coração a eterna gratidão e o desejo que a Graça do Bom Deus os acompanhe por onde quer que andarem. Muito obrigado!

RESUMO

É proposta uma metodologia de análise de disponibilidade para sistemas complexos baseada num modelo de inferência para as taxas falha e de reparo de um processo markoviano. Para tanto, influências causais entre variáveis monitoradas que influenciam direta ou indiretamente o tempo entre falhas e de reparo são analisadas, tais como: qualidade do equipamento instalado, tipo da última manutenção realizada, fatores que causam a falha dos equipamentos, a forma como os equipamentos falham e outras variáveis monitoradas do sistema. Para representar as relações causais dessas variáveis, serão utilizadas as redes bayesianas no contexto da extração do conhecimento de base de dados. De acordo com essa abordagem, o modelo proposto irá fazer uso de uma base de dados para determinar a topologia da rede formada pelas variáveis monitoradas e o tempo entre falhas e de reparo dos equipamentos. Com as redes bayesianas estruturadas, torna-se possível, por meio do teorema de Bayes, atualizar as distribuições de probabilidade dos tempos dado um evento de manutenção ou uma nova evidência em alguma variável da rede. Uma base de dados de um sistema complexo no campo de produção de óleo será utilizada como um exemplo de aplicação.

Palavras-chave: Engenharia de Produção; Redes bayesianas; Processos markovianos; Confiabilidade de sistemas.

ABSTRACT

A methodology of availability analysis for complex systems is proposed based on inference model for equipments failure and repair rates to transition in a Markov process. To do so, causal influences among measurable system variables that influence directly or indirectly the failure and repair times will be analyzed, such as: quality of the installed equipments, kind of the last realized maintenance, factors that cause the failure of the equipments, the way how equipments fail and monitored variables of the system. To represent the causal influences of these variables, Bayesian Belief Networks (BBNs) will be used in a Knowledge Discovery in Databases (KDD) context. According to this approach, the proposed model will make use of a database to determine the topology of the network comprised by the measured variables and the failure and repair times of equipments. With that BBN, through Bayes's theorem, it is possible to update the times probability distribution given a maintenance event or a new piece of evidence on any variable in the network. A real database of a complex system in the context of oil production will be used as an example application.

Keywords: Production Engineering; Bayesian belief networks, Markovian process; Systems reliability.

SUMÁRIO

AGRADECIMENTOS	iv
RESUMO	iii
ABSTRACT	iv
1 INTRODUÇÃO	1
1.1 Justificativa	4
1.2 Objetivos do Trabalho	4
1.2.1 Objetivo Geral	4
1.2.2 Objetivos Específicos	5
1.3 Organização da Dissertação	5
2 BASE CONCEITUAL	7
2.1 Introdução	7
2.2 Extração de Conhecimento via Base de Dados	7
2.3 Elementos Probabilísticos para a Construção e Atualização de uma Rede Bayesiana	9
2.3.1 Variáveis Aleatórias	9
2.3.2 Distribuição de Probabilidade	9
2.4 Redes Bayesianas	10
2.4.1 Representação de uma Rede Bayesiana	12
2.4.2 Probabilidades Condicionais	12
2.4.3 Condição Markoviana	14
2.4.4 Independência Condicional	14
2.4.5 Equivalência Markoviana	16

2.5	Cálculo de Atualização da Rede Bayesiana	17
2.6	Mineração de Dados com Redes Bayesianas	19
2.7	Algoritmos de Busca e Pontuação	21
2.7.1	Função de Avaliação via <i>Minimum Description Length</i> (MDL)	22
2.7.2	Um Algoritmo de Aprendizagem – <i>Sum Of Products</i> (SOP)	24
2.8	As Taxas de Falha e Reparo	27
2.9	Parametrização dos Tempos de Falha e Reparo	28
2.9.1	As Distribuições de Probabilidades	29
2.10	Processos Markovianos	32
2.11	Análise de Disponibilidade com Modelos Markovianos	34
3	DESENVOLVIMENTO DA METODOLOGIA PROPOSTA	36
3.1	Introdução	36
3.2	Identificação dos Equipamentos e Variáveis Relevantes que Influenciam os Tempos entre Falhas e de Reparo do Sistema	40
3.3	Realização de uma Limpeza e Integração dos Dados	40
3.4	Estruturação das Relações Causais entre as Variáveis do Sistema e os Tem- pos entre Falhas e de Reparo	41
3.5	Estimação dos Parâmetros das Redes Bayesianas Estruturadas	42
3.6	Estruturação da Cadeia Markoviana do Sistema	43
3.7	Análise de Disponibilidade com a Cadeia Markoviana	44
3.8	Atualização do Estudo de Disponibilidade via Evidências nas Variáveis das Redes Bayesianas dos Equipamentos do Sistema	52
4	APLICAÇÃO DA METODOLOGIA PROPOSTA	54
4.1	Introdução	54
4.2	Intervenção em Poços de Petróleo	54
4.3	Os Tipos de Poços Analisados	55
4.3.1	Poços de Completação Convencional	55
4.3.2	Métodos de Elevação Artificial	55
4.3.3	Bombeio Mecânico com Hastes	56
4.3.4	Bombeio por Cavidades Progressivas	57
4.4	Aplicação da Metodologia	58

4.4.1	Equipamentos e Variáveis Relevantes	59
4.4.2	Limpeza e Integração dos Dados (Pré-Processamento)	66
4.4.3	Estruturação das Relações Causais entre as Variáveis do Sistema e os Tempos entre Falhas e de Reparo	66
4.4.4	Estimação dos Parâmetros das Redes Bayesianas Estruturadas	69
4.4.5	Estruturação da Cadeia Markoviana do Sistema	70
4.4.6	Análise de Disponibilidade com a Cadeia de Markov	74
4.4.7	Atualização do Estudo de Disponibilidade via Evidências nas Var- iáveis das Redes Bayesianas dos Equipamentos do Sistema	77
5	CONCLUSÕES E TRABALHOS FUTUROS	81
5.1	Conclusões	81
5.1.1	O Aprendizado das Redes Bayesianas	82
5.1.2	Algoritmos de Estimação da disponibilidade de Sistemas	82
5.1.3	A Metodologia Proposta	83
5.1.4	O Estudo de Caso	85
5.2	Sugestões para Trabalhos Futuros	86
	REFERÊNCIAS BIBLIOGRÁFICAS	88

LISTA DE FIGURAS

2.1	Etapas do processo de extração de conhecimento da base de dados	7
2.2	Exemplo de tabela de probabilidade condicional	13
2.3	Conexão serial	15
2.4	Conexão divergente ou do tipo causa comum	16
2.5	Conexão convergente ou do tipo efeito comum	16
2.6	Exemplo de estruturas equivalentes markovianas	17
2.7	Teste de ruptura do fluxo de informação	25
2.8	Curvas de comportamento das taxas de falha por tipo de equipamento	28
2.9	Exemplo de densidades de probabilidade Weibull	30
2.10	Exemplo de densidades de probabilidade exponencial	31
2.11	Exemplo de densidades de probabilidade Normal	31
2.12	Exemplo de densidades de probabilidade LogNormal	32
3.1	Passos metodológicos da abordagem proposta	38
3.2	Esquema da metodologia proposta	39
3.3	Exemplo de rede para cálculo da TPC	42
3.4	Exemplo de disponibilidade para um processo não-homogêneo de Markov	48
3.5	Simulação da disponibilidade de um sistema em série	50
3.6	Cadeia de markov de um sistema em série	50
3.7	Comparação dos resultados fornecidos pelo algoritmo proposto 2 e pela Equação 2.21	51
3.8	Disponibilidade fornecida pelos algoritmos 1 e 2	52
4.1	Poço com método de elevação artificial de bombeio mecânico com hastes	57
4.2	Poço com método de elevação artificial de bombeio por cavidades progres- sivas	58

4.3	Cadeia de Markov para o sistema de subsuperfície de um poço de produção de óleo	59
4.4	BBN da bomba BCP para o TBF	67
4.5	BBN da bomba BCP para o TTR	67
4.6	BBN da bomba BM para o TBF	68
4.7	BBN da bomba BM para o TTR	68
4.8	BBN da haste para o TBF	69
4.9	BBN da haste para o TTR	69
4.10	BBN do tubo para o TBF	70
4.11	BBN do tubo para o TTR	70
4.12	Cadeia de Markov para o sistema de subsuperfície de um poço	71
4.13	Distribuição de probabilidade para o TBF e o TTR da bomba BCP	72
4.14	Distribuição de probabilidade para o TBF e o TTR da bomba BM	73
4.15	Distribuição de probabilidade para o TBF e o TTR da haste	73
4.16	Distribuição de probabilidade para o TBF e o TTR do tubo	73
4.17	Disponibilidade para o sistema poço com método de elevação centrífugo	74
4.18	Disponibilidade para o sistema poço com método de elevação mecânico	75
4.19	Disponibilidade para o sistema poço com método de elevação mecânico (20 anos)	76
4.20	Disponibilidade para o sistema poço com método de elevação centrífugo (20 anos)	76
4.21	Disponibilidade para o sistema poço com método de elevação mecânico completamente sem renovação (20 anos)	77
4.22	Disponibilidade para o sistema poço com método de elevação centrífugo completamente sem renovação (20 anos)	77
4.23	Distribuição de probabilidade do TBF da bomba BCP antes da evidência	79
4.24	Distribuição de probabilidade do TBF da bomba BCP após evidência	79
4.25	Disponibilidade para o sistema poço com método de elevação centrífugo após evidência	80
4.26	Disponibilidade para o sistema poço com método de elevação mecânico após evidência	80

LISTA DE TABELAS

2.1	Exemplo de distribuição marginal	10
2.2	Probabilidades conjuntas do exemplo tempo de uso do equipamento	18
2.3	Probabilidades <i>a posteriori</i> do exemplo tempo de uso do equipamento	18
3.1	Exemplo de TPC	43
3.2	Exemplo de um sistema em série	50
4.1	Lista de equipamentos encontrados na base de dados	60
4.2	Lista de variáveis relevantes para os tempos entre falhas e de reparo	61
4.3	Lista de variáveis a serem inseridas nas BBN's da bomba BCP	63
4.4	Lista de variáveis a serem inseridas nas BBN's da bomba BM	63
4.5	Lista de variáveis a serem inseridas nas BBN's da Haste	64
4.6	Lista de variáveis a serem inseridas nas BBN's do Tubo	65
4.7	TPC do tempo até o reparo dado o nível de perda para a BBN do tubo	71
4.8	Caracterização dos nós da cadeia de Markov	71
4.9	Indicadores para a análise de disponibilidade do sistema BCP	78
4.10	Indicadores para a análise de disponibilidade do sistema BM	78

1 INTRODUÇÃO

A decisão de investir em negócios é calcada nos retornos esperados das aplicações. Algumas teorias, tal como a de Markowitz, propõem medidas de análise de investimentos baseadas no retorno esperado e no risco que o portfolio destes propiciam, sendo este risco freqüentemente representado como o desvio padrão que a carteira de investimentos traz (NETO, 2003). Esses investimentos, por sua vez, possuem retornos esperados que são definidos como uma função dos ganhos por unidade de tempo, os quais são fortemente influenciados pelos seus níveis de disponibilidade.

Tem-se assim que, quanto maior a disponibilidade de um sistema, maior será seu ganho médio por unidade de tempo, uma vez que menos tempo será necessário para se atingir os ganhos esperados. Por outro lado, é verdade que existe um ponto ótimo de disponibilidade, no qual além desse, os ganhos são reduzidos, visto que o custo marginal dos recursos necessários para trazer um incremento na disponibilidade se torna maior do que o ganho marginal trazido por esse incremento na disponibilidade. Mas, nesse trabalho, será assumido que tal ponto ainda não fora alcançado.

Na literatura da área de confiabilidade e disponibilidade de sistemas, tem-se a definição de disponibilidade como sendo a probabilidade de que um sistema esteja em condição operacional num instante t (LAFRAIA, 2001). E, após decorrido certo tempo, em que os efeitos transitórios do tempo são eliminados, a disponibilidade pode ser entendida como a fração de tempo em que o sistema estará operacional conforme foi planejado (disponibilidade estacionária ou assintótica) (COCHRAN; MURUGAN; KRISHNAMURTHY, 2001).

Com essa definição inicial do que vem a ser disponibilidade, torna-se possível ilustrar porque é importante se investir tempo nessa análise. Considerando um caso real de um poço produtor de petróleo na Bacia de Campos, o qual tem uma capacidade média de produção de 4,2 mil barris por dia¹, e considerando ainda o preço de venda do barril a US\$65,47², pode-se ter um raciocínio monetário para a disponibilidade da seguinte forma: suponha que um poço esteja em operação com uma disponibilidade de 98% e que, após registros de falhas nos equipamentos desse sistema, a disponibilidade seja reduzida

¹Dados de produção da Bacia de Campos referentes ao mês de junho – 400 poços com produção média diária de 1.682 mil barris por dia (PETROBRAS, 2006).

²Cotação de venda do barril de petróleo do tipo *pesado* no dia 10 de agosto de 2006 (GLOBO, 2006).

para o patamar de 95%, assim, ter-se-á ao ano um prejuízo aproximado de US\$ 3 milhões de dólares em conseqüência dos dias parados provenientes da queda no patamar da disponibilidade. Fica então registrado, por meio de uma breve ilustração em um caso possível na indústria petrolífera, a importância de se realizar uma análise mais detalhada da disponibilidade dos equipamentos.

Para a análise de disponibilidade, as ferramentas matemáticas de cunho probabilístico têm sido largamente utilizadas, pois permitem analisar os sistemas considerando as incertezas do “mundo real” (DROGUETT, 2002). Dentre essas ferramentas, destacam-se os processos de Markov, os quais têm sido utilizados para tratar problemas em que o reparo através da manutenção é possível.

Alguns exemplos de uso dessa ferramenta podem ser enumerados:

- Lai et al. (2002) faz uso dos processos homogêneos de Markov para analisar a disponibilidade de sistemas computacionais distribuídos (*software* e *hardware*). Uma comparação também é feita caso o sistema não fosse distribuído.
- Ouhbi e Linnios (1997) faz uso dos modelos semi-markovianos³ para a análise de disponibilidade de um rotor de um gerador em uma das unidades geradoras de energia da França. Nesse trabalho é feita uma comparação com os resultados obtidos de um trabalho com abordagem bayesiana de Milano (1995).

De forma semelhante, tentando captar as mudanças ocorrentes no “mundo real”, estão as redes bayesianas, as quais fazem uso de um forte raciocínio bayesiano. Nessa forma de pensar, os resultados *a posteriori* de um dado estado são uma conseqüência do que já se sabe a seu respeito (estimativas *a priori*) e das observações realizadas no transcorrer da operação do sistema (funções de verossimilhança). Com essa ferramenta, é possível não só se estimar a disponibilidade com dados ocorridos no passado, mas também com fatos que ocorrem no presente.

Por volta dos últimos dez anos é que as redes bayesianas começaram a ser utilizadas na área de confiabilidade (SIGURDSSON; WALLS; QUIGLEY, 2001). Da mesma forma, algumas aplicações nesse área podem ser enumeradas:

- Webera e Jouffe (2006) mostra por meio das redes bayesianas, o impacto de decisões na política de manutenção sob a disponibilidade de um sistema de aquecimento de

³Processos markovianos que consideram o tempo de permanência nos estados como fator influenciante nas probabilidades de transição.

água num processo industrial. Nesse artigo, as redes bayesianas são utilizadas, como uma forma alternativa à árvore de falhas, para modelar modos de falha dependentes e para propagar as ocorrências desses modos no modelo de confiabilidade do sistema.

- Bai et al. (2005) ilustra um exemplo de aplicação das redes bayesianas na análise de disponibilidade em sistemas de aviação. Nesse artigo, é utilizado um modelo de predição de falhas em *software*, onde as redes bayesianas atuam como uma ferramenta de captação de ocorrência de novas falhas para atualização dos parâmetros de confiabilidade do modelo.
- Milano (1994) aplica as redes bayesianas como um modelo de validação em tempo real de dados provenientes de diversos sensores que captam falhas nos motores de foguetes da NASA;
- Sigurdsson, Walls e Quigley (2001) expõe um exemplo de elicitación de conhecimento de especialistas para estruturar a topologia e os parâmetros de uma rede bayesiana. Nessa aplicação, busca-se estimar a probabilidade de sucesso de um determinado produto que está em fase de desenvolvimento. Além disso, esse trabalho traz uma visão geral de diversas aplicações de redes bayesianas em confiabilidade.

Ainda hoje, a maioria das aplicações na área de confiabilidade ora somente fazem uso dos processos markovianos para a análise da disponibilidade, deixando de lado a atualização dos parâmetros por meio de observações do que ocorre no “mundo real”, ora somente utilizam as redes bayesianas para a recepção de evidências, desprezando a simplicidade trazida pelo modelo markoviano (SMYTH, 2005). Com isso, observa-se na literatura uma lacuna de trabalhos que tratem o problema do hibridismo e da *interface* entre os processos markovianos e as redes bayesianas.

Calcado na necessidade de modelar e estimar a disponibilidade dos sistemas complexos, é que se pauta o desenvolvimento desse trabalho, fazendo uso de uma base de dados e de um modelo híbrido para tanto. Inicia-se, assim, o processo de preenchimento da lacuna literária, até o momento existente, para essa abordagem híbrida.

1.1 Justificativa

Com o crescente aumento da complexidade dos sistemas utilizados nas empresas, tem surgido a necessidade de se monitorar cada equipamento que formam seus processos. Aliado a isto, tem crescido a capacidade de armazenamento de dados, o que faz ser imprescindível a utilização de ferramentas que possibilitem a extração de conhecimentos desses dados a fim de suportar os decisores.

No campo de aplicação da confiabilidade, essas bases de dados têm se apresentado de maneira relacional, em que cada cada variável monitorada do sistema compõe uma variável dessas bases.

Muitas abordagens têm sido utilizadas para tentar extrair conhecimento dessas bases de dados, quer seja com a utilização de métodos descritivos (estatística descritiva), quer seja via métodos de inferência mais sofisticados (análise de *clusters*, mineração de dados, análise de regressão, etc.). Mas, o grande problema de tais abordagens é o de retratar o comportamento dos sistemas dinâmicos de forma genérica e simples.

Esse trabalho vem, portanto, desenvolver uma metodologia para avaliação de sistemas complexos. A metodologia proposta é híbrida por natureza, fazendo uso de processos markovianos para a descrição do comportamento dinâmico de um sistema e utilizando redes bayesianas para a caracterização das relações causais entre as variáveis/atributos que influenciam os tempos de transição (tempo entre falhas e de reparo) do processo markoviano de um sistema. Além do mais, a metodologia faz uso de uma base de dados para extrair conhecimento necessário à construção da topologia das redes bayesianas, caracterizando assim as distribuições de probabilidade dos tempos entre falhas e de reparo.

1.2 Objetivos do Trabalho

Atentando para a necessidade de estimar a disponibilidade dos sistemas, este trabalho buscará atingir os seguintes objetivos seguir.

1.2.1 Objetivo Geral

Este trabalho tem como objetivo geral:

- Elaborar uma metodologia para estimação dinâmica da disponibilidade de sistemas

complexos em degradação por meio do hibridismo entre os processos markovianos e as redes bayesianas.

1.2.2 Objetivos Específicos

Este trabalho tem como objetivos específicos:

- Realizar uma revisão literária sobre extração de conhecimento de base de dados, enfatizando o aprendizado via redes bayesianas;
- Fazer uma revisão dos conceitos utilizados na análise de disponibilidade de sistemas;
- Definir passo a passo a forma de interação entre as redes bayesianas e a cadeia de Markov com distribuição de probabilidade não-paramétrica dos tempos de transição para um modelo de análise de disponibilidade de sistemas;
- Desenvolver e implementar em linguagem de programação um algoritmo para cálculo de disponibilidade de sistemas complexos via processos markovianos com distribuições de probabilidade não-paramétricas para os tempos de transição;
- Ilustrar a metodologia proposta em um exemplo de aplicação.

1.3 Organização da Dissertação

No segundo capítulo, será realizado um embasamento teórico que dará suporte ao trabalho a ser realizado nos outros capítulos, sendo abordados tópicos sobre extração de conhecimento de base de dados, redes bayesianas e processos markovianos no contexto da análise da disponibilidade de sistemas. No terceiro capítulo, uma metodologia de como estruturar e analisar sistemas no contexto da disponibilidade, levando em consideração os tempos entre falhas e de reparo de seus componentes fornecidos por uma base de dados, será desenvolvida. Ainda neste mesmo capítulo, será proposto um algoritmo para cálculo das probabilidades de estados de uma cadeia de Markov via simulação de eventos discretos, em que não será necessário assumir tempos exponenciais de transição e, especialmente, nenhuma outra distribuição paramétrica. No quarto capítulo, uma contextualização da metodologia proposta será realizada em sistemas de produção de óleo (poços de petróleo).

E, por fim, no último capítulo, será realizada uma relação dos principais pontos abordados neste trabalho, avaliando os pontos a ponderar e os a considerar.

2 BASE CONCEITUAL

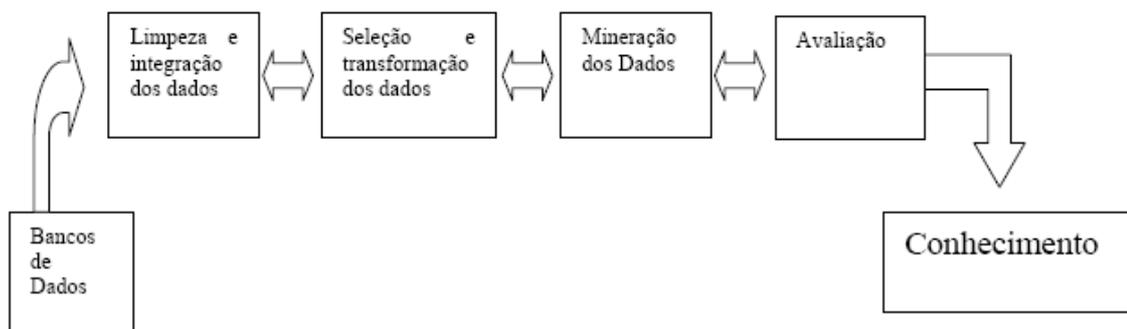
2.1 Introdução

O presente capítulo abordará os conceitos fundamentais a serem utilizados neste trabalho. Para tanto, serão realizadas discussões a respeito do que vem a ser o conhecimento a ser extraído via base de dados e seus elementos, o que vem a ser uma rede bayesiana, algoritmos de aprendizagem de estruturas de redes bayesianas e elementos probabilísticos essenciais para a análise de disponibilidade de sistemas (taxas de falha e reparo, curva da banheira, distribuições de probabilidade e modelos markovianos).

2.2 Extração de Conhecimento via Base de Dados

Quando se fala em extração ou descoberta do conhecimento através de uma base de dados (*Knowledge Discovery in Databases – KDD*) não está se reportando ao conhecimento extraído através do pensamento hipotético-dedutivo. O autor acredita que a nomenclatura mais adequada seria informação extraída através de uma base de dados, mas como o termo conhecimento é adotado nesta área de pesquisa, o mesmo será utilizado.

O processo de extração de conhecimento da base de dados (KKD) conhecido na literatura é subdividido conforme a Figura 2.1.



Fonte: Hrudchka (2003, p.5).

Figura 2.1: Etapas do processo de extração de conhecimento da base de dados

É importante destacar que é na etapa de mineração de dados que os padrões são detectados e o conhecimento é gerado, sendo assim o motivo pelo qual esse processo de

extração de conhecimento é amplamente divulgado como mineração de dados.

Limpeza e Integração de Dados (passo 1). Fase que é também conhecida como a de pré-processamento. Nela, realiza-se uma limpeza nos dados objetivando filtrar os ruídos que fazem com que os dados não representem a amostra analisada.

Outro fator importante que deve ser analisado é ausência de dados num determinado ponto da amostra. Frequentemente, adota-se a prática de eliminar o caso da amostra que possui algum dado ausente, o que pode reduzir significativamente o tamanho da amostra, dificultando a extração do conhecimento. Segundo Hrudchka (2003), algumas causas que levam à ausência de dados são: valor não informado pela fonte de dados, erro de operação ou digitação ao realizar o registro dos dados, falha no sistema de manuseio dos dados e impossibilidade de efetuar o registro do dado devido a não presença de campo apropriado na base de dados (exemplo: tem-se numa amostra o campo complemento para endereço, mas o mesmo não é preenchido pois trata-se de uma casa).

Seleção e Transformação dos Dados (passo 2). A seleção e transformação dos dados objetiva identificar quais os atributos (variáveis) que são relevantes para a geração do conhecimento, reduzindo, inclusive, o tempo de processamento dos algoritmos.

Mineração dos Dados (passo 3). Esta é a etapa fundamental do processo, pois como dito anteriormente, objetiva encontrar padrões, associações ou *clusters* existentes na base de dados.

Avaliação e Apresentação (passo 4). Nem todos os resultados oriundos da mineração de dados apresentam relevância para o processo de captação de conhecimento. Segundo Han e Kamber (2001), os motivos pelos quais os resultados são irrelevantes são:

- Apresenta dificuldade para a compreensão humana;
- Possui incerteza sobre o padrão encontrado;
- Há dificuldade para a utilização do padrão;
- O padrão descoberto não evidencia aspectos inovadores.

Segundo Munteanu e Bendou (2001), o aprendizado de redes bayesianas via dados é uma das mais ambiciosas abordagens de extração de conhecimento de base de dados. E, seguindo este pensamento, a seguir será realizada uma explanação a respeito das redes bayesianas, mas antes, considerações sobre elementos probabilísticos que compõem as redes bayesianas se fazem necessárias.

2.3 Elementos Probabilísticos para a Construção e Atualização de uma Rede Bayesiana

2.3.1 Variáveis Aleatórias

Uma variável aleatória X é uma função que associa cada elemento do espaço amostral S a um número real de seu contradomínio S_X : $X : S \mapsto S_X$. Por exemplo, seja o experimento E lançar uma moeda duas vezes e verificar o número de coroas (K), tem-se:

- O conjunto de possíveis valores $S = \{CC, CK, KC, KK\}$;
- Os valores da variável aleatória (número de coroas) $S_X = \{0, 1, 2\}$ com $X(CC) = 0$, $X(KC) = X(KC) = 2$ e $X(KK) = 2$.

Assim, as respectivas probabilidades de ocorrência de cada evento são: $p(X = 0) = p(X = 2) = 1/4$ e $p(X = 1) = 1/2$.

Assume-se uma variável como discreta quando os elementos de seu domínio podem ser contados, e contínua quando não. É importante realizar esta distinção pois mais adiante serão descritos métodos próprios para variáveis discretas e contínuas.

2.3.2 Distribuição de Probabilidade

“Uma distribuição de probabilidade é a coleção de todas as probabilidades dos valores possíveis de uma variável aleatória”.

(DEGROOT; SCHERVISH, 2002, p.97)

Tabela 2.1: Exemplo de distribuição marginal

Fumante	câncer		Marginal de fumante
	sim	não	
sim	0,3	0,15	0,45
não	0,1	0,45	0,55
Marginal de câncer	0,4	0,6	

Distribuição Conjunta

Faz-se uso da distribuição conjunta quando tem-se interesse em calcular a probabilidade de um evento conjunto que considera o acontecimento de situações específicas em duas ou mais variáveis aleatórias. Por exemplo: suponha que quer-se saber a probabilidade que uma pessoa seja fumante e tenha câncer. Nesse caso, tem-se interesse no acontecimento simultâneo de duas variáveis, o que faz necessário o uso da distribuição conjunta $P(X, Y)$ ou $P(X \cap Y)$.

Distribuição Marginal

A idéia de distribuição marginal decorre da distribuição conjunta, em que uma das variáveis é marginalizada na análise. Por exemplo: suponha que tem-se a distribuição conjunta de probabilidade de uma pessoa ter câncer e ela ser fumante. Assim, a distribuição marginal de uma pessoa ter câncer seria aquela que desprezaria o fato da pessoa ser fumante ou não, ver Tabela 2.1.

Distribuição Condicional

A distribuição marginal $P(X|Y)$ é utilizada para representar a distribuição de probabilidade de uma variável X condicionada ao valor de uma outra variável Y . A forma de cálculo da distribuição condicional é $P(X|Y) = P(X, Y)/P(Y)$

2.4 Redes Bayesianas

As redes bayesianas em seu estado da arte é uma representação probabilística do conhecimento. Em sua representação qualitativa elas evidenciam o conhecimento via grafo (conjunto de arcos e nós) acíclico e orientado (GAO) o qual mostra as variáveis e

suas interrelações. Em sua representação quantitativa o grau de dependência entre as variáveis é representado por distribuições de probabilidades condicionais.

Assume-se que, quando há arco não direcionado entre dois estados, trata-se do mesmo estado.

Cada variável ou estado é denotado por $X_i = \{x_1, \dots, x_k\}$ em que X_i assume seus possíveis valores de forma mutuamente exclusivos.

Cada variável possui uma probabilidade condicional que quantifica o efeito de seus pais¹ sobre ela. Essas probabilidades condicionais, por sua vez, são atualizadas via regra de Bayes ou pelo teorema da probabilidade total.

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \quad (2.1)$$

$$p(x) = \sum_{\theta} p(x|\theta)\pi(\theta) \quad (2.2)$$

As redes bayesianas são pouco influenciadas por amostras de pequenos tamanhos (LEE; ABBOTT, 2003). Isto é fácil entender visto que se trata de um conhecimento formado por uma composição que inclui a opinião a priori de especialista, o que não necessita de dado para formá-la.

Lee e Abbott (2003) sugerem alguns passos a serem seguidos na estruturação de uma rede bayesiana:

1. Identificar os estados relevantes para a análise;
2. Identificar as relações causais entre as variáveis;
3. Eduzir² a probabilidade a priori de ocorrência dos estados ou uma tabela condicional de probabilidades;

É importante destacar que o item 3 retrata a elaboração quantitativa de uma rede bayesiana via opinião de especialista. Mas, como será discutido mais adiante, é possível estimar essas probabilidades por intermédio de um banco de dados. Como não é do escopo desse trabalho tratar o problema de elaboração quantitativa da rede bayesiana via opinião de especialistas, logo não será abordado aqui. Para uma maior discussão sobre esse assunto, ver Droguett et al. (2006).

¹Estados de onde partem arcos direcionados apontando para a variável filho

²Também referenciado como elicitar.

2.4.1 Representação de uma Rede Bayesiana

Uma rede bayesiana é formada por nós que representam variáveis aleatórias e arcos que evidenciam as relações causais (relações de causa e efeito) entre os nós. A variável de onde parte o arco é denominada de pai da variável para onde o arco aponta e da mesma forma o pai é considerado uma provável causa do filho. Generalizando esta idéia, tem-se os ancestrais de um nó A, aqueles que, seguindo os sentidos das setas, permitem chegar ao nó A, e os descendentes de A aqueles que são alcançados seguindo o sentido dos arcos partindo de A.

As redes bayesianas são representadas formalmente como um par $N = \{(V, E), P\}$ em que (V, E) é um grafo direcionado o qual representa os nós (V) e os arcos (E) e P é a distribuição de probabilidade a respeito de V. É importante lembrar que quando se parte de um nó não é possível retornar ao mesmo (condição acíclica).

2.4.2 Probabilidades Condicionais

Até este ponto já foram identificadas os passos 1 e 2 do sugerido por Lee e Abbott (2003), seguindo então para o passo 3.

Como já explicitado, a relação quantitativa entre as variáveis é definida através de distribuições de probabilidade marginais que para o caso de variáveis discretas serão representadas por tabelas de probabilidade condicionais (TPC).

Como aplicação da TPC, tem-se um exemplo adaptado de Menezes (2005) de como o tempo de uso da máquina influencia o tempo de reparo e o custo unitário de manutenção, como ilustrado na Figura 2.2. Neste exemplo é assumido que o tempo de uso de equipamento leva a um baixo ou alto nível de tempo e custo de manutenção. Para estruturar as TPC's é necessário evidenciar todas as combinações possíveis entre os valores de cada filho e seus pais, para daí estimar-se a probabilidade de ocorrência dessas combinações, quer pela opinião de especialistas quer de forma freqüentista utilizando uma base de dados. Assim, as variáveis foram dicotomizadas para facilitar o cálculo do exemplo. Um exemplo de como se estimar essas probabilidades condicionais de forma freqüentista será mostrado na seção 3.5.

Para a variável tempo de uso de equipamento assumiu-se que a variável receberia valor 0 quando seu verdadeiro valor fosse inferior a x' dias e 1 em caso contrário. A variável tempo de manutenção foi dicotomizada em valor baixo (0 para valores inferiores ou iguais

a y' dias) e alto (1 para valores superiores a y' dias). E a variável custo de manutenção também foi dicotomizada em valor baixo (0 para valores inferiores a z' reais) e alto (1 para valores superiores a z' reais).

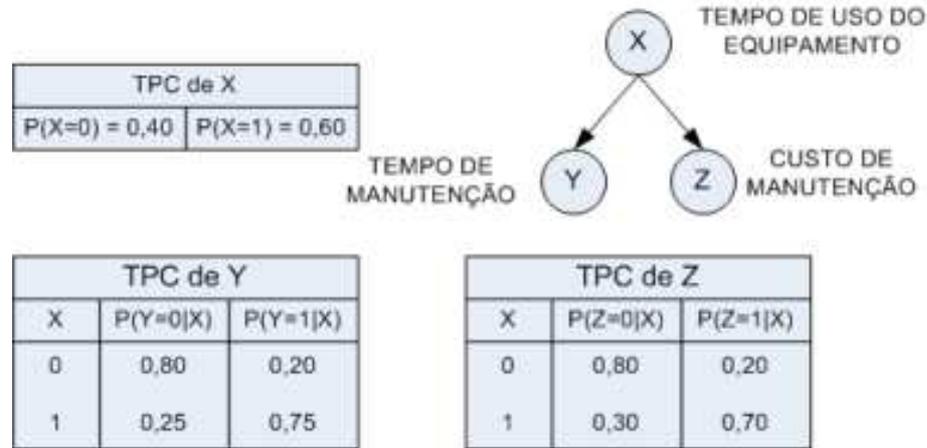


Figura 2.2: Exemplo de tabela de probabilidade condicional

Aplicando as Equações 2.1 e 2.2 neste caso tem-se:

$$P(z|x) = \frac{P(z \cap x)}{P(x)} \quad \therefore P(z \cap x) = P(z|x)P(x) \quad \therefore P(z) = \sum_x P(x)P(z|x)$$

Substituindo os valores da figura 2.2 na equação anterior, tem-se:

$$P(Z = z_0) = 0,4 \cdot 0,2 + 0,6 \cdot 0,7 = 0,5 \quad (2.3)$$

Para y :

$$P(y|x) = \frac{P(y \cap x)}{P(x)} \quad \therefore P(y \cap x) = P(y|x)P(x) \quad \therefore P(z) = \sum_x P(x)P(y|x)$$

Assim:

$$P(Y = y_0) = 0,4 \cdot 0,2 + 0,6 \cdot 0,75 = 0,53 \quad (2.4)$$

2.4.3 Condição Markoviana

Definição 2.4.1 (Condição Markoviana) *Seja P a distribuição de probabilidade conjunta das variáveis aleatórias de um conjunto V e seja (V, E) um GAO. É dito que $N = \{(V, E), P\}$ satisfaz à condição Markoviana se para cada V , em que $X \in V$, X é condicionalmente independente de todas as variáveis que compõem o conjunto dos seus não descendentes, dados os valores das variáveis que compõem seu conjunto de pais. Ou, em notação probabilística, $\{X\} \text{ ind } ND(X)|PA(X)$, em que $ND(X)$ refere-se ao conjunto de não descendentes de X e $PA(X)$ ao seu conjunto de pais.*

Dito de outra maneira, é dito que não há dependência direta entre quaisquer dois nós, a não ser que exista um arco unindo-os.

A conjectura a seguir, expande o conceito de dependência acima citado de uma forma mais ampla.

Conjectura 2.4.1 (Princípio da Causa Comum) *Se duas variáveis são probabilisticamente dependentes, então ou uma causa a outra (diretamente ou indiretamente) ou elas possuem um ancestral comum.*

Este princípio diz que, em essência, por trás de uma dependência probabilística existe uma dependência causal que a explica. Os algoritmos que fazem uso da mineração de dados irão buscar uma estrutura de dependência para explicar as dependências probabilísticas.

2.4.4 Independência Condicional

“É importante entender a relação entre a independência condicional e a estrutura da rede bayesiana para saber como ela funciona”.

(KORB; NICHOLSON, 2004)

Além disso, é importante vislumbrar esta relação para poder entender as heurísticas³ dos algoritmos de montagem de estruturas de redes bayesianas via dados.

A independência condicional entre variáveis pode reduzir a inferência da rede bayesiana para o cálculo da probabilidade de um dado estado. Para verificar se dois estados são

³Método de investigação baseado na aproximação progressiva de um dado problema (HOUAISS, 2001).

condicionalmente independentes basta verificar a propriedade *d-separação*. Diz-se que dois nós X e Y são *d-separados* se existe um nó intermediário Z que obedeça a uma das seguintes propriedades:

1. Z é um nó intermediário numa conexão serial ($X \rightarrow Z \rightarrow Y$ ou $Y \rightarrow Z \rightarrow X$) e Z é instanciado pela evidência.
2. Z tem uma conexão divergente entre X e Y e então Z é instanciado pela evidência ($X \leftarrow Z \rightarrow Y$);
3. Z é um nó intermediário numa conexão convergente ($X \rightarrow Z \leftarrow Y$) e nem Z e nenhum de seus descendentes tem recebido evidência.

Na situação 1, tem-se uma estrutura como mostrado na Figura 2.3 (caso $X \rightarrow Z \rightarrow Y$), com X como uma provável causa de Z e Z como uma provável causa de Y . Com base na condição markoviana, a probabilidade de $Y|Z$ é a mesma de $Y|(Z \cap X)$, ou seja, o fato de saber que X ocorreu não influencia a probabilidade de ocorrência ou crença de Y se já é sabido que Z ocorreu. Assim, Y é condicionalmente independente de X , dado que se sabe que Z ocorreu, ou em notação formal: $C \text{ ind } A|B$.

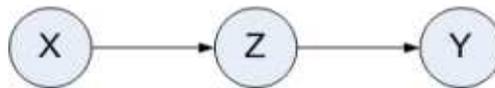


Figura 2.3: Conexão serial

A situação 2, como verificado na Figura 2.4, evidencia que quando duas variáveis X e Y têm causa comum Z e não existe evidências sobre Z , o fato de uma das conseqüências serem verificadas, X por exemplo, modifica a probabilidade de Z ocorrer, o que irá também alterar a probabilidade da conseqüência Y estar presente. Porém, se já se tem alguma evidência a respeito de Z , o fato da conseqüência X estar presente não diz nada sobre a probabilidade de se ter também a conseqüência Y . Assim, X é condicionalmente independente de Y , dado que Z ocorreu (em notação matemática: $X \text{ ind } Y|Z$).

Na situação 3, como verificado na Figura 2.5, a condição markoviana é violada devido à perda da independência condicional dada a observância do valor de uma variável que seja efeito comum de ao menos duas variáveis não descendentes entre si. A esse caso dar-se o nome de ativação de dependências (inglês: *explaining away*). Assim, uma variável

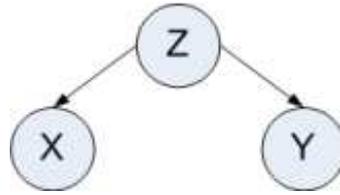


Figura 2.4: Conexão divergente ou do tipo causa comum

descendente é dita ser de efeito comum de um conjunto de variáveis não descendentes entre si se há ligações entre as variáveis desse conjunto e a mesma.

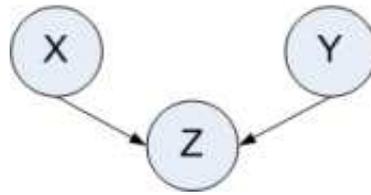


Figura 2.5: Conexão convergente ou do tipo efeito comum

2.4.5 Equivalência Markoviana

Este tema será muito abordado no momento em que se estiver comentando a respeito da aprendizagem topológica ou estrutural das redes bayesianas. Neste ponto, será dito que o aprendizado da estrutura é fidedigno até o ponto das classes equivalentes, pois estas apresentam propriedades que não se torna possível a distinção via base de dados.

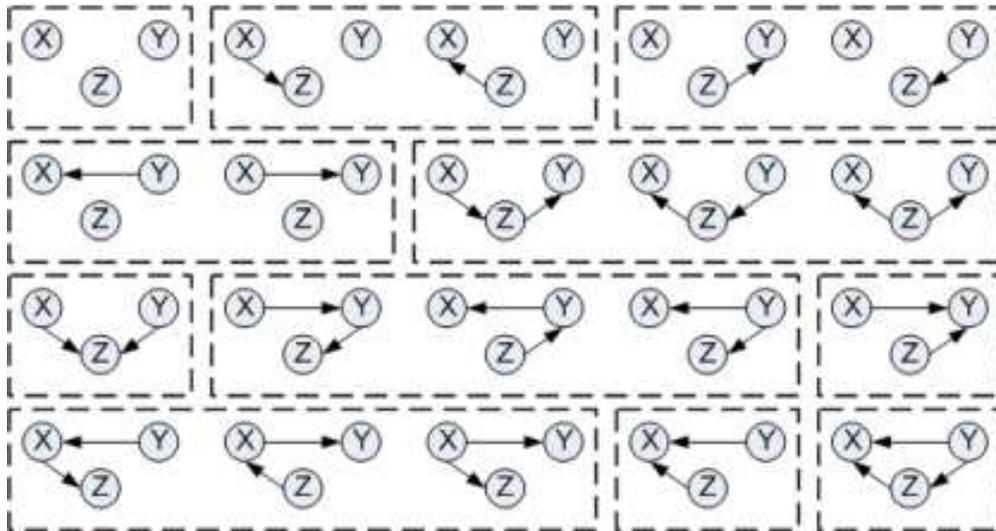
O que vem a ser a equivalência markoviana foi demonstrado por Verma e Pearl (1990) no teorema a seguir. O mesmo não será demonstrado pois foge ao escopo do trabalho.

Teorema 2.4.1 *Dois modelos causais que possuem as mesmas variáveis, o mesmo esqueleto e a mesma v-estrutura são equivalentes markovianos (VERMA; PEARL, 1990).*

Alguns exemplos são mostrados na itemização a seguir e na Figura 2.6. Nesta figura, as estruturas que estão dentro da mesma caixa tracejada são equivalentes markovianos.

- Todo modelo completamente conectado e que possuam as mesmas variáveis são equivalentes markovianos;
- $A \rightarrow B \rightarrow C$ e $A \leftarrow B \leftarrow C$ são equivalentes markovianos e não há a presença de v-estruturas;

- $A \rightarrow B \rightarrow D \leftarrow C$ e $A \leftarrow B \rightarrow D \leftarrow C$ são equivalentes markovianos e apenas a v-estrutura centrada em D é mantida.



Fonte: Korb e Nicholson (2004).

Figura 2.6: Exemplo de estruturas equivalentes markovianas

2.5 Cálculo de Atualização da Rede Bayesiana

Para realizar o cálculo de atualização das redes, é necessário calcular as distribuições conjuntas para cada conjunto de pais e filhos:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | pa(X_i)) \quad (2.5)$$

onde $P(X_i | pa(X_i))$ são as probabilidades de X_i em relação a seus pais.

Com a distribuição conjunta agora é possível calcular a distribuição a posteriori, o que será feito através da normalização da soma das probabilidades favoráveis à ocorrência do evento. Para melhor entender o processo de atualização, será dado continuidade ao exemplo iniciado na seção 2.4.2 quando se falava sobre o tempo de uso do equipamento como fator condicionante para o tempo de reparo e custo de manutenção. A seguir são evidenciadas as probabilidades conjuntas da v-estrutura apresentada anteriormente, quando calculadas a partir das TPC's e da expressão da probabilidade conjunta:

$$P(X, Y, Z) = P(Y|X)P(Z|X)P(X) \quad (2.6)$$

Tabela 2.2: Probabilidades conjuntas do exemplo tempo de uso do equipamento

$P(x_0, y_0, z_0) = 0,256$	$P(x_1, y_0, z_0) = 0,045$
$P(x_0, y_0, z_1) = 0,064$	$P(x_1, y_0, z_1) = 0,105$
$P(x_0, y_1, z_1) = 0,064$	$P(x_1, y_1, z_0) = 0,135$
$P(x_1, y_0, z_0) = 0,016$	$P(x_1, y_1, z_1) = 0,315$

Supondo a observância de um valor inferior a y' ($Y = y_0$) para a variável tempo de manutenção, prossegue-se para o cálculo das probabilidades a posteriori de toda a rede utilizando a Tabela 2.2 e a Equação 2.1. Nesse caso, como montamos uma rede com apenas três estados, será realizada apenas a atualização para os outros dois estados ($P(x_0|y_0)$, $P(x_1|y_0)$, $P(z_0|y_0)$ e $P(z_1|y_0)$), mas se existissem mais nós na rede, o restante também seria atualizado. Para $P(x_0|y_0)$ tem-se:

$$P(x_0|y_0) = \frac{P(x_0, y_0, z_0) + P(x_0, y_0, z_1)}{P(x_0, y_0, z_0) + P(x_0, y_0, z_1) + P(x_1, y_0, z_0) + P(x_1, y_0, z_1)} \quad (2.7)$$

Assim, os valores das probabilidades são mostrados na Tabela 2.3.

Tabela 2.3: Probabilidades a posteriori do exemplo tempo de uso do equipamento

$P(x_0) = 0,681$	$P(x_1) = 0,319$
$P(z_0) = 0,640$	$P(z_1) = 0,360$

Para comparação das probabilidades *a priori* e *a posteriori* de z_0 , por exemplo, após a observação de y_0 , deve-se utilizar os resultados da Equação 2.3 e Tabela 2.3 e notar que houve um acréscimo da ocorrência de baixos tempos de uso de equipamento.

Após discussão a respeito das redes bayesianas, será discutido o passo 3 da metodologia apresentada por Hrudchka (2003) na seção 2.2. O processo de seleção das variáveis, o passo 2, será discutido na metodologia proposta e no exemplo de aplicação no capítulo 4. A parte de limpeza e integração será exposta no capítulo de apresentação da metodologia proposta, quando se falará a respeito da imputação de dados.

É importante destacar que a metodologia apresentada por Hrudchka (2003) não detalha as formas de se executar a mineração de dados, mas apenas o propósito dessa etapa, uma vez que trata-se de um trabalho de imputação de dados. Porém, neste trabalho, será apresentada uma forma de realizar essa mineração via redes bayesianas.

2.6 Mineração de Dados com Redes Bayesianas

“Com o demonstrado potencial das redes bayesianas, interesses em métodos para automatizar a sua construção tem crescido enormemente nos últimos anos”.

(KORB; NICHOLSON, 2004, p.151)

Para utilizar estas duas ferramentas de maneira conjugada, a literatura tem mostrado duas fases distintas de desenvolvimento: a primeira em que é construído o grafo direcionado acíclico, chamado de estrutura da rede bayesiana, e a segunda que corresponde à avaliação das probabilidades à priori e o treinamento e teste da rede bayesiana com o cálculo das probabilidades à posteriori por intermédio das evidências. Para Silva e Ladeira (2002), enquanto o aprendizado dos parâmetros (análise das probabilidades) tem se mostrado relativamente simples, o aprendizado da estrutura gráfica se apresenta de forma muito complexa.

A elaboração da estrutura da rede bayesiana pode ser feita de duas diferentes formas:

- Manual, utilizando a experiência dos especialistas na área de análise;
- Automática ou semi-automática, via algoritmos de aprendizagem aplicados a uma base de dados relacional existente.

Na primeira maneira de criar a estrutura da rede, o conhecimento do especialista é fortemente necessitado, exigindo grande experiência no problema em questão. As relações causais nesse caso são confiáveis apenas no domínio de conhecimento do especialista.

Na segunda forma, a estruturação da rede bayesiana é suportada automaticamente ou semi-automaticamente por algoritmos de aprendizagem aplicados em descoberta de conhecimento em grandes bases de dados, em que a dependência ou independência das variáveis é avaliada pela relação direta entre os estados. As relações causais nesse caso são garantidas apenas no domínio dos dados.

Um dos grandes problemas do aprendizado da estrutura gráfica ou descoberta das relações de causa e efeito é justamente a explosão combinatória, que é tratada na literatura como uma busca gulosa, com espaço de busca de dimensão exponencial do número de variáveis envolvidas (ROBINSON, 1977 apud SILVA; LADEIRA, 2002).

Só para se ter uma idéia do problema que se tem em mão, Robinson (apud SILVA; LADEIRA, 2002) fornece uma estimativa da quantidade de redes no espaço de busca: $f(n) = \sum_{i=1}^n (-1)^{i+1} \frac{n!}{(n-i)!} f(n-i)$. Como exemplo, para uma rede com apenas dez variáveis, seria necessária a geração de $4,2 \cdot 10^{18}$ redes candidatas.

Além desse fato, Pearl (2000) mostrou que geralmente não se pode inferir sobre relações de causalidade baseado apenas em distribuições de probabilidade.

Existem duas abordagens distintas de algoritmos para montar uma estrutura ótima de rede bayesiana: a primeira são os algoritmos baseados em busca e pontuação em que são utilizados *scores* ou uma função de avaliação para quantificar o grau de adequação da estrutura⁴ aos dados, e o segundo são os algoritmos baseados em restrições os quais buscam uma melhor estrutura entre todas as possíveis dependências e independências condicionais via testes estatísticos.

Segundo Cheng et al. (2002), os algoritmos baseados em restrições são aconselhados para base de dados com grande quantidade de variáveis e casos, e os algoritmos baseados em métricas são aconselhados para redes com poucos nós e dados ruidosos. Uma primeira leitura nesse sentido, pode-se pensar que os métodos baseados em restrições são os mais aconselháveis para grandes base de dados pois supõe-se que os mesmos realizam os cálculos mais rapidamente, mas não. Este conselho é dado pois, como explica Munteanu e Bendou (2001), os algoritmos baseados em restrições são mais sensíveis à presença de “buracos”, o que numa base de dados grande esse efeito é dirimido. O mesmo raciocínio é posto para o conselho de utilizar os algoritmos baseados em pontuação em bases pequenas, uma vez que estes são mais insensíveis a falta de dados.

Utilizando a abordagem de algoritmos baseados em restrições é possível o especialista alimentar o algoritmo de aprendizagem para que o leve para o melhor modelo possível. Por outro lado, Heckerman, Geiger e Chickering (1995) compararam essas duas abordagens de aprendizado e mostraram que os métodos baseados em pontuação geralmente apresentam vantagens sob os baseados em testes de dependência. Mas, Cowell (2001) provou que para todo algoritmo de pontuação existe um equivalente na classe dos algoritmos de testes de dependência, e vice-versa.

A mineração de dados vem, portanto, ser uma metodologia de aplicação na primeira fase desse processo, auxiliando na construção da estrutura da rede bayesiana com uma

⁴Surge aqui o conceito de solução candidata, aquela rede utilizada como possível solução para representar a base de dados, a qual será avaliada via função de avaliação.

base de dados relacional existente, deixando de forma explícita a relação de causalidade entre as variáveis.

Além dessa aplicação, a integração dessas duas ferramentas tem sido utilizada para o preenchimento de dados faltantes em base de dados (imputação bayesiana), ver Hrudchka (2003) para maiores detalhes.

2.7 Algoritmos de Busca e Pontuação

Serão discutidas as idéias principais dos algoritmos de busca e pontuação e a seguir um algoritmo específico denominado SopLEQ.

Os algoritmos de busca e pontuação realizam uma varredura “gulosa” no espaço das estruturas equivalentes das redes bayesianas e terminam quando falham em revelar alguma relação de independência entre as variáveis, podendo, por esse motivo, apresentar discrepância com a verdadeira rede que representa a base de dados.

Grande parte dos trabalhos que criam algoritmos para a construção da estrutura de redes fazem uso de estruturas equivalentes de redes e, principalmente, utilizam estruturas de classes equivalentes ordenadas para checar a consistência e dar a pontuação, o que deixa os algoritmos significativamente lentos.

Sendo assim, estes algoritmos tratam o problema de aprendizagem como um problema de pesquisa clássico, em que são definidas uma função de avaliação que pontua as redes candidatas ao dados existentes, uma estratégia de pesquisa baseada nessa função e um conjunto de operadores que são aplicados para cobrir todo o espaço de busca.

Notadamente, no caso desses algoritmos, as estratégias são centradas nos operadores de acréscimo, supressão ou inversão de arcos. E, as funções de avaliação são diversas: *Minimum Description Length* (MDL), *Minimum Message Length* (MML), *Bayesian Dirichlet equivalent* (BDE), etc.

Essas funções de avaliação possuem uma característica matemática interessante, em que a pontuação de toda a rede é definida como a soma dos scores para cada nó e as classes equivalentes possuem a mesma pontuação (ver Teorema 2.4.1).

Um grande problema encontrado no aprendizado da estrutura topológica causal de uma rede bayesiana são justamente os ótimos locais, que por sua grande maioria, são causados por essas estruturas equivalentes. O algoritmo a ser descrito mais adiante irá tratar desse problema de equivalência, o que tradicionalmente é resolvido pela orientação

aleatória dos arcos das estruturas equivalentes, podendo gerar redes bastante diferentes das ótimas procuradas.

2.7.1 Função de Avaliação via *Minimum Description Length* (MDL)

Grunwald (2005), em seu recente trabalho, mostra diversas formas de interpretação e aplicações do comprimento de descrição mínimo (*Minimum Description Length* – MDL), mas é a forma de como pontuar redes bayesianas candidatas para a solução de um problema de busca que interessa a este trabalho.

A forma de pontuar uma rede causal via MDL teve grandes avanços baseados na idéia de Lam e Bacchus (1994), mas esta inferência foi originariamente desenvolvida por Rissanen (1978) baseada na idéia do comprimento mínimo da mensagem (*Minimum Message Length* – MML) criada por Wallace e Boulton (apud KORB; NICHOLSON, 2004).

A idéia básica por trás do MDL é que o melhor modelo que representa uma base de dados é aquele no qual a **soma do detalhamento da codificação do modelo** e a **soma do detalhamento da codificação dos dados dado o modelo** são minimizados. Dito de outra forma, é a escolha do compromisso (inglês: *tradeoff*), ótimo entre a simplicidade do modelo e o ajustamento aos dados que minimizam o detalhamento de uma descrição conjunta do modelo e dos dados. É importante destacar que as duas medidas devem ser representadas em *bits*.

Por exemplo, se é permitido ao modelo que ele cresça de forma arbitrariamente complexa, é certo que o seu poder de representação dos dados será grande.

Para mensurar a soma do detalhamento do modelo, foi sugerido por Lam e Bacchus (1994) uma medida de eficiência da rede representada levando em consideração a quantidade de nós N , a quantidade de pais k_i do i -ésimo nó, o tamanho da palavra d em *bits* para armazenar um valor numérico no computador e s_i o número de estados do i -ésimo nó. Essa medida é mostrada a seguir para cada instanciação:

$$I_{LB}(h) = \sum_{i=1}^N \left[k_i \log_2 N + d(s_i - 1) \prod_{j \in \mathcal{F}_i} s_j \right] \quad (2.8)$$

Para a codificação da rede é necessário instanciar cada pai do nó i e este, por sua vez, irá ocupar $\log N$ *bits* por cada pai. Para cada instanciação dessa é necessário especificar uma distribuição de probabilidade sob os filhos (CPT) com $s_i - 1$ parâmetros e com

$d(s_i - 1)$ bits para cada instanciação. Com isto, tem-se uma idéia de quão custoso é representar a rede computacionalmente. A Equação anterior nada mais é do que uma representação em bits da Equação (2.5).

A Equação (2.9) fornece uma maneira computacionalmente implementável de calcular a soma do detalhamento da codificação dos dados dado o modelo. A idéia básica por trás desta medida de informação é a de fornecer o ganho que se tem pela adição de nós ao conjunto dos pais $\mathcal{F}(i)$ do filho i .

$$I_{LB}(e|h) = K \left[\sum_{i=1}^N H(X_i) - \sum_{i=1}^N H(X_i, \mathcal{F}_i) \right] \quad (2.9)$$

Nesta equação, K representa a quantidade de observações conjuntas nos dados, ou dito de outra forma, a quantidade de combinações possíveis para os valores do filho com o conjunto de seus pais, $H(X_i)$ a entropia da variável X_i e $H(X_i, \mathcal{F}_i)$ a informação mútua entre a variável X_i e o conjunto de seus pais.

A entropia do nó X é definida pela Equação (2.10). Para essa definição, é assumido que os valores x da variável X possuem uma probabilidade de ocorrência $P(X = x)$; e, ainda, segundo uma medida largamente adotada na literatura, a qual foi apresentada no algoritmo de Huffman, que cada valor x será representado por um código de tamanho aproximadamente igual a $-\log_2 P(X = x)$. Assim, com K observações, sendo K de valor muito alto, espera-se obter $K \cdot P(X = x)$ ocorrências para o evento $X = x$. Desse modo, a base de dados possuirá uma codificação de tamanho aproximado igual a $-K \sum_x [P(X = x) \log_2 P(X = x)]$.

$$H(X) = - \sum_x P(X = x) \log_2 P(X = x) \quad (2.10)$$

Chow e Liu (1968) mostram que quando se define a informação mútua entre X_i e \mathcal{F}_i segundo a Equação (2.11) e, além disso, quando todos os possíveis valores de X_i e \mathcal{F}_i estão sendo somados nesta mesma equação, está se atribuindo o peso $H(X_i, \mathcal{F}_i)$ a todo arco entre X_i e \mathcal{F}_i . Neste caso generalizado, todos os valores estão sendo somados, pois trata-se do conjunto dos nós e seus respectivos pais.

$$H(X_i, \mathcal{F}_{X_i}) = H(X_i) - H(X_i | \mathcal{F}_{X_i}) = \sum_{x_i, F_i} P(x_i, F_i) \log_2 \frac{P(x_i, F_i)}{P(x_i)P(F_i)} \quad (2.11)$$

Como dito no início da seção, o MDL busca a escolha ótima de um compromisso (*tradeoff*) entre a simplicidade da rede e a capacidade da mesma representar os dados. Com isso, utiliza-se o indicador aditivo da Equação (2.12) para representar a função de pontuação da rede candidata que irá balizar o algoritmo em busca da solução ótima.

$$I_{LB}(h, e) = I_{LB}(h) + I_{LB}(e|h) \quad (2.12)$$

Na seção a seguir, sempre que for citada a expressão função de avaliação, *score* ou pontuação, estar-se referindo ao valor fornecido pela Equação (2.12) no cálculo do que se quer obter.

2.7.2 Um Algoritmo de Aprendizagem – *Sum Of Products* (SOP)

O SOP é um algoritmo que se enquadra na família de algoritmos de busca e pontuação, mas faz uso de alguns testes que são notadamente utilizados na classe dos algoritmos de independência condicional. Ele foi desenvolvido por Jouffe (2002) sob a óptica de que as *v*-estruturas são os principais elementos perturbadores de pesquisa no espaço das redes bayesianas, buscando assim evitar certos ótimos locais devido a estruturas equivalente.

O SOP é entendido como uma algoritmo de pesquisa de pais ótimos de um nó, efetuando assim uma busca gulosa para analisar as conseqüências de decisão de orientação dos arcos a cada passo da busca.

Esse algoritmo realiza uma pesquisa de pais do nó alvo t como se houvesse uma competição entre os nós candidatos através de dois operadores: a adição e a supressão de arcos para o nó alvo $\rightarrow t$.

No início, considera-se a rede totalmente desconectada e todos os nós são avaliados como possíveis pais de t . A cada passo, avalia-se a operação de adição e supressão de arcos via a função de avaliação que, como resultado, tem-se a adição dos pais ao nó alvo que obtiveram maior variação da pontuação no seu conjunto. No caso de um acréscimo de arco na fase de testes de supressão e adição, o vencedor w submete a sugestão de mudança para um analisador que irá classificá-lo como um integrante do conjunto de pais do no alvo (\mathcal{F}_t) ou do conjunto de filhos (\mathcal{S}_t) ou do conjunto dos nós indiretos (\mathcal{I}_t) ou do conjunto em que não se sabe se o nó é pai ou filho do no alvo (\mathcal{X}_t). A seguir são descritos cada um desses conjuntos.

Conjunto dos Nós Pais ou Filhos (\mathcal{X})

Este conjunto é formado pelos nós em que não se tem informação suficiente para decidir se um nó é pai ou filho do nó alvo. Como descrito anteriormente, neste caso, a ausência total de informação gera uma orientação aleatória do arco.

Conjunto dos Nós Indiretos (\mathcal{I})

Se um nó não é conectado diretamente através de um arco com a variável alvo, há a possibilidade desse nó ter alguma dependência marginal com o alvo, o que justifica a conexão do nó com alguma outra variável que se liga ao nó alvo direta ou indiretamente. Fazendo analogia com o fluxo de informação, os nós pertencentes ao conjunto anterior \mathcal{X} não interrompem nenhuma fonte de informação proveniente dos pais ou dos filhos, o que geraria uma rede com os nós totalmente interligados, direta ou indiretamente. Para realizar a interrupção do fluxo de informação e possibilitar a construção de redes não totalmente conectadas é utilizado um procedimento de detecção de nós indiretos que tentam interromper o fluxo de informação do nó vencedor w para o nó $x \in \mathcal{X} \cup \mathcal{S}$. A ruptura do fluxo informacional é testada simplesmente analisando a pontuação obtida das subredes da Figura 2.7. Se o *score* da Figura 2.7(b) não é tão bom quanto o da Figura 2.7(a) o fluxo de informação que sai de w é interrompido por x e $\mathcal{I} = \mathcal{I} \cup \{w\}$.

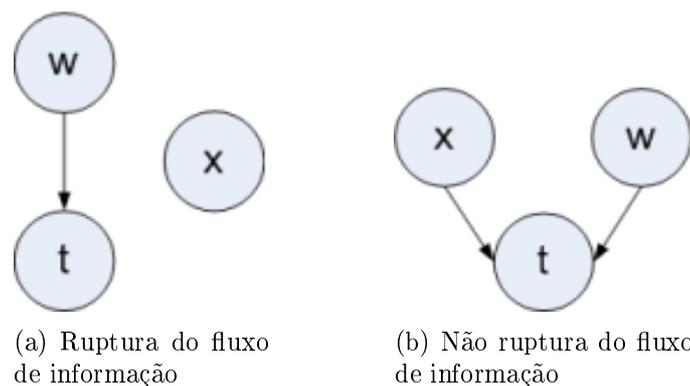


Figura 2.7: Teste de ruptura do fluxo de informação

Conjunto dos Nós Pais (\mathcal{F}) e Filhos (\mathcal{S})

Para a formação do conjunto dos pais de um nó t , parte-se do pressuposto que a v-estrutura deve ser previamente detectada. Em caso de não se ter detectado nenhum pai

($\mathcal{F} = \phi$), realiza-se um teste para verificar se o nó vencedor w forma uma v-estrutura com algum nó $x \in \mathcal{X}$, comparando-se os *scores* obtidos pelas estruturas $w \rightarrow t \leftarrow x$ e $w \rightarrow t \rightarrow x$. Caso a primeira configuração apresente melhor *score*, deve-se ter $\mathcal{F} = \mathcal{F} \cup \{x, w\}$. A cada passo realizado no espaço de busca, outros nós do conjunto \mathcal{X} devem ser orientados através da comparação dos *scores* das duas estruturas possíveis: $x_i \rightarrow t$ e $x_i \leftarrow t$. Caso a primeira estrutura apresente melhor pontuação, a relação $x_i \rightarrow t$ é adicionada à rede e $\mathcal{F} = \mathcal{F} \cup \{x_i\}$. Caso contrário, nenhum arco é adicionado mas armazena-se a relação $\mathcal{S} = \mathcal{S} \cup \{x_i\}$. No instante em que o nó alvo for direcionado, os nós vencedores restantes são analisados da mesma forma.

Após definidos os conjuntos a serem utilizados na classificação dos nós, parte-se para a formação da estrutura da rede.

A Elaboração da Estrutura Topológica da Rede

O algoritmo SOP é utilizado para se ter uma visão global da estrutura da rede bayesiana e se evitar os ótimos locais. O processo de estruturação da topologia da rede será dividida em três etapas para melhor entendimento do procedimento:

1. Aplica-se o algoritmo SOP para todos os nós t_i sem a realização de testes de aciclicidade. Como já visto, o resultado dessa fase é a caracterização dos nós em diferentes conjuntos (\mathcal{S}_{t_i} , \mathcal{F}_{t_i} , \mathcal{X}_{t_i} , e \mathcal{I}_{t_i}).
2. Define-se o ganho potencial do nó como a diferença entre a pontuação obtida do mesmo sem pais e com o conjunto de pais sugeridos pelo SOP. É importante destacar que o conjunto sugerido pelo SOP não apresenta compromisso com a aciclicidade, sendo assim, este ganho potencial deve representar o ganho máximo obtido pelo nó e os nós que apresentam maiores ganhos potenciais terão assim prioridade na escolha dos pais potenciais.
3. Seguindo a ordem definida na etapa anterior, do nó t_i de maior ganho potencial para o de menor, deve-se conectar os pais definidos em \mathcal{F}_{t_i} ao nó t_i , atentando agora para a necessidade de não haver ciclos.

O resultado destas etapas é um gráfico formado por v-estruturas, o qual representa a rede bayesiana procurada com maior valor da função de avaliação dada a restrição de aciclicidade.

2.8 As Taxas de Falha e Reparo

De uma forma intuitiva as taxas de falha $\lambda(t)$ e reparo $\mu(t)$ podem ser definidas como a intensidade com que um sistema passe de um estado para outro. Elas serão chamadas de taxas de falha quando o sistema muda de estado devido à falha de algum componente ou conjunto de componentes e de taxas de reparo quando no sistema há uma mudança de estado devido ao reparo de algum equipamento ou conjunto de equipamentos que não estavam em condições operacionais.

Utilizando a notação matemática tem-se $\lambda(t)\Delta t$ como a probabilidade de que o sistema irá falhar em algum tempo $t < t' + \Delta t$

$$\lambda(t)\Delta t = P\{t < t' + \Delta t | t > t'\}$$

e, utilizando a idéia de probabilidade condicional e conjunta apresentada nas seções 2.3.2 e 2.3.2, tem-se:

$$P\{t < t' + \Delta t | t > t'\} = \frac{P\{(t > t') \cap t < t' + \Delta t\}}{P\{t > t'\}}$$

Escrevendo o numerador de uma outra forma tem-se:

$$P\{(t > t') \cap t < t' + \Delta t\} \equiv P\{t' < t < t' + \Delta t\} = f(t)\Delta t$$

e como $P(t > t')$ é definido como a confiabilidade do sistema $R(t)$, obtém-se:

$$\lambda(t) = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - \int_0^t f(t)dt} \quad (2.13)$$

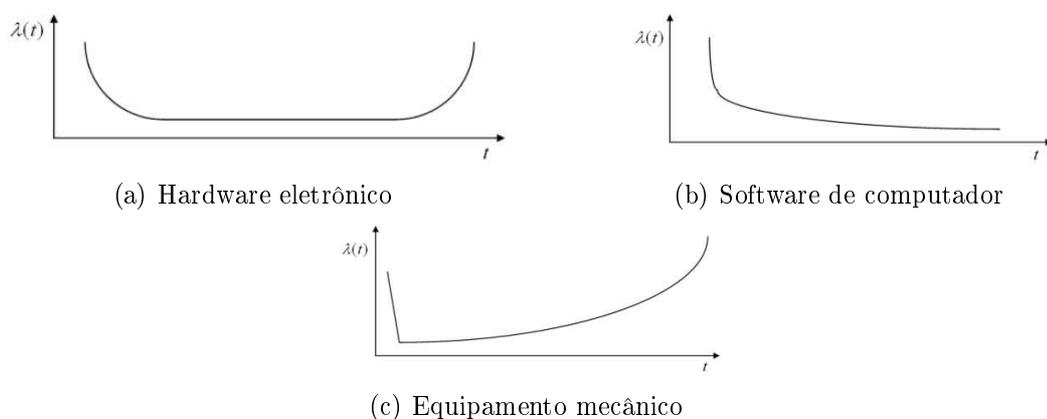
onde $f(t)$ é a densidade de probabilidade de falha $f(t) = \frac{-dR(t)}{dt}$.

A mesmo desenvolvimento pode ser realizado para se chegar a expressão da taxa de reparo, só que em vez de usar o tempo até que a falha ocorra, deve-se usar o tempo até que o reparo termine.

A seguir serão descritos algumas distribuições de probabilidade que podem ser utilizadas para modelar os tempos de falha e reparo assumindo certas taxas de falha e reparo.

2.9 Parametrização dos Tempos de Falha e Reparo

Tradicionalmente, antes de se realizar um ajustamento dos tempos de falha a uma distribuição de probabilidade de um equipamento, é comum analisar em que período encontra-se a taxa de falha do mesmo. Lewis (1987) evidencia alguns tipos de comportamento de taxas de falha a depender do tipo do sistema analisado (mecânico, hardware eletrônico e softwares de computador). Mas, o mais importante disso é que em todos os sistemas analisados, os mesmos apresentam um comportamento de taxa de falha decrescente, constante e, em quase todos, crescente em função do tempo, ver Figura 2.8.



Fonte: Lewis (1987, p.27).

Figura 2.8: Curvas de comportamento das taxas de falha por tipo de equipamento

Estas curvas são chamadas de curva da banheira, pois como visto na Figura 2.8(a), a mesma possui uma forma de banheira. É a partir dessas curvas que se decide qual distribuição de probabilidade utilizar para parametrizar as falhas dos componentes.

Para as taxas de reparo não faz sentido falar em curva da banheira, uma vez que o tempo de reparo não é influenciado por características intrínsecas do equipamento. Bukowski (2006) sugere que o tempo de reparo de um equipamento é formado pela soma de muitas partes: tempo para a equipe de manutenção chegar ao local do equipamento falho, tempo para identificar a falha, tempo para adquirir os equipamentos sobressalentes e, por fim, o tempo para o reparo em si. Assim, basta apenas averiguar a que tipo de distribuição o tempo de reparo ajusta-se para então parametrizar o processo de Markov.

2.9.1 As Distribuições de Probabilidades

Seguindo o apresentado na forma da curva da banheira, pode-se modelar as falhas dos componentes assumindo que os mesmos estão na área central da curva, em que o comportamento da taxa de falha é constante ou aleatória, independente do tempo de uso do equipamento, ou nas extremidades, onde eles estão em processo de deterioração (taxa de falha crescente) ou falhas prematuras (taxa de falha decrescente).

Distribuição Weibull

A distribuição Weibull é largamente utilizada para representar tempos de falha e de reparo de equipamentos pois ela é bastante flexível, permitindo modelar todas as fases da curva da banheira. Uma variável aleatória contínua t segue uma distribuição de Weibull quando a sua PDF é dado por:

$$f(t) = \frac{\beta t^{\beta-1}}{\alpha^\beta} \exp \left[- \left(\frac{t}{\alpha} \right)^\beta \right] \quad (2.14)$$

A Figura 2.9 mostra a PDF de uma distribuição de Weibull com o parâmetro α constante igual a 3 e diferentes valores de β . A Figura 2.9 mostra que o parâmetro β é que determina o formato da distribuição e, de acordo com o seu valor, a distribuição pode representar uma grande variedade de formatos ou comportamento do equipamento, no caso de interesse. Já o parâmetro α é referido como o parâmetro de escala, pois o mesmo controla a escala da distribuição.

Distribuição Exponencial

Uma distribuição comumente utilizada na modelagem da fase intermediária da curva da banheira é a distribuição exponencial, em que as taxas do modelo é constante e a falta de memória⁵ está presente.

Para definir uma distribuição exponencial, é necessário apenas estimar um parâmetro, que no contexto da confiabilidade seria a taxa λ constante. Assim, a distribuição expo-

⁵Este é o processo em que se assume que a ocorrência de um evento só depende do estado atual, desconsiderando os tempos passados, ou seja, $P\{X \geq x' + \Delta x | X \geq x'\} = P\{X \geq \Delta x\}$. Assim, equipamentos modelados por essa distribuição não apresentarão deterioração ao longo do tempo de operação.

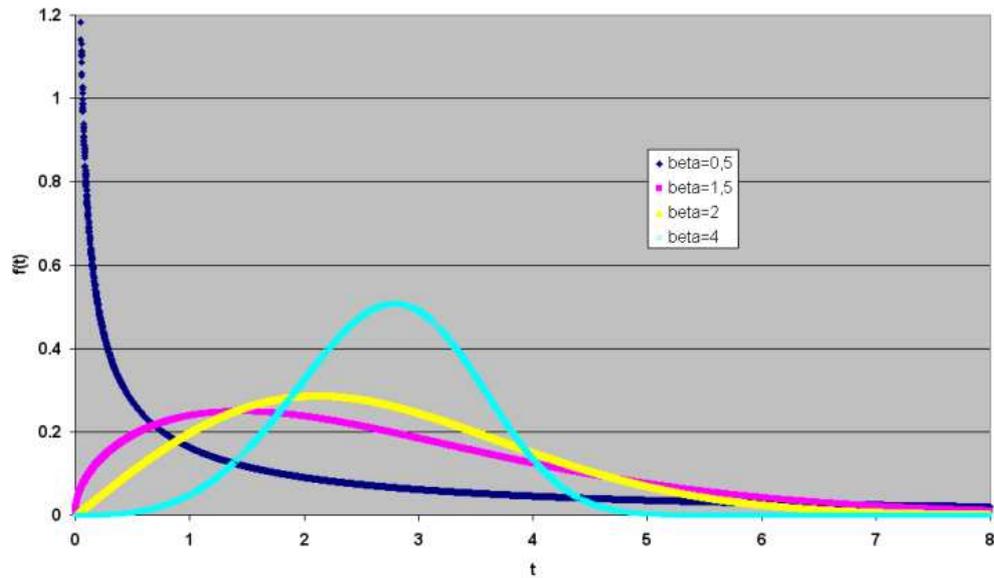


Figura 2.9: Exemplo de densidades de probabilidade Weibull

nencial é definida como:

$$f(t) = \lambda e^{-\lambda t} \quad (2.15)$$

com média igual a $1/\lambda$ e variância igual a $1/\lambda^2$.

A Figura 2.10 evidencia exemplos de distribuição exponencial para três valores distintos de seus parâmetros.

É importante observar que a distribuição exponencial é um caso particular da distribuição Weibull, em que o parâmetro β é igual a 1 e o parâmetro α é igual a $1/\lambda$.

Distribuição Normal

A mais conhecida distribuição de probabilidade contínua é a distribuição Normal, muitas vezes referida como distribuição gaussiana. Uma distribuição normal é definida por dois parâmetros: a média μ e a variância σ^2 , como mostra a equação a seguir:

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]; \quad -\infty < t < \infty \quad (2.16)$$

A distribuição normal tem a característica de ser simetricamente distribuída em torno da sua média, como mostra a Figura 2.11 para média assumida igual a 0.

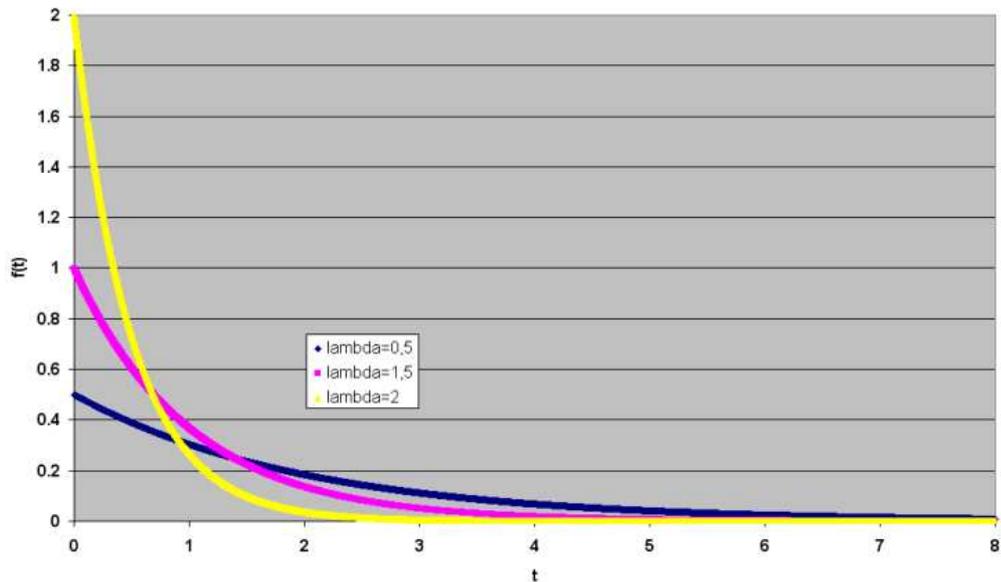


Figura 2.10: Exemplo de densidades de probabilidade exponencial

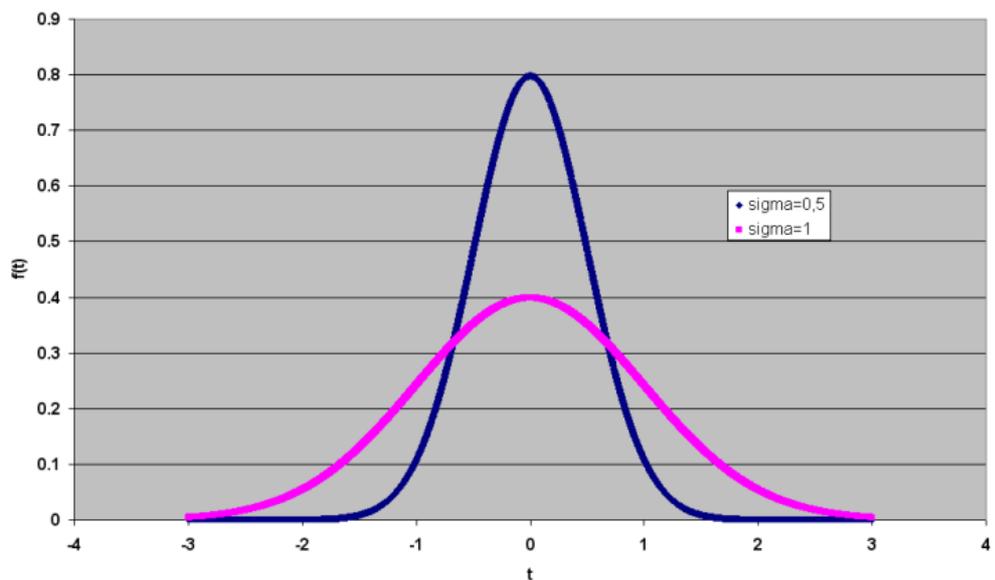


Figura 2.11: Exemplo de densidades de probabilidade Normal

Essa distribuição tem uso limitado na confiabilidade e análise de risco, pois só permite a modelagem de sistemas com taxas de ocorrência crescentes, sendo só então aplicadas na fase de deterioração dos sistemas. Além disso, ela permite tempos de falha ou reparo negativos, o que não é apropriado em análises de confiabilidade e disponibilidade.

Distribuição LogNormal

Uma variável aleatória T segue uma distribuição LogNormal se o logaritmo natural de T segue uma distribuição Normal. Assim, uma distribuição LogNormal está apenas distribuída para valores positivos de T , o que a torna mais apropriada em análises de confiabilidade e disponibilidade. A sua PDF é representada por:

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{(\log t - \mu)^2}{2\sigma^2} \right]; \quad t > 0 \quad (2.17)$$

É importante destacar que μ e σ são, respectivamente, a média e o desvio padrão do logaritmo natural do tempo ($\ln T$), e não a média e desvio padrão do tempo T . A Figura 2.12 mostra distribuições LogNormais para três diferentes valores de σ e média constante igual a 0,8.

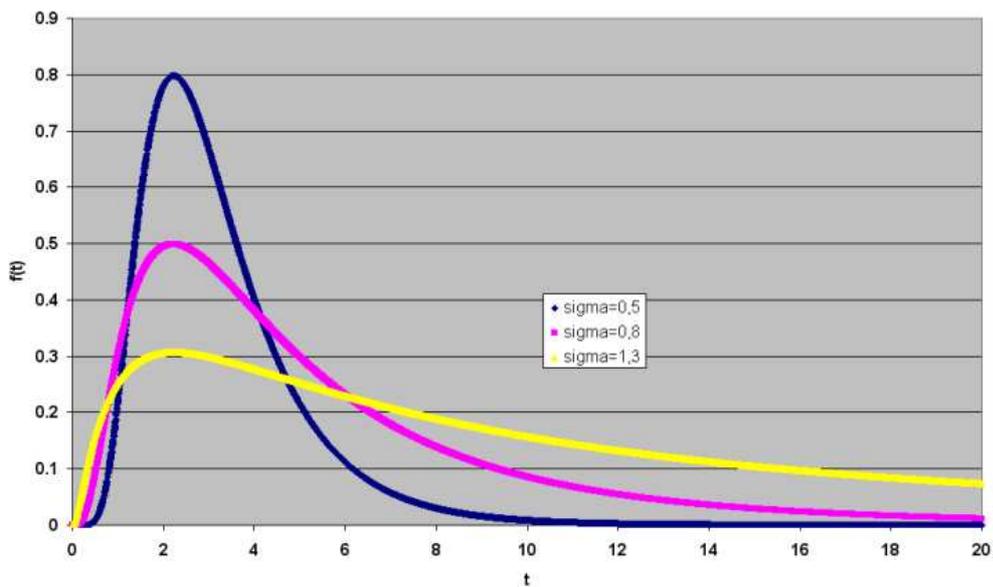


Figura 2.12: Exemplo de densidades de probabilidade LogNormal

2.10 Processos Markovianos

Um processo estocástico é definido como uma variável aleatória que para cada instante de tempo t há uma distribuição de probabilidade indexada. Um exemplo pode ser definido como uma variável aleatória $X(t)$ para $t \in T$, que representa o número de clientes na fila de um banco num determinado instante t . Nesse exemplo, a variável aleatória que representa

os estados do sistema é discreta e o tempo contínuo, o que será assumido a partir de agora.

Os processos de Markov são um dos processos estocásticos mais utilizados na modelagem de sistemas probabilísticos dinâmicos e apresentam muita robustez na análise de disponibilidade.

Um processo estocástico é dito ser markoviano se possui a seguinte propriedade:

$$P[X(t) = j | X(v) = i, X(u) = k, \dots, X(0) = s] = P[X(t) = j | X(v) = i] \quad (2.18)$$

onde $0 \leq u \leq v \leq t$. Para maiores detalhes ver Droguett (2002).

Pode-se entender com a equação acima que se o presente é conhecido, o passado não terá nenhuma influência no comportamento futuro do sistema.

No processo de Markov, o sistema é graficamente representado por uma cadeia (cadeia de Markov) constituída de arcos e nós, como nas redes bayesianas, só que a condição de aciclicidade não é mais necessária. Aqui as variáveis representam o estado do sistema e os arcos as transições entre os estados. Na aplicação em confiabilidade essas transições frequentemente são representadas por falhas ou reparos nos equipamentos que caracterizam os estados.

Neste último ponto é que se costuma realizar a classificação dos processos markovianos em homogêneo e não-homogêneo. No primeiro, as transições são representadas por taxas de valores constantes ao longo do tempo (λ, μ , etc.). Já no segundo caso, essas são representadas como funções no tempo ($\lambda(t), \mu(t)$, etc.).

Os tempos que regem as transições entre estados no modelo homogêneo são distribuídos segundo uma distribuição exponencial e, como consequência disso, as probabilidades de transição não dependem do tempo já percorrido pelo sistema, mas sim apenas do tempo atual para a transição⁶.

Acabou-se de mostrar que as probabilidades de transição entre os estados da cadeia são representadas pelas probabilidades condicionais do tipo $p_{ij}(v, t) = P[X(t) = j | X(v) = i]$ (probabilidade de ir para o estado j em t dado que em v o sistema estava em i). Essas por sua vez possuem a seguinte propriedade: $\sum_{j=0}^r P_{ij}(t) = 1$. A equação de Chapman-Kolmogorov diz que a probabilidade de sair de um estado i e ir até um estado j é igual ao somatório das probabilidades de transição de sair de i e ir até um estado j passando

⁶Essas probabilidades de transição em que $P[X(t+v) = j | X(v) = i] = P_{ij}(v)$ são chamadas de probabilidades de transição estacionárias

por todos os estados intermediários possíveis k :

$$P_{ij}(v, t) = \sum_{k=0}^r P_{ik}(v, u) \cdot P_{kj}(u, t) \quad \text{para } t, v > 0 \quad (2.19)$$

Definindo-se as taxas de transição do estado i ao estado j como:

$$q_{ij}(t) = \lim_{\Delta t \rightarrow 0} \frac{P[X(t + \Delta t) = j | X(t) = i]}{\Delta t} \quad (2.20)$$

por meio da equação de Chapman–Kolmogorov, pode-se chegar as equações das probabilidades de estado (2.21). Para as demonstrações ver Howard (1971).

$$\frac{d}{dt} P_j(t) = \frac{d}{dt} P[X(t) = j] = \sum_i q_{ij}(t) \cdot P[X(t) = i] \quad (2.21)$$

2.11 Análise de Disponibilidade com Modelos Markovianos

A análise de disponibilidade pode ser realizada estimando-se a própria disponibilidade do sistema bem como alguns indicadores que dizem respeito aos estados do sistema.

Se o sistema é composto por dois subconjuntos: D (conjunto dos estados disponíveis) e I (conjunto dos estados indisponíveis) a disponibilidade instantânea do sistema será dada pela probabilidade do sistema está no subconjunto D no tempo t . Assim:

$$A(t) = \sum_{k \in D} P_k(t) \quad (2.22)$$

Além disso, muitas vezes quer-se achar uma disponibilidade independente do estado inicial do sistema $X(t_0)$, para isso faz-se $t \rightarrow \infty$ e chama o $\lim_{t \rightarrow \infty} P_j(t)$ de probabilidade estacionária do estado j . Assim:

$$A_\infty = \lim_{t \rightarrow \infty} \sum_{k \in D} P_k(t) \quad (2.23)$$

Outro indicador a ser utilizado na análise de disponibilidade é o número de visitas que o sistema faz aos estados. É importante realizar tal observação para, por exemplo, identificar quais os estados que estão sendo mais visitados, sendo possível com isso dizer

quais os equipamentos que caracterizam os estados falhos demandam mais reparos. Uma decisão possível de ser tomada com tal análise é a de criar redundância dos equipamentos ou até mesmo trocar a tecnologia do sistema. Matematicamente, o número esperados de visitas a um dado estado indisponível num intervalo de tempo t é representado como o número de transições de um estado disponível a esse estado indisponível:

$$f_j = \sum_{i \in I} \int_0^t P_i(\lambda) \cdot \mu_{ij} \cdot d\tau \quad (2.24)$$

Além dos indicadores até então definidos, pode-se ainda querer saber quanto tempo em média o sistema passa no estado indisponível. Ainda mais, pode-se desejar saber quanto tempo em média demora-se para realizar o reparo de um dado equipamento, avaliando assim as equipes e a atividade de manutenção. Para definir tal indicador é preciso antes assumir que a taxa de permanência num estado é igual à taxa total de saída desse estado:

$$q_{jj}(t) = \sum_{\substack{i=0 \\ i \neq j}}^n q_{ji}(t) \quad (2.25)$$

Assim, matematicamente, a duração média de permanência do sistema no estado j é dado por:

$$\bar{T}_j = \frac{1}{\frac{1}{t} \int_0^t q_{jj}(\tau) d\tau} \quad (2.26)$$

Com esses indicadores, é possível realizar uma análise mais profunda do nível de disponibilidade de um sistema.

No capítulo a seguir, é proposta uma metodologia de como realizar uma análise de disponibilidade de um sistema complexo utilizando os conceitos até aqui apresentados.

3 DESENVOLVIMENTO DA METODOLOGIA PROPOSTA

3.1 Introdução

O estudo desenvolvido trata-se de uma metodologia para avaliar sistemas complexos que possuem grandes bases de dados de história de funcionamento dos mesmos. É imprescindível para a aplicação da metodologia que fiquem registrados na base os tempos entre falhas e de reparo dos equipamentos e/ou modos de falha destes sistemas, bem como variáveis monitoradas que de algum modo venham a influenciar os eventos de falha e reparo.

A disponibilidade do sistema neste trabalho é modelada via um processo de Markov não-homogêneo. Para a quantificação de tal modelo, é necessária a estimação da distribuição não-paramétrica dos tempos entre falhas e de reparo dos equipamentos ou de seus modos de falha envolvidos, ou seja, a estimação das distribuições dos tempos de transição entre os estados da cadeia markoviana.

Parte-se do princípio então que os tempos de transição da cadeia de Markov são influenciados por outros atributos como qualidade dos componentes substituídos, qualidade do reparo, etc. Esses atributos serão discutidos detalhadamente no próximo capítulo no exemplo de aplicação.

A metodologia proposta procura, então, identificar e caracterizar as relações de cada efeito entre os tempos de falha e de reparo de cada equipamento e/ou modos de falha e atributos considerados relevantes que influenciem tais tempos. Essas relações causais são caracterizadas por meio de redes bayesianas em que suas topologias e parâmetros são estimados a partir de base de dados via algoritmos de aprendizagem (ver capítulo anterior).

Uma vez obtidas as topologias dessas redes e as respectivas distribuições de probabilidade condicionais *a priori*, inicia-se a resolução da cadeia markoviana. Nesta cadeia, as BBNs são acopladas e quantificadas em tempo de simulação da cadeia, a fim de se obterem distribuições atualizadas dos tempos de transição de Markov.

Ressalta-se ainda outros aspectos importantes da metodologia proposta:

- Capacidade de modelar processos em fases de deterioração ou melhoria (diferentes fases da curva da banheira);
- Flexibilidade ao considerar reparos imperfeitos como também renovações em pontos específicos durante o tempo de missão considerado. Por exemplo, pode-se especificar previamente que se a probabilidade de falha de um componente atingir um valor maior do que x , um reparo perfeito ocorrerá e o componente estará completamente renovado, retornando à operação “tão bom quanto novo”.
- Possibilidade de atualizar em tempo de simulação as distribuições dos tempos de falha e reparo das redes bayesianas à medida que novas informações sobre o sistema são disponibilizadas. Por exemplo, se informações quanto à qualidade dos equipamentos substituídos num reparo forem fornecidas, poderá haver uma variação na distribuição de probabilidade do tempo entre falhas do equipamento ou de ocorrência do modo de falha, uma vez que o impacto dessa evidência poderá ser propagada através da rede bayesiana até atingir o nó de interesse, que é o tempo entre falhas.

Os passos metodológicos de como tornar essa modelagem factível no “mundo real” são listados a seguir e descritos nas seções seguintes.

1. Identificação dos equipamentos e variáveis relevantes que influenciam os tempos entre falhas e de reparo do sistema;
2. Realização de uma limpeza e integração dos dados;
3. Estruturação das relações causais entre as variáveis do sistema e os tempos entre falhas e de reparo;
4. Estimação dos parâmetros das redes bayesianas estruturadas;
5. Estruturação da cadeia markoviana do sistema;
6. Análise de disponibilidade com a cadeia markoviana;
7. Atualização do estudo de disponibilidade via evidências nas variáveis das redes bayesianas dos equipamentos do sistema.

A Figura 3.1 ilustra esquematicamente os passos citados acima.

Para melhor entender a forma de interação entre as ferramentas utilizadas na metodologia proposta, é desenhada, na Figura 3.2, a estrutura de relacionamento entre as ferramentas utilizadas. Nessa figura os dados servem como um ponto de partida para a estruturação topológica e parametrização das probabilidades condicionais das redes bayesianas. Com essas parametrizações, as redes bayesianas geram tempos de transição baseados nas distribuições de probabilidade dos nós “tempos entre falhas” e “tempos de reparo” e, assim, têm-se definidas as transições da cadeia de Markov via simulação discreta de eventos. Além desse sequencialmente, é possível a qualquer momento realizar atualizações nos parâmetros das redes bayesianas ao passo que novas variações nas variáveis monitoradas forem sendo observadas, com isso obtendo uma análise continuada da disponibilidade do sistema.

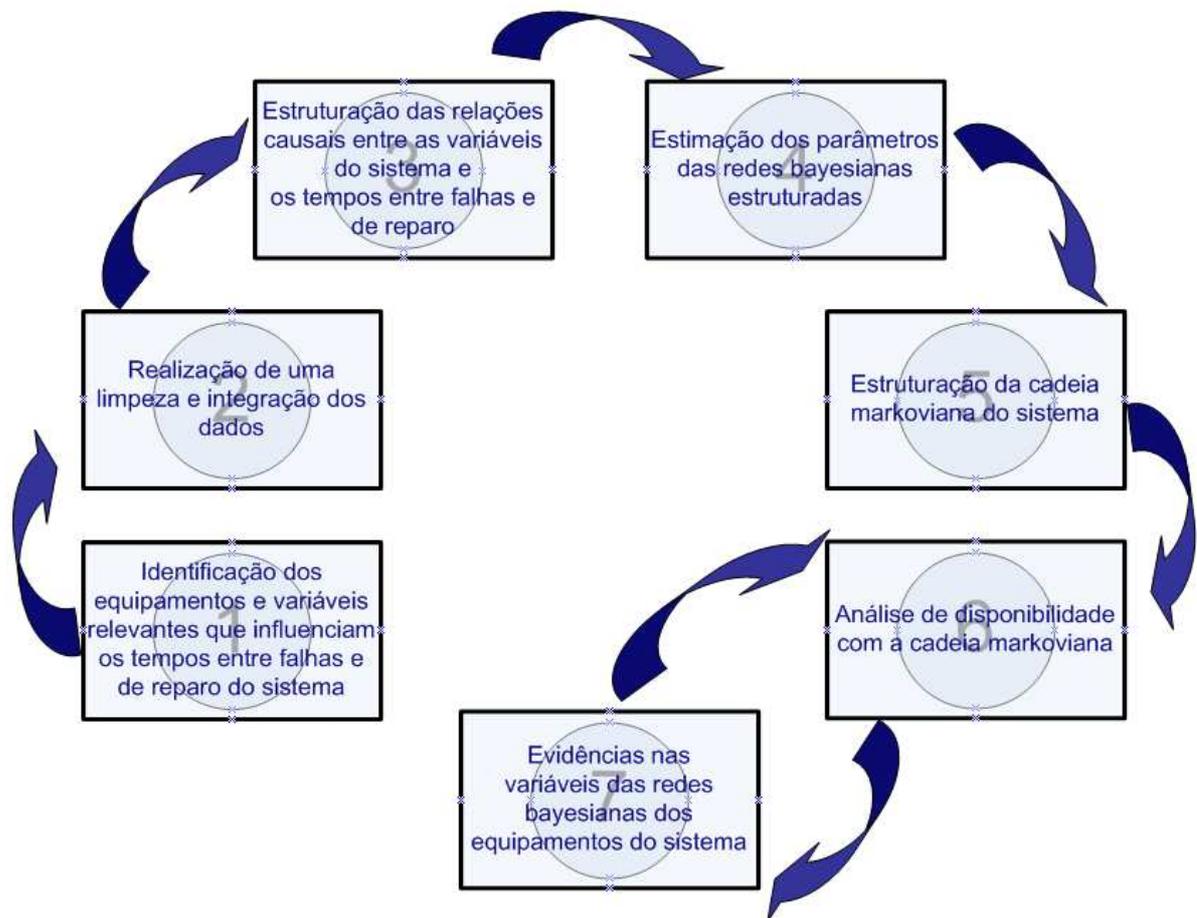


Figura 3.1: Passos metodológicos da abordagem proposta

Como mostrado na seção 2.2, Hrudchka (2003) apresenta passos metodológicos genéricos para a extração de conhecimento de base de dados de um sistema qualquer. A metodologia proposta neste trabalho, considerando essa abordagem genérica, faz uso de

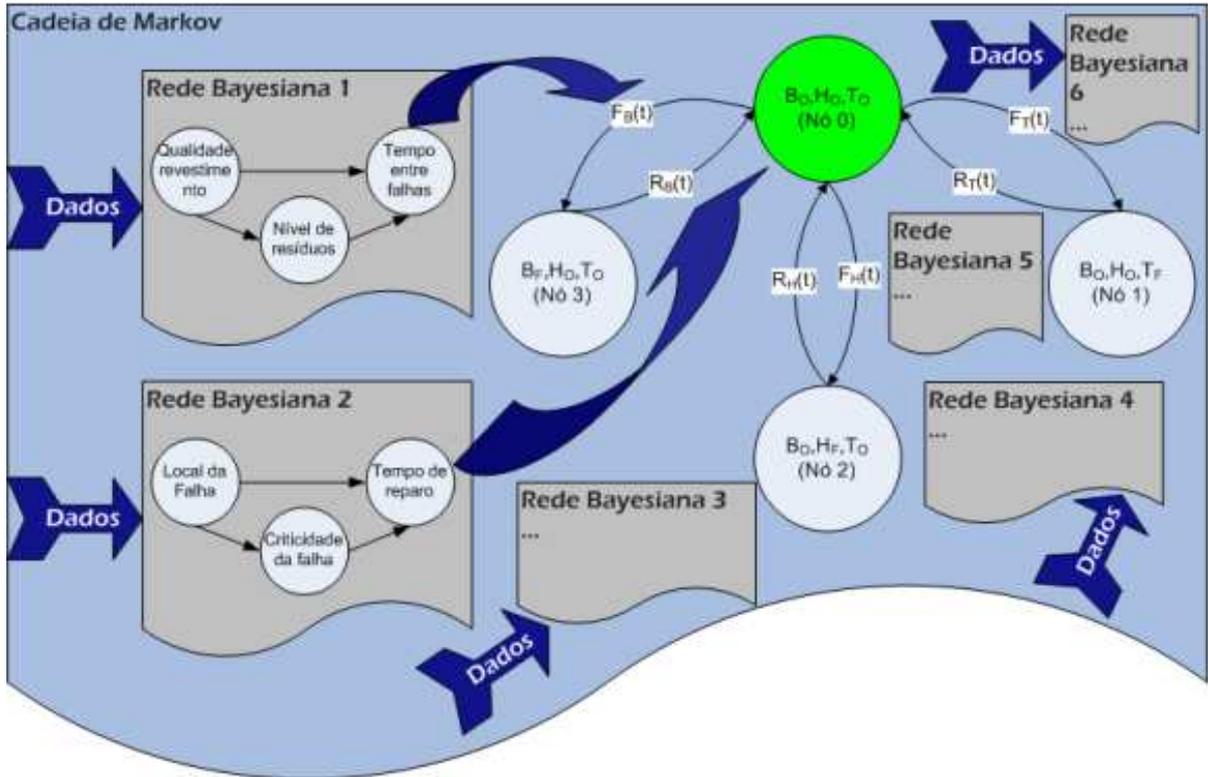


Figura 3.2: Esquema da metodologia proposta

ferramentas de cunho probabilístico para servir de detalhamento da fase da mineração de dados, etapa na qual a “inteligência do processo” é desenvolvida.

Nesse detalhamento, busca-se, de uma maneira bastante precisa, a análise de disponibilidade com uma ferramenta comumente utilizada nessa área do conhecimento, a cadeia de Markov, mas, além disso, incorpora à cadeia a possibilidade de atualização dos parâmetros de transição, uma vez que considera mudanças nas variáveis monitoradas do sistema utilizando as redes bayesianas. Adicionalmente, com a não-homogeneidade utilizada num dos algoritmos desenvolvidos a ser mostrado, dá uma característica nitidamente robusta à análise, já que incorpora à análise o tempo de operação do sistema no passado. A seguir são detalhados cada um dos passos anteriormente definidos para a análise de disponibilidade de sistemas complexos que são monitorados e registros são realizados em base de dados.

3.2 Identificação dos Equipamentos e Variáveis Relevantes que Influenciam os Tempos entre Falhas e de Reparo do Sistema

Nesta etapa são definidos os equipamentos que apresentam relevante impacto na estimação da disponibilidade do sistema, bem como as variáveis monitoradas relevantes.

A opinião de um especialista deve ser utilizada para que apenas as variáveis relevantes sejam adicionadas as redes bayesianas. Pode-se utilizar também técnicas estatísticas de correlação entre variáveis para identificar aquelas variáveis de elevada influência nos tempos de reparo e entre falhas de cada equipamento ou componente do sistema.

Aqui, faz-se uso de uma base de dados existente que freqüentemente registra o tempo cronológico da ocorrência da falha (data, hora, etc.) e do reparo e os valores das variáveis monitoradas. Calculando-se a diferença entre os tempos de ocorrência de falha, tem-se o tempo entre falhas (*inglês: time between failures*) de um determinado equipamento do sistema. O mesmo deve ser feito para o caso dos reparos (a diferença entre o tempo cronológico do término do reparo e o início do mesmo), deste modo tem-se o tempo até o reparo (*inglês: time to repair*).

É importante destacar que em muitas situações práticas, a forma adotada nas bases de dados não possibilita a determinação das taxas de falha e reparo dos equipamentos em si, pois não são armazenados os dados históricos de falha e reparo dos equipamentos, mas sim o histórico do sistema. Desta forma, a exemplo do que ocorre com o ser humano, em que o seu comportamento é influenciado pelo meio em que se está inserido, é assumido que os equipamentos apresentam um comportamento padrão a depender do sistema em que estão inseridos, o que na literatura tem sido tratado como um problema de estimação dos parâmetros do equipamento no *socket* (ASCHER; FEINGOLD, 1984).

3.3 Realização de uma Limpeza e Integração dos Dados

Da mesma forma que foi explicitado na seção 2.2, esta etapa também é denominada como a etapa de pré-processamento.

É importante a realização de uma limpeza nos dados de modo a filtrar os ruídos que

fazem com que os dados não representem a amostra analisada. Limites de confiança podem ser estabelecidos para uma distribuição empírica das variáveis analisadas a fim de realizar a retirada de ruídos (HAHN; SHAPIRO, 1967).

A integração dos dados é realizada com o objetivo de preencher espaços vazios na base de dados¹. Alguns métodos são sugeridos, tais como: imputação bayesiana, valor médio ou modal, completação dinâmica, baseados na estrutura, entre outros (HRUDCHKA, 2003).

Essas imputações são realizadas devido à necessidade dos algoritmos de estruturação das rede bayesianas necessitarem de dados completos e/ou devido à existência de poucos dados para realizar a análise. Assim, de outra forma, poderia-se ocorrer o processo de limpeza.

É importante destacar que inverteu-se a ordem desses dois primeiros passos devido à possibilidade de se despender esforços em variáveis que poderiam não ser utilizadas na análise.

Com a prontidão da base de dados, segue-se para o passo da estruturação da rede bayesiana.

3.4 Estruturação das Relações Causais entre as Variáveis do Sistema e os Tempos entre Falhas e de Reparo

Nesta etapa, busca-se definir a topologia da rede bayesiana, deixando clara as relações causais entre as variáveis monitoradas e os tempos entre falhas e de reparo. Pode ocorrer que ao se aplicar o algoritmo de estruturação das redes² os tempos entre falhas e de reparo não sejam incluídos como nós da rede. O que se sugere neste caso é atribuir maior peso ao componente da soma do detalhamento da codificação dos dados dado o modelo da função de pontuação do MDL, a fim de que maior quantidade de arcos sejam adicionados e os tempos sejam interligados as variáveis. Com isso tem-se uma maior representação do modelo aos dados.

Com essa primeira estruturação, sugere-se a apresentação do modelo a um especialista para validá-lo, visto que a incerteza recai sobre as redes equivalentes markovianas³. Caso

¹Conhecido na literatura como o fenômeno da imputação.

²Para consulta de exemplo de algoritmo de estimação topológica de rede bayesiana, vide seção 2.7.2.

³Vide seção 2.4.5.

a opinião do especialista seja favorável a mudança de sentido de arco, deve-se buscar primeiro uma rede equivalente na classe de equivalências markovianas. Caso não seja possível, devido a uma restrição da rede, realizar a mudança de orientação forçadamente, mas sempre observando a restrição de aciclicidade.

Colocações e retiradas de arcos também devem ser admitidas como possíveis, uma vez que se está trabalhando com uma amostra de uma população e o viés pode estar inserido.

Após a estruturação da rede, segue-se para a etapa de estimação das probabilidades condicionais.

3.5 Estimação dos Parâmetros das Redes Bayesianas Estruturadas

Neste momento, são estimados os parâmetros das redes bayesianas que são as probabilidades condicionais. Como já dito anteriormente, esta estimação pode ser realizada via dados ou via opinião de especialistas, mas como o pressuposto é de que existam dados para formar a estrutura da rede, parte-se para a estimação dos parâmetros via dados.

A idéia básica é a de estimar as probabilidades condicionais de forma frequentista, realizando o processo de contagem para os valores da variável alvo condicionado aos possíveis valores do conjunto de seus pais. Desta forma, a probabilidade condicional de um valor de um nó filho é estimada dividindo a quantidade de casos para esse valor e para um determinado conjunto valores dos nós pais pela quantidade em que este conjunto de valores dos nós pais é verificado na base de dados.

Essas probabilidades condicionais serão armazenadas numa tabela de probabilidade condicional (TPC), e cada nó que possua pelo menos um pai terá a sua. Como exemplo de cálculo de uma TPC, tem-se a rede da Figura 3.3 e sua respectiva TPC na Tabela 3.1.

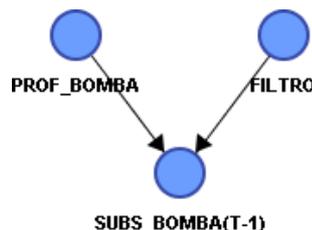


Figura 3.3: Exemplo de rede para cálculo da TPC

Tabela 3.1: Exemplo de TPC

PROF_BOMBA	FILTRO	SUBS_BOMBA(T-1)	
		N	S
≤ 430	C	0,07712	0,92288
≤ 430	F	0,0906	0,9094
≤ 430	N	0,41534	0,58466
≤ 430	S	0,55555	0,44445
≤ 546,5	C	0,15956	0,84044
≤ 546,5	F	0,09349	0,90651
≤ 546,5	N	0,30462	0,69538
≤ 546,5	S	0,45965	0,54035
≤ 743	C	0,09367	0,90633
≤ 743	F	0,09222	0,90778
≤ 743	N	0,12702	0,87298
≤ 743	S	0,14352	0,85648
> 743	C	0,08723	0,91277
> 743	F	0,06074	0,93926
> 743	N	0,09267	0,90733
> 743	S	0,10212	0,89788

Por exemplo, para se calcular o valor de $P(\text{SUBS_BOMBA}(T-1)=N \mid \text{PROF_BOMBA} \leq 430, \text{FILTRO} = C)^4$, deve-se realizar a contagem dos casos na base em que $\text{SUBS_BOMBA}(T-1)=N$ e $\text{PROF_BOMBA} \leq 430$ e $\text{FILTRO}=C$ e dividir este valor pela quantidade de ocorrências em que $\text{PROF_BOMBA} \leq 430$ e $\text{FILTRO}=C$ ocorrem na base de dados. Com isso tem-se a probabilidade condicional desejada, ou seja, $P = 0,07712$.

3.6 Estruturação da Cadeia Markoviana do Sistema

Identifica-se neste momento as configurações de funcionamento dos equipamentos, como em série ou em paralelo, com o intuito de estruturar a cadeia de Markov. São definidos os estados do sistema em indisponível ou disponível com base na operacionalidade ou falha do conjunto de seus equipamentos, bem como as possíveis transições entre esses estados. Para melhor entendimento deste procedimento um exemplo será exposto no capítulo seguinte.

Com a distribuição não-paramétrica obtida com as redes bayesianas para os tempos, é possível ajustá-los a distribuições paramétricas (exponencial, weibull, lognormal, normal, etc.) e realizar os cálculos das taxas de falha e reparo conforme equação 2.13, definindo

⁴Nomenclatura, vide tabela 4.2 na página 61.

assim o grau de intensidade com que ocorre as transições entre os estados.

Neste trabalho, ao invés da utilização de distribuições paramétricas de probabilidade para se estimar as taxas de transição, utiliza-se um algoritmo proposto, a ser mostrado na próxima seção, que faz uso das distribuições não-paramétricas de probabilidade dos TBF's e TTR's para realizar o cálculo das probabilidades transientes e estacionárias dos estados da cadeia de Markov. Assim, não se faz necessário ajustes dos TBF's e dos TTR's a uma distribuição paramétrica conhecida. Tudo isso será discutido na seção seguinte e, além disso, a forma de cálculo dos limites destas probabilidades será exposta.

A título de entendimento, até o presente momento, tem-se:

1. Os dados servindo de entrada (inglês: *input*) para um algoritmo estruturação topológica de redes bayesianas;
2. Os dados sendo utilizados para estimar as probabilidades condicionais das redes bayesianas;
3. As redes bayesianas, por meio das distribuições não-paramétricas de probabilidade totais dos TBF's e TTR's, sendo usadas para gerar tempos aleatórios de permanência nos estados da cadeia de Markov.

Como será visto mais adiante, ao passo que novas variações nas variáveis monitoradas forem sendo observadas, executar-se-á o passo 2 e 3 para se ter estimações em tempo real da disponibilidade do sistema em análise.

3.7 Análise de Disponibilidade com a Cadeia Markoviana

A fim de realizar a análise de disponibilidade de um sistema, propõe-se algoritmos para cálculo das probabilidades dos estados do sistemas e seus intervalos de confiança. A justificativa para desenvolvimento destes algoritmos está calcada na necessidade de se avaliar as probabilidades dos estados sem ter que assumir alguma distribuição de probabilidade para os TBF's e TTR's, fazendo com que a análise seja mais flexível.

Com o proposto, busca-se simular o sistema não mais utilizando as taxas de transição, mas sim os tempos gerados por distribuições não-paramétricas de probabilidade, e, para isso, utiliza-se a simulação de eventos discretos.

Como ponto negativo para essa abordagem, está o fato da necessidade de se ter uma base dados para representar o sistema que, por si só, já torna o processamento computacional mais lento no momento da simulação.

Como ponto positivo, está a flexibilidade da análise decorrente da não necessidade da parametrização das distribuições dos tempos. Além disso, por meio da integração com as redes bayesianas, o modelo markoviano possibilita a análise em tempo real da disponibilidade do sistema.

Ross (2002) mostra a importância da simulação discreta de eventos como um apoio no acompanhamento de um modelo ao longo do tempo e na determinação de quantidades de interesse. Esse mesmo autor descreve algumas variáveis de interesse a serem utilizadas ao longo da simulação, a saber:

Variáveis de tempo. Referem-se a quantidade de tempo simulado decorrido;

Variáveis contadoras. São usadas para contar o número de vezes que certo evento ocorreu até um tempo t ;

Variáveis de estado do sistema. Descrevem o estado do sistema num tempo t .

Utilizando a idéia de simulação discreta de eventos, descreve-se o algoritmo a seguir, mas antes é preciso definir alguns entes: $S = \{S_1, S_2, S_3, \dots, S_n\}$ – conjunto dos estados do sistema, $P_{S_i} = \{S_1, S_2, \dots, S_{i-1}, S_{i+1}, \dots, S_r\}$ – conjunto dos filhos do nó S_i , $C = \{c_1, c_2, \dots, c_l\}$ – conjunto dos componentes que formam os estados, $E(S_i) = \{E_{S_i}(c_1), E_{S_i}(c_2), \dots, E_{S_i}(c_l)\}$ – o estado do nó S_i baseado nas condições dos componentes que o constituem, t_{S_i} – tempo de permanência no estado S_i , v_{S_i} – o número de visitas ao estado S_i e t_{c_i} – soma do tempo operacional do componente c_i antes do reparo perfeito⁵ (soma dos tempos entre falhas antes do reparo perfeito). Além disso, a deterioração do sistema ao longo do tempo é modelado por meio de atualizações sucessivas das probabilidades condicionais de ocorrências dos tempos:

$$U_j = F_j(t|t_{c_j}) = \Pr(T \leq t | T \geq t_{c_j}) = \frac{\Pr(t_{c_j} \leq T \leq t)}{\Pr(T \geq t_{c_j})} = \frac{F_j(t) - F_j(t_{c_j})}{1 - F_j(t_{c_j})} \quad (3.1)$$

onde T é a variável aleatória do tempo de falha do componente c_j e $F_j(t)$ é a função de distribuição de probabilidade acumulada (CDF) do tempo de falha do componente c_j .

⁵É o reparo realizado de tal modo que a probabilidade de ocorrência de uma falha após o reparo seja igual à probabilidade de ocorrência de uma falha quando o produto se encontra em estado novo. Diz-se que nesse tipo de reparo o produto retorna à operação na condição “tão bom quanto novo”.

Assim:

$$F_j(t) = [1 - F_j(t_{c_j})] \cdot U_j + F_j(t_{c_j}) \quad (3.2)$$

Algoritmo 1 Algoritmo de cálculo das probabilidades de estado para o processo de Markov não-homogêneo

1. Fazer $t_{S_i} = 0$, $t_{c_j} = 0$ e $v_{S_i} = 0 \forall i, j$;
 2. Definir o estado inicial do sistema: $A = S_{i=k} = S_k$;
 3. Se $E_A(c_j)$ for operacional, para $j = 1, 2, \dots, l$ fazer:
 - (a) Se $F_j(t_{c_j}) = 1$, fazer $t_{c_j} = 0$ (reparo perfeito);
 - (b) Se $t_{c_j} = 0$, fazer $F_j(t_{c_j}) = 0$. Caso contrário, calcular $F_j(t_{c_j})$ de acordo com a distribuição de probabilidade do tempo entre falhas do componente c_j ;
 - (c) Gerar um número aleatório $U_j \in]0, 1[$;
 - (d) Fazer $p = F_j(t)$ de acordo com a Equação (3.2);
 - (e) Fazer $t_{S_k}^j = F_j^{-1}(p) - t_{c_j}$, onde $t_{S_k}^j$ é o tempo candidato de permanência no nó S_k provocado por alteração do estado do componente c_j ;
 4. Caso contrário, se $E(c_j)$ for falho, para $j = 1, 2, \dots, l$, fazer:
 - (a) Gerar um número aleatório $U \in]0, 1[$;
 - (b) Calcular $t_{S_k}^j = R_j^{-1}(U)$ baseado na distribuição de probabilidade do tempo de reparo do componente c_j ;
 5. Se $E_A(c_j) = E_{S_i}(c_j) \forall S_i \in P_A$, fazer $t_{S_k}^j = \infty$. Fazer isso para $j = 1, 2, \dots, l$;
 6. Se $E_A(c_j)$ for operacional, fazer $t_{c_j} = t_{c_j} + \min(t_{S_k}^j)$ (menor valor de $t_{S_k}^j$);
 7. Para $\min(t_{S_k}^j)$, realizar a mudança de condição do componente c_j em A . Se ele estiver operacional mudar para falho e vice-versa;
 8. Se $\sum_{i=1}^n t_{S_i}$ for menor do que o tempo de simulação pré-definido, então $t_{S_k} = t_{S_k} + \min(t_{S_k}^j)$; $prob(S_k) = t_{S_k} / \sum_{i=1}^n t_{S_i}$; $v_{S_k} = v_{S_k} + 1$; $k = i$ para $E(A) = E(S_i)$ e retornar para o passo 3. Caso contrário pare.
-

O algoritmo 1 trata-se de uma simulação discreta de eventos para um processo estocástico markoviano não-homogêneo⁶ com um tempo cronológico para cada equipamento. Pode-se assim caracterizá-lo, pois a cada execução da renovação de um componente que forma a cadeia no passo 3(a), há um reinício na contagem do tempo de operação para o

⁶Se o presente é conhecido, o passado não terá nenhuma influência no comportamento futuro do sistema, mas sim o tempo percorrido pelo sistema.

equipamento que possui uma frequência de falhas elevada. Com isso, CDF do tempo de falha para o componente j passa a ser representada por $F_j(t - T_R^j)$, em que t é o tempo total de vida do sistema e T_R^j é o tempo corrido pelo sistema desde o último reparo perfeito do equipamento j .

Os passos de 1 a 2 referem-se a definição dos valores dos estados e variáveis iniciais. Todos os passos do item 3 assumem que o sistema está num estado operacional. O 3(a) realiza um reparo perfeito quando a indisponibilidade instantânea tende a 1. É importante dizer que esse valor 1 é flexível e pode assumir outros valores. Por exemplo, quando a indisponibilidade instantânea for superior a 0,9, pode-se admitir que o reparo é perfeito, fazendo que com isso não sejam consideradas situações de extrema degradação do sistema. O passo 3(b) considera uma convenção clássica em que a indisponibilidade instantânea em $t = 0$ é igual a 0. Os passos 3(c) a 3(e) geram conjuntamente o tempo em que sistema permaneceu no atual estado, bem como um estado candidato para ser o próximo em que o sistema irá se encontrar na cadeia de Markov. O passo 4 realiza o mesmo que o anterior só que considerando o sistema em um atual estado falho. O passo 5 é um importante ponto, pois viabiliza a transição de estados quando os tempos de reparo e os tempos de falha encontram-se em ordens de grandeza diferentes. O passo 6 atualiza o tempo operacional dos componentes considerando o tempo de permanência no estado atual. O passo 7 realiza a transição do estado em si e o último passo executa o armazenamento das variáveis de interesse.

A título de exemplificação, se não for executado o passo 3(a) desse algoritmo, tem-se um processo markoviano não-homogêneo com o resultado para a disponibilidade⁷ de um sistema composto por três equipamentos em série sofrendo deterioração mostrado na Figura 3.4. Esse sistema trata-se do modelo ilustrado na Figura 4.12 que representa a operação dos três principais equipamentos de um poço de produção de óleo com o método de elevação mecânico a ser detalhado no próximo capítulo.

A Figura 3.4 mostra claramente que ao passo que o tempo avança a disponibilidade desse sistema tende ao valor zero. Isso acontece devido ao elevado nível de deterioração dos equipamentos com o transcorrer do tempo.

É fácil de verificar que tais resultados distanciam-se do chamado “mundo real”, uma vez que não é justificável manter um sistema apenas para está-lo reparando a quase todo

⁷A disponibilidade de um sistema num instante t pode ser definida como a soma das probabilidades dos estados disponíveis nesse mesmo momento.

tempo (disponibilidade instantânea tendendo a zero). Para sanar essa incompatibilidade, propõe-se justamente a renovação trazida pelo passo 3(a) do algoritmo 1.

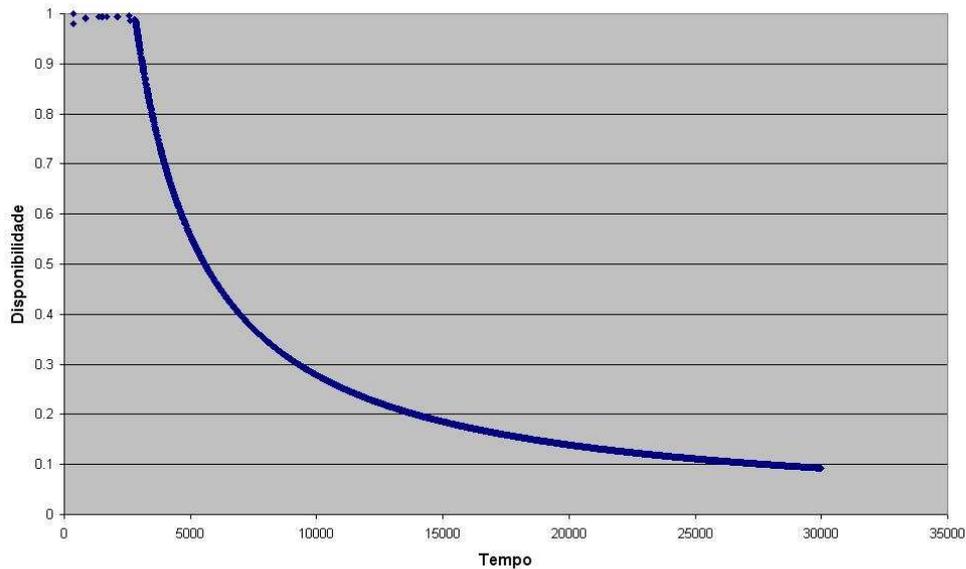


Figura 3.4: Exemplo de disponibilidade para um processo não-homogêneo de Markov

Alguns autores sugerem a adoção de critérios de reparo nem tão conservadores, visto que muitas vezes os equipamentos não são completamente renovados, e nem tão pessimistas. Yanez, Joglar e Modarres (2002) expõem uma metodologia de análise para modelar esses reparos, em que o equipamento não retorna à operação simplesmente “tão bom quanto novo” e nem somente “tão ruim quanto velho”, mas também, a depender da eficácia da equipe de manutenção, algo entre essas duas situações (“melhor do que velho mas pior do que novo”). Este será um dos pontos de sugestão para desenvolvimento de trabalhos futuros.

Com esse algoritmo, busca-se estimar probabilidades de estado mais próximas da realidade, uma vez que a deterioração do equipamento é considerado.

Complementando a abordagem até aqui utilizada, elabora-se o algoritmo 2 utilizando os mesmos entes até então apresentados. Esse algoritmo é utilizado para se ter uma idéia da variação dessas probabilidades de estado por meio da adição de uma visão mais otimista à análise, ou seja, assumindo-se também que após qualquer reparo os equipamentos sempre retornam à operação na condição tão bons quanto novos, desprezando assim a deterioração dos mesmos. Assim, o algoritmo 1 fornece uma visão mais pessimista (ou mais realista para casos em que a deterioração é relevante) para a disponibilidade, con-

siderando o desgaste dos equipamentos e renovando-os apenas num nível de alto desgaste, e o algoritmo uma visão mais otimista assumindo a renovação dos equipamentos a cada evento de manutenção (renovação por oportunidade).

Algoritmo 2 Algoritmo de cálculo das probabilidades de estado

1. Fazer $t_{S_i} = 0$ e $v_{S_i} = 0 \forall i$;
 2. Definir o estado inicial do sistema – $A = S_{i=k} = S_k$;
 3. Se $E_A(c_j)$ for operacional gerar um número aleatório U_j baseado na distribuição de falha de c_j , e se $E_A(c_j)$ for falho gerar U_j baseado na distribuição de reparo de c_j . Fazer isso para $j = 1, 2, \dots, l$;
 4. Se $E_A(c_j) = E_{S_i}(c_j) \forall S_i \in P_A$, fazer $U_j = \infty$. Fazer isso para $j = 1, 2, \dots, l$;
 5. Para o menor valor de U_j ($\min(U_j)$), realizar a mudança de condição do componente c_j em A , se ele estiver operacional mudar para falho e vice-versa;
 6. Se $\sum_{i=1}^n t_{S_i}$ for menor do que o tempo de simulação pré-definido, então $t_{S_k} = t_{S_k} + \min(U_j)$; $prob(S_k) = t_{S_k} / \sum_{i=1}^n t_{S_i}$; $v_{S_k} = v_{S_k} + 1$; $k = i$ para $E(A) = E(S_i)$ e retornar para o passo 3. Caso contrário pare.
-

Da mesma forma que o anterior, o algoritmo 2 faz uma modelagem por meio de simulação discreta de eventos, em que o caminho a ser seguido pelo sistema é determinado a partir de um número gerado de uma distribuição de probabilidade, que pode variar a depender da semente utilizada para a função aleatória *Random*. Sendo assim, torna-se necessário realizar diversas simulações e verificar o comportamento das probabilidades dos estados em termos do valor esperado.

Não será detalhado cada passo desse segundo algoritmo, pois acredita-se que com a explicação realizada no primeiro fica claro para o leitor o que cada etapa desse segundo algoritmo significa.

Para se ter uma idéia da diferença das probabilidades de estado em cada simulação, é evidenciada a disponibilidade ao longo do tempo de um sistema nas Figuras 3.5(a) e 3.5(b). Para tanto, definiu-se um sistema em série formado por dois equipamentos com taxas de falha e reparo constantes (tempos de transição exponenciais), sendo elas $\lambda_1 = 0,001$ e $\mu_1 = 0,07$ para o equipamento 1 e $\lambda_2 = 0,006$ e $\mu_2 = 0,03$ para o equipamento 2, em que o sistema só se encontra disponível quando os dois componentes estão em estado operacional. A Tabela 3.2 mostra o estado do sistema em função das condições operacionais de seus equipamentos e a Figura 3.6 a cadeia de markov homogênea que representa este sistema.

Assim, foram gerados 10.000 tempos aleatórios de falha e 10.000 de reparo para cada equipamento via o método da inversa ($X_{ij} = -\log(1 - U)/\lambda_j$ e $Y_{ij} = -\log(1 - U)/\mu_j$, para $i=1, 2, \dots, 10.000$ e $j=1,2$, onde U é um número aleatório entre 0 e 1) e agrupados em 20 intervalos para formar distribuições não-paramétricas destes pontos, e, para só então, aplicar o algoritmo 2 em um tempo de simulação de 500.000 unidades de tempo.

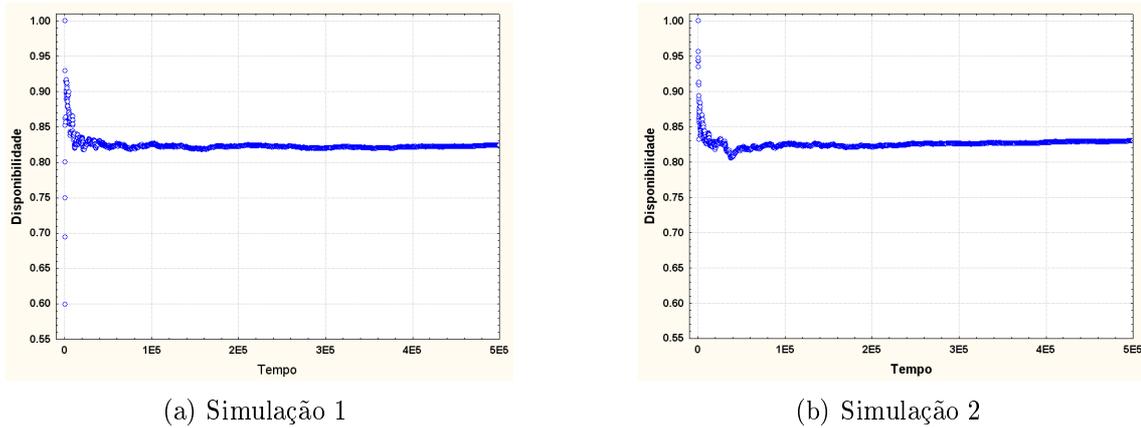


Figura 3.5: Simulação da disponibilidade de um sistema em série

Tabela 3.2: Exemplo de um sistema em série

Nó	Equipamento 1	Equipamento 2	Estado do sistema
0	Operacional	Operacional	Disponível
1	Operacional	Falho	Indisponível
2	Falho	Operacional	Indisponível

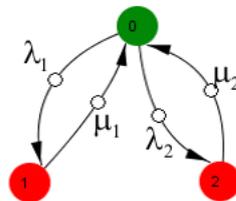


Figura 3.6: Cadeia de markov de um sistema em série

Para validar o algoritmo 2, foram realizadas simulações com várias configurações de rede com o algoritmo e comparadas com a resolução fornecida pela Equação 2.21, tendo sido observada a igualdade na convergência em todas elas. Como exemplo, é calculada a disponibilidade do sistema mostrada na Tabela 3.2 ao longo do tempo utilizando a

Equação 2.21 e confrontada com a disponibilidade média fornecida pelo algoritmo sugerido. Este foi implementado em linguagem de programação C++ para as mesmas situações do parágrafo anterior e numa quantidade de rodadas de 1.000. A Figura 3.7 mostra esses dois comportamentos.

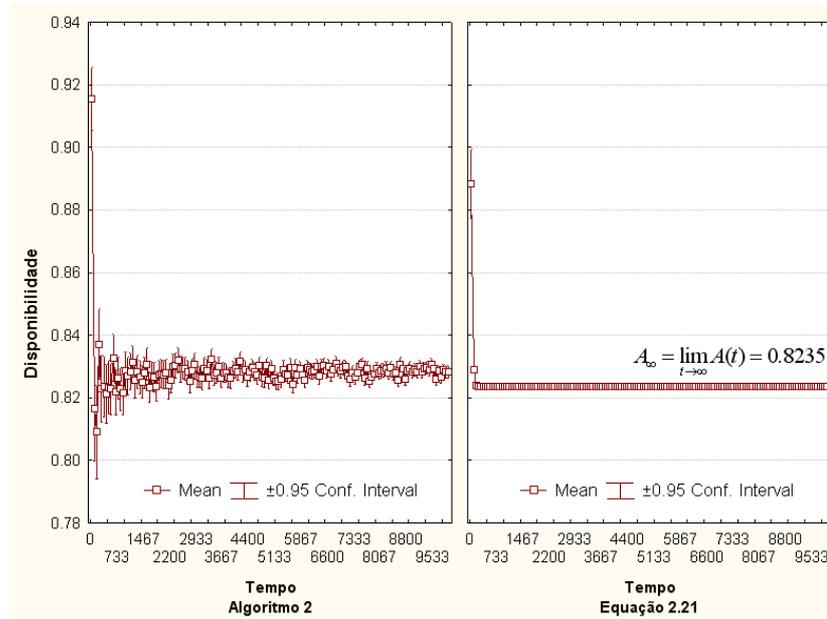


Figura 3.7: Comparação dos resultados fornecidos pelo algoritmo proposto 2 e pela Equação 2.21

Para validar o algoritmo 1, devido à inacessibilidade a um *software* que calcule a disponibilidade ou as probabilidades de estado de um sistema fazendo as suposições aqui apresentadas, é analisada apenas a lógica dos resultados. Se para o algoritmo 2 é admitida uma substituição do equipamento após qualquer reparo, independente dele estar deteriorado ou não, e no algoritmo 1 esse reparo perfeito só ocorre após um alto nível de deterioração, é de se esperar que a disponibilidade fornecida pelo algoritmo 1 seja inferior à calculada pelo algoritmo 2. Com isso, por meio do algoritmo 1, tem-se uma visão mais realista do sistema a ser analisado e com o algoritmo 2 uma visão mais idealista, a qual pode servir de meta para o gestor da área de manutenção.

O comportamento esperado desses resultados foi verificado e, como exemplo, tem-se o exposto na Figura 3.8. É fácil notar que o algoritmo 2 apresenta realmente uma disponibilidade média ao longo do tempo superior a fornecida pelo algoritmo 1.

O cálculo das probabilidades de estado fornece uma idéia sobre o tempo relativo de permanência em cada estado, possibilitando uma avaliação das equipes e da atividade de manutenção. Mas, além disso, é importante calcular as métricas que trazem informações

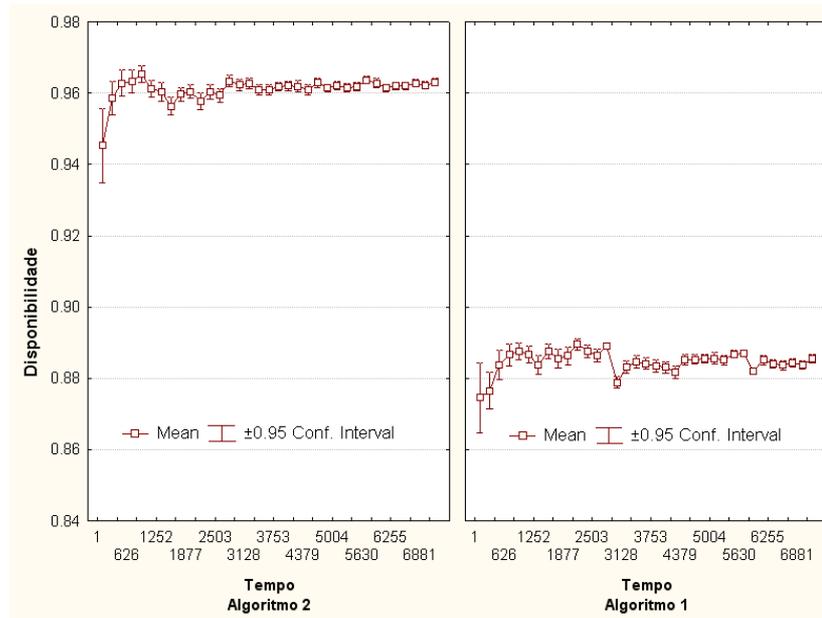


Figura 3.8: Disponibilidade fornecida pelos algoritmos 1 e 2

sobre o número de visitas a cada estado, o qual fornecerá uma estimativa do nível de ocorrência de falhas para cada equipamento.

Ao fim desta etapa, é permitido ao analista tomar decisões quanto à necessidade de redundância de equipamentos ou mudança de tecnologia dada a baixa disponibilidade do sistema. Um exemplo de aplicação dessas medidas será fornecido no capítulo seguinte.

3.8 Atualização do Estudo de Disponibilidade via Evidências nas Variáveis das Redes Bayesianas dos Equipamentos do Sistema

É nesta fase que as redes bayesianas são efetivamente utilizadas. Devido à capacidade de atualização das probabilidades condicionais das redes bayesianas é que a metodologia aqui apresentada possui a característica de ser uma ferramenta de análise em tempo real.

As atualizações são realizadas por meio de evidências nas variáveis monitoradas e propagadas para os TBF's e TTR's das redes bayesianas. Com essas atualizações nos tempos, obtém-se uma nova distribuição não-paramétrica para os mesmos. Após isso, a cadeia de Markov é retroalimentada com esses novos tempos e uma nova análise de disponibilidade é realizada utilizando os algoritmos propostos.

O cálculo de atualização da rede deve ser realizado conforme a Equação 2.5 e feito o passo 6 da metodologia, a fim de que se tenha uma nova análise de disponibilidade com as novas distribuições geradas para os TBF's e TTR's. Para um melhor entendimento do procedimento de atualização das redes bayesianas, pode-se rever o exemplo ilustrado na seção 2.4.2 que tem seu processo de atualização exibido na seção 2.5 do mesmo capítulo.

Enfim, com esta possibilidade de atualização, busca-se sair do lugar comum em que as taxas de transição da cadeia de markov são mantidas constantes em uma análise de fase, onde cada fase é parametrizada com as taxas de transição fornecidas, para uma cadeia que possui uma dinâmica baseada nos tempos de transição e, por sua vez, serão atualizados com mudanças nas probabilidades condicionais dos tempos entre falhas e de reparo da rede bayesiana devido a uma evidência numa variável monitorada do sistema.

A seguir será exposto um exemplo de aplicação da metodologia aqui apresentada no contexto dos sistemas complexos de produção de óleo, os poços de petróleo.

4 APLICAÇÃO DA METODOLOGIA PROPOSTA

4.1 Introdução

Buscando-se sistemas que apresentam oportunidades para aplicação da metodologia proposta, consideram-se poços de produção de petróleo. Tal sistema é escolhido devido à existência de um banco de dados relacional de tamanho considerável oriundo de uma empresa de extração de petróleo para a realização da mineração de dados (32.052 casos de ocorrência de falhas e reparos distribuídos em 1883 poços ao longo de 23 anos – 1983 a 2006), no qual os tempos entre falhas e de reparo estão relacionados com fatores intrínsecos do sistema.

Uma característica peculiar dessa base de dados é que os registros armazenam o histórico dos equipamentos instalados no poço e não dos equipamentos em si, o que na literatura é referido como *socket*. Para entender melhor essa diferença, é como se estivesse analisando o comportamento de indivíduos inseridos numa dada comunidade, o que é de se esperar um padrão de comportamento para os indivíduos ali inseridos.

A seguir, são descritas de maneira sucinta as formas de intervenção em poços de petróleo e os tipos de poços utilizados na aplicação da metodologia e, logo a seguir, os resultados encontrados fazendo uso dos passos metodológicos do capítulo anterior.

4.2 Intervenção em Poços de Petróleo

Segundo Silva (2004), intervenções em poços de petróleo (inglês: *workover*) são necessárias para manter a produção ou casualmente realizar melhorias na produção.

Geralmente, essas intervenções ocorrem com a utilização de uma sonda para realizar reparos na bomba, coluna de produção ou hastes, em que problemas com tais equipamentos geram perda de produtividade do poço, produção excessiva de gás ou água e produção de areia. Mas, em casos de manutenções mais simples, tais como substituição de uma válvula e/ou registro de pressão, o reparo pode ser realizado com um arame, não necessitando da instalação da sonda.

Assim, é importante destacar que na base de dados em análise, os casos de falha só foram registrados quando se fez uso da sonda, a qual é freqüentemente utilizada para

realizar reparos em equipamentos de subsuperfície.

4.3 Os Tipos de Poços Analisados

Os poços são comumente classificados quanto ao tipo de completção¹ utilizado em sua preparação para produção e ao tipo de bombeio da produção (método de elevação). A seguir será descrito o tipo de completção utilizado nos poços da base de dados e os tipos de bombeio.

4.3.1 Poços de Completção Convencional

Os poços de completção convencional ou de completção “burra” são aqueles poços em que, na fase de completção, tiveram equipamentos instalados que não permitem a monitoração e o controle de produção com precisão.

Poços de completção convencional não fornecem nenhuma informação sobre o poço e são incapazes de controlar o processo sem intervenções (acesso remoto para aquisição de dados e alterações de fluxo das zonas de produção) (SILVA, 2004). Poços com esse tipo de completção diferenciam-se dos poços de completção inteligente devido aos equipamentos de controle e monitoramento instalados nesses últimos.

Nesse trabalho, apenas os poços de completção convencional serão analisados, pois não há base de dados referentes aos poços de completção inteligentes de tamanho suficiente para o uso da ferramenta de estruturação automática da topologia das redes bayesianas como mineração de dados. A seguir são descritos os métodos de elevação.

4.3.2 Métodos de Elevação Artificial

Chama-se método de elevação todo o conjunto de operações, após a completção, necessárias para que o óleo extraído atinja a superfície. Dentro do conjunto de métodos de elevação é comum encontrar a seguinte classificação: método de elevação natural e método de elevação artificial.

¹A fase de completção ocorre após o estágio de perfuração do poço e é nela em que são instaladas as tubulações do poço (revestimento e filtro), o cascalho (pré-filtro) e o cimento (cimentação). Trata-se de um conjunto de operações para equipar um poço de maneira a deixá-lo em condições de produzir petróleo de forma segura e econômica, reduzindo a necessidade de futuras intervenções para manutenção e elevando a vazão de produção. É nessa fase em que são instalados os equipamentos de controle inteligente. Para maiores detalhes ver Thomas (2000).

Quando a pressão do óleo no interior do poço é suficientemente alta para alcançar a superfície, o poço é chamado de surgente, e a sua elevação é chamada de natural. Já o método de elevação artificial é aquele utilizado em poços os quais a pressão do reservatório é relativamente baixa e não consegue alcançar a superfície naturalmente.

Na aplicação da metodologia proposta, somente serão utilizados casos de métodos de elevação artificial, mais especificamente dois deles: Bombeio Mecânico com Hastes (BM) e Bombeio por Cavidades Progressivas (BCP), uma vez que esses são os únicos métodos de elevação que são registrados em quantidades relevantes na base de dados. Em ambos os métodos, a coluna de tubos, haste e bomba são os principais equipamentos de subsuperfície.

O tubo no sistema poço possui função semelhante para os dois tipos de métodos de elevação: revestem-se as paredes do poço evitando desmoronamentos e vazamento de gás por entre os poros das rochas, e permite-se a entrada da haste por dentro dele. Porém, as hastes e a bomba trabalham de modos diferentes em cada método de elevação, sendo descritos nas seções seguintes.

4.3.3 Bombeio Mecânico com Hastes

O BM é um dos métodos de elevação artificial mais utilizados em todo o mundo. Nesse método de elevação, o movimento rotativo de um motor é transformado em movimento alternativo por uma unidade de bombeio na superfície, o qual é transmitido para a coluna de hastes. A bomba, por sua vez, é responsável por transmitir a energia para o fluido, elevando-o para a superfície através de dois cursos, o ascendente (inglês: *upstroke*) e o descendente (inglês: *downstroke*). No curso ascendente o fluido no reservatório vai para o interior da bomba e no descendente o fluido que já estava no interior da bomba é passado para o interior do tubo acima da bomba, subindo assim para a superfície. A Figura 4.1 ilustra um poço com o método de elevação BM.

As hastes e os tubos estão sujeitos a cargas cíclicas, uma vez que o fluido acima da bomba é sustentado pela coluna de hastes no curso ascendente e pela coluna de tubos no curso descendente.



Figura 4.1: Poço com método de elevação artificial de bombeio mecânico com hastes

4.3.4 Bombeio por Cavidades Progressivas

Nesse método de elevação artificial, a energia rotativa de um motor na superfície é transmitida para a coluna de hastes que gira com um torque constante, fazendo com que a energia dispendida pelo motor seja menor do que a necessária no bombeio mecânico. A Figura 4.2 ilustra um poço com o método de elevação artificial BCP.

A bomba, nesse método de elevação artificial, fica imersa no poço e é composta unicamente por um camisa estacionária – o estator de material macio e de forma espiral fêmea – e por um rotor – constituído de aço e de forma espiral macho. Quando os dois são encaixados, uma série de cavidades seqüenciais permanece entre eles, possibilitando a passagem do fluido. O fluido é bombeado por meio da rotação da coluna de haste, a qual transmite a força rotacional para o rotor e provoca o deslocamento das cavidades seqüenciais de uma extremidade da bomba para a outra.

Atualmente, existem tecnologias que acoplam um motor elétrico diretamente na bomba, fazendo com que o rotor gire sem a necessidade de rotação da haste. Isso diminui não só as cargas nas hastes – que muitas vezes têm dificuldade de rotação por causa da viscosidade do fluido ou pela fricção com a coluna de tubos – como também o consumo de energia.



Figura 4.2: Poço com método de elevação artificial de bombeio por cavidades progressivas

4.4 Aplicação da Metodologia

Deste ponto em diante, até ao final do capítulo, serão descritos em cada seção os principais resultados obtidos com a aplicação da metodologia apresentada no capítulo anterior. Será, portanto, cada passo metodológico discutido numa seção.

A título de entendimento, torna-se interessante mostrar os principais pontos que serão tratados mais adiante:

1. Serão descritas as variáveis e os equipamentos relevantes para a análise. Nesta etapa será dito que os equipamentos relevantes a serem investigados para uma análise do sistema de subsuperfície serão as bombas BCP e BM, a depender do tipo de método de elevação em análise, os tubos e as hastes. Além disso, variáveis de monitoramento serão descritas, tais como: nível de água e sólidos no interior do poço, nível de parafina, profundidade do poço, etc.
2. Serão mostradas as redes construídas com o algoritmo de aprendizagem de topologia (SOP, seção 2.7.2) e comparadas com a opinião do especialista, quando houver divergências na construção. No total serão exibidas 8 redes bayesianas, sendo uma para cada equipamento e tipo de tempo em análise, a exceção das bombas que

serão diferenciadas se é do método de elevação BM ou BCP. Por exemplo: para as hastes, será uma BBN para análise das relações causais que influenciam o tempo entre falhas e outra para análise das relações que influenciam o tempo de reparo, o mesmo ocorre para os tubos. Já no caso das bombas, elas serão segregadas em BM e BCP, tendo, a partir disso, a mesma estrutura de análise das hastes e dos tubos.

3. Será construída uma cadeia de Markov, conforme Figura 4.3, para avaliar a disponibilidade do sistema de subsuperfície de um poço de produção de óleo. Cada nó dessa cadeia é representado pelo conjunto dos estados operacionais dos equipamentos. Por exemplo: O nó 0 é caracterizado quando todos os equipamentos (bomba, haste e tubo) encontram-se operacional. Já o nó 1 é caracterizado por bomba operacional, haste operacional e tubo falho. A mesma idéia é adotada para os outros dois nós da rede.

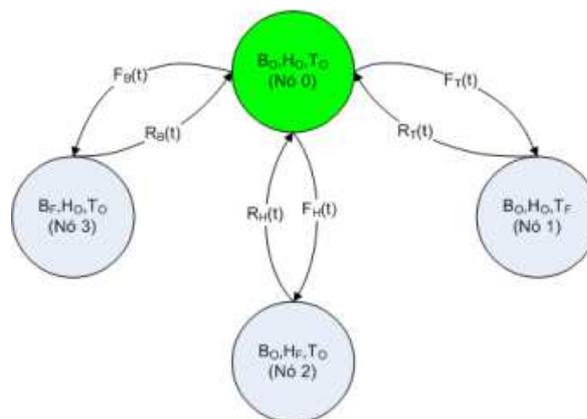


Figura 4.3: Cadeia de Markov para o sistema de subsuperfície de um poço de produção de óleo

4. Após isso, será ilustrada uma análise de disponibilidade assumindo hipoteticamente evidências em uma das variáveis das BBNs anteriormente descritas.

4.4.1 Equipamentos e Variáveis Relevantes

Analisando os 10 equipamentos existentes na base de dados, constata-se que apenas 3 deles apresentam relevância em termos do número de falhas e do tempo em que os mesmos permanecem em estado falho (bomba, haste e tubo). O restante dos equipamentos apesar de apresentarem tempos médios de reparo relativamente altos quando comparados com os 3 citados, a quantidade de ocorrências de falha é bastante pequena, o que torna irrelevante

para a análise do sistema como um todo. Esse resultado pode ser averiguado na Tabela 4.1. É importante destacar que todos os equipamentos listados na Tabela 4.1 apresentam o mesmo nível de risco² para o processo.

Tabela 4.1: Lista de equipamentos encontrados na base de dados

Equipamento	Num. Falhas	% Num. Falhas	Soma Tempo Reparo	Média Tempo Reparo
BOMBA	13.833	43,16	81.540	5,90
BOMBA BCS	1	0,00	0	0,00
DRENADOR	4	0,01	22	5,50
FILTRO	253	0,79	669	2,64
HASTE	10.377	32,38	48.097	4,64
HASTE POLIDA	116	0,36	522	4,50
ON OFF	6	0,02	13	2,17
PACKER	14	0,04	134	9,57
PONY ROD	47	0,15	132	2,81
REVESTIMENTO	16	0,05	246	15,38
SEPARADOR	1	0,00	4	4,00
TUBO	7.384	23,04	79.642	10,80
Total	32.052	100,00	211.021	6,59

Devido ao diferente funcionamento do poço, a depender do tipo de método de elevação utilizado, serão considerados dois tipos de sistema em funcionamento, o sistema com bombeio BCP e o sistema com bombeio BM, onde em cada sistema será utilizada a haste e o tubo mais o tipo de bomba correspondente, BCP ou BM, a depender do caso.

Assim, ao final da aplicação de toda a metodologia, ter-se-á duas cadeias de Markov para análise da disponibilidade, uma para cada tipo de sistema.

Após a definição dos equipamentos que farão parte da análise, torna-se necessário definir as variáveis intrínsecas de funcionamento do sistema que servirão para a construção das redes bayesianas.

Na Tabela 4.2 é evidenciada a lista de variáveis a serem adotadas nas redes bayesianas. A escolha das variáveis foi baseada na opinião de especialista, sendo discriminada toda variável que apresentasse alguma relevância para a ocorrência do tempo entre falhas e de reparo.

²Risco = Probabilidade de ocorrência de um evento \times consequência trazida pelo evento (quantidade de dinheiro gasto).

Tabela 4.2: Lista de variáveis relevantes para os tempos entre falhas e de reparo

Nome da variável	Descrição da variável
BSWSOT	Percentual de H ₂ O e sólidos no poço.
CINCO	Possui hastes de 5/8" (1-sim ou 0-não).
CLASSE1(T-1)	Classificação da haste de 1" instalada na última manutenção (0-nova, 1-semi-nova, 2-meia-vida, 3-sucata e 9-nenhuma das opções).
CLASSE34(T-1)	Classificação da haste de 3/4" instalada na última manutenção (0-nova, 1-semi-nova, 2-meia-vida, 3-sucata e 9-nenhuma das opções).
CLASSE58(T-1)	Classificação da haste de 5/8" instalada na última manutenção (0-nova, 1-semi-nova, 2-meia-vida, 3-sucata e 9-nenhuma das opções).
CLASSE78(T-1)	Classificação da haste de 7/8" instalada na última manutenção (0-nova, 1-semi-nova, 2-meia-vida, 3-sucata e 9-nenhuma das opções).
COMPRIMENT	Comprimento da Bomba.
DUASVPAS	Existência ou não de duas válvulas de passeio na bomba.
FILTRO	Classificação do filtro instalado.
FOLGA	Folga entre pistão e camisa do poço (geralmente determinada pela viscosidade do óleo).
GRAUTUBO(T-1)	Material do tubo descido na manutenção anterior.
H2S	Nível de H ₂ S.
HASTES	Combinação das hastes descidas. Exemplo: 76=7/8 e 3/4 86=1", 7/8, 3/4.
LOCAL	Local de falha do equipamento na ocorrência atual.
LOCAL(T-1)	Local de falha do equipamento na última ocorrência.
ME	Método de elevação (Bombeio mecânico com hastes (BM) ou bombeio por cavidades progressivas (BCP)).
MODO	Efeito de falha.
MODO(T-1)	Efeito de falha da última manutenção.

continua na próxima página

Tabela 4.2: Lista de variáveis relevantes para os tempos entre falhas e de reparo (continuação)

Nome da variável	Descrição da variável
OITO	Possui hastes de 1" (1-sim ou 0-não).
PARAFINA	Nível de parafina.
PERDASOT	Perda de produção. Quanto o poço está perdendo por dia.
PESOREV	Peso do revestimento por unidade de comprimento.
PRODSOT	Produção normal de óleo do poço.
PROF_BOMBA	Profundidade do local da bomba.
PROFUNDIDADE	Profundidade do poço.
REV	Diâmetro do revestimento.
SEIS	Possui hastes de 3/4" (1-sim ou 0-não).
SETE	Possui hastes de 7/8" (1-sim ou 0-não).
SPEC	Especificação da bomba em termos de material.
SPT	Número da sonda utilizada na manutenção.
SUBS_HASTE	Substituição ou não das hastes: S-Sim, P- Parcial e N-Nenhuma.
SUBS_HASTE(T-1)	Substituição ou não das hastes no último reparo: S-Sim, P-Parcial e N-Nenhuma.
SUBS_BOMBA(T-1)	Se houve substituição da bomba na última intervenção para reparo da bomba.
TBetweenFailures	Tempo entre falhas por tipo de equipamento.
TIPO(T-1)	Tipo de manutenção realizada na última falha.
TIPOLUVA	Tipo do material da luva.
TtoRepair	Tempo de reparo por tipo de equipamento.
TUBOESTADO(T-1)	Classificação do tubo descido na manutenção anterior (escala de novo a sucata: NOVO->AMARELO->AZUL->VERDE->CINZA->VERMELHO).
TUBOS	Diâmetro da tubulação.

Como será construída uma rede bayesiana para os tempos entre falhas e outra para o tempo de reparo de cada equipamento, as Tabelas 4.3 a 4.6 fazem esta distinção para as variáveis descritas na Tabela 4.2, imputando o valor "Sim" caso a variável influencie no

tempo da coluna (falha ou reparo) e “Não” caso contrário.

Tabela 4.3: Lista de variáveis a serem inseridas nas BBN's da bomba BCP

Variável	Tempo entre falhas	Tempo de reparo
BSWSOT	Sim	Sim
FILTRO	Sim	Sim
H2S	Sim	Não
LOCAL	Não	Sim
LOCAL(T-1)	Sim	Não
MODO	Não	Sim
MODO(T-1)	Sim	Não
PARAFINA	Sim	Sim
PERDASOT	Não	Sim
PRODSOT	Sim	Sim
PROF_BOMBA	Sim	Sim
SPT	Não	Sim
SUBS_BOMBA(T-1)	Sim	Não
TIPO_BOMBA	Sim	Sim

Tabela 4.4: Lista de variáveis a serem inseridas nas BBN's da bomba BM

Variável	Tempo entre falhas	Tempo de reparo
BSWSOT	Sim	Sim
COMPRIMENT	Sim	Sim
DUASVPAS	Sim	Sim
FILTRO	Sim	Sim
FOLGA	Sim	Não
H2S	Sim	Não
LOCAL	Não	Sim
LOCAL(T-1)	Sim	Não
MODO	Não	Sim

continua na próxima página

Tabela 4.4: Lista de variáveis a serem inseridas nas BBN's da bomba BM (continuação)

Variável	Tempo entre falhas	Tempo de reparo
MODO(T-1)	Sim	Não
PARAFINA	Sim	Sim
PERDASOT	Não	Sim
PRODSOT	Sim	Sim
PROF_BOMBA	Sim	Sim
SPEC	Sim	Não
SPT	Não	Sim
SUBS_BOMBA(T-1)	Sim	Não
TIPO_BOMBA	Sim	Sim

Tabela 4.5: Lista de variáveis a serem inseridas nas BBN's da Haste

Variável	Tempo entre falhas	Tempo de reparo
BSWSOT	Sim	Sim
CINCO	Sim	Sim
CLASSE1(T-1)	Sim	Não
CLASSE34(T-1)	Sim	Não
CLASSE58(T-1)	Sim	Não
CLASSE78(T-1)	Sim	Não
FILTRO	Sim	Sim
H2S	Sim	Não
HASTES	Sim	Sim
LOCAL	Não	Sim
LOCAL(T-1)	Sim	Não
ME	Sim	Sim
MODO	Não	Sim
MODO(T-1)	Sim	Não
OITO	Sim	Sim
PARAFINA	Sim	Sim

continua na próxima página

Tabela 4.5: Lista de variáveis a serem inseridas nas BBN's da Haste (continuação)

Variável	Tempo entre falhas	Tempo de reparo
PERDASOT	Não	Sim
PRODSOT	Sim	Sim
PROFUNDIDADE	Sim	Sim
SEIS	Sim	Sim
SETE	Sim	Sim
SPT	Não	Sim
SUBS_HASTE	Não	Sim
SUBS_HASTE(T-1)	Sim	Não

Tabela 4.6: Lista de variáveis a serem inseridas nas BBN's do Tubo

Variável	Tempo entre falhas	Tempo de reparo
FOLGA	Sim	Não
ME	Sim	Sim
PRODSOT	Sim	Sim
BSWSOT	Sim	Sim
FILTRO	Sim	Sim
H2S	Sim	Não
PARAFINA	Sim	Sim
TIPO_BOMBA	Sim	Sim
PERDASOT	Não	Sim
LOCAL	Não	Sim
LOCAL(T-1)	Sim	Não
TIPO(T-1)	Sim	Sim
MODO	Não	Sim
MODO(T-1)	Sim	Não
REV	Sim	Sim
PESOREV	Sim	Sim
SPT	Não	Sim

continua na próxima página

Tabela 4.6: Lista de variáveis a serem inseridas nas BBN's do Tubo (continuação)

Variável	Tempo entre falhas	Tempo de reparo
TUBOS	Sim	Não
TUBOESTADO(T-1)	Sim	Não
PROFUNDIDADE	Sim	Sim
GRAUTUBO(T-1)	Sim	Não
TIPOLUVA	Sim	Não

4.4.2 Limpeza e Integração dos Dados (Pré-Processamento)

Para a realização da limpeza dos dados, com o intuito de retirar valores discrepantes da amostra, adota-se um ponto de corte para valores superiores a 97,5% de uma distribuição não-paramétrica dos dados. Esta medida é necessária devido aos dados apresentarem poucos pontos “soltos” com valores bastante superiores aos apresentados pelas ocorrências em geral e, além disso, tais pontos serem provocados por outros motivos que não tenham sido devido a reparo realizado ou falha ocorrida.

Para a realização da integração, adota-se o algoritmo EM (inglês: *Expectation-Maximization algorithm*) que é uma forma geral e iterativa de se estimar o valor de máxima verossimilhança dos parâmetros de uma distribuição associada quando se têm valores faltantes na base de dados. Para isso, assume-se a distribuição conjunta $p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)p(x|\Theta)$ com parâmetros Θ a serem estimados, em que χ é o conjunto dos dados incompletos e $\mathcal{Z} = (\chi, \gamma)$ é o conjunto de dados completos após a imputação. Para maior detalhe ver Dempster, Laird e Rubin (1977).

4.4.3 Estruturação das Relações Causais entre as Variáveis do Sistema e os Tempos entre Falhas e de Reparo

Para evidenciar as relações causais entre as variáveis monitoradas, mencionadas na seção 4.4.1, e os tempos entre falhas e de reparo definidos para os dois sistemas em análise, são estruturadas as topologias de oito redes bayesianas.

Essas BBN's são mostradas aos pares, com a solução fornecida pelo algoritmo de aprendizagem (SOP, seção 2.7.2) e com as correções realizadas pelo especialista. As que não venham a estar pareadas é devido ao especialista confirmar a topologia sugerida pelo algoritmo. Essas BBN's são mostradas da Figura 4.4 à Figura 4.11.

Cabe aqui mais uma vez lembrar, que serão as distribuições não-paramétricas dos Tempos (TBFs e TTRs) que servirão de entrada para o modelo markoviano ilustrado no início da seção 4.4 e que será discutido mais em detalhes na seção 4.4.5.

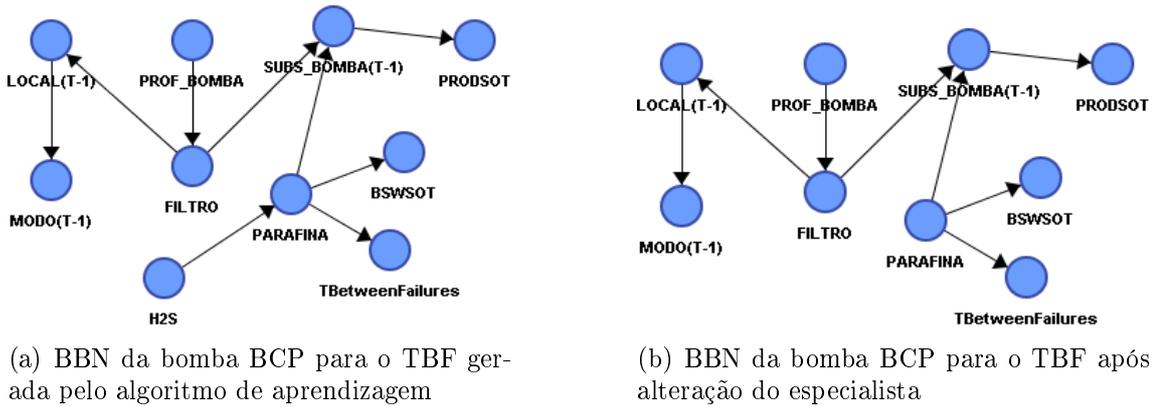


Figura 4.4: BBN da bomba BCP para o TBF

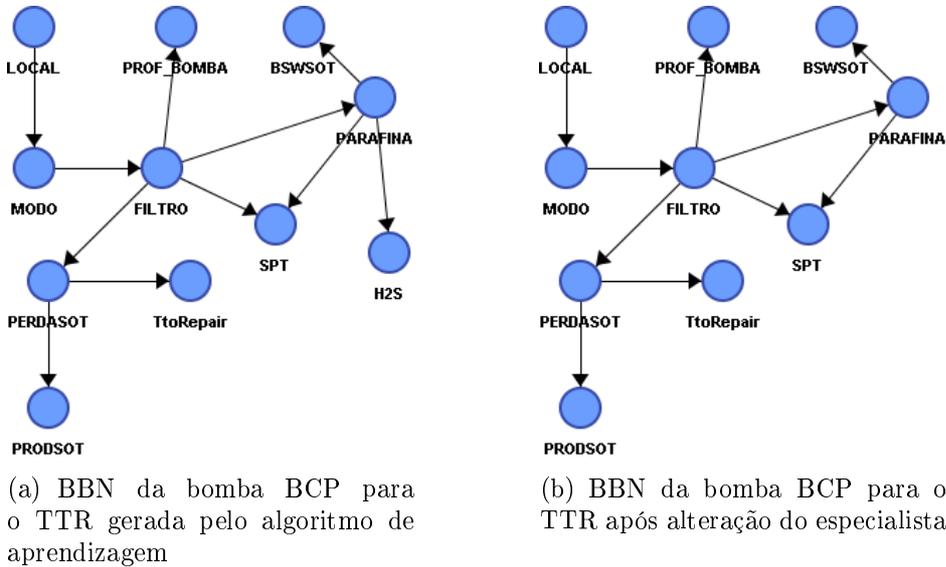


Figura 4.5: BBN da bomba BCP para o TTR

Pôde-se notar que em algumas redes existem nós que estão desconectados de toda a rede. Isso é devido a tais nós não apresentarem relações relevantes com os outros nós das redes. É verdade que tais nós poderiam ser incluídos se fosse modificado o peso

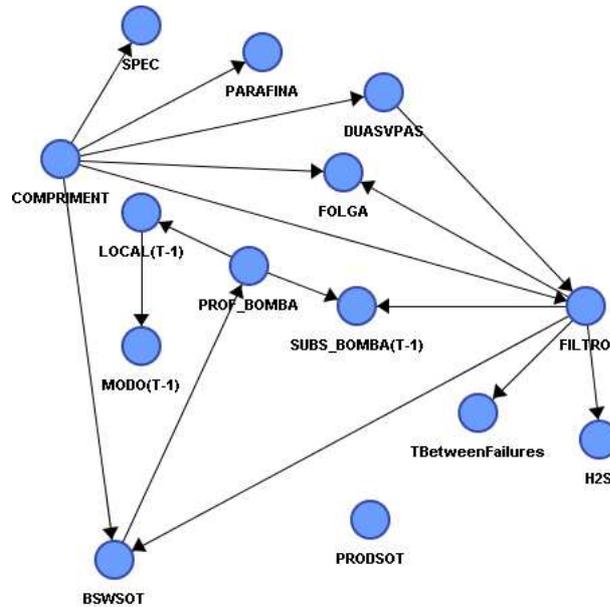


Figura 4.6: BBN da bomba BM para o TBF

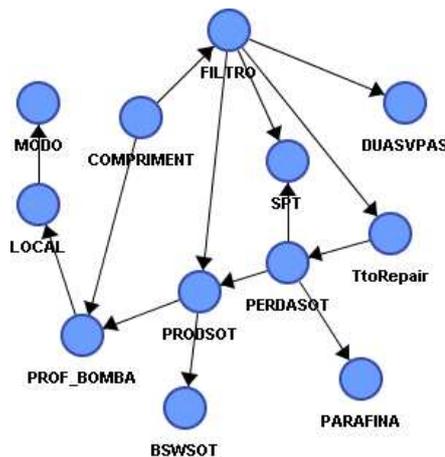


Figura 4.7: BBN da bomba BM para o TTR

do componente da soma do detalhamento da codificação dos dados dado o modelo da função de pontuação do MDL, mas o interesse nesse momento é o de conseguir incluir o tempo dentro da estrutura da rede e, comitadamente, não deixá-la complicada para o entendimento do profissional que a for utilizar.

No desenvolvimento do problema, de acordo com a metodologia proposta nesse trabalho, serão utilizadas as redes que foram modificadas pelo especialista.

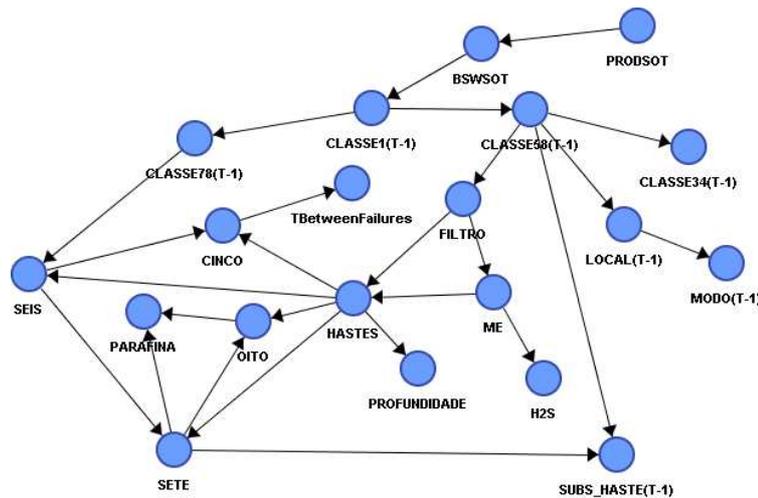


Figura 4.8: BBN da haste para o TBF

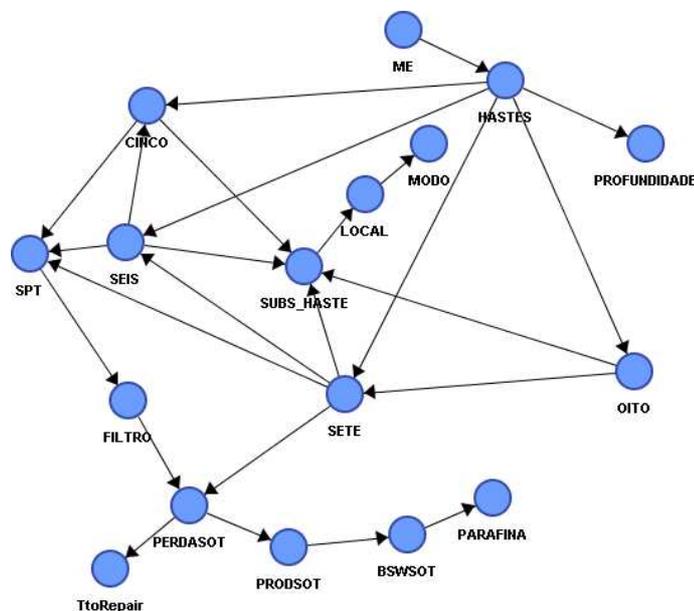


Figura 4.9: BBN da haste para o TTR

4.4.4 Estimação dos Parâmetros das Redes Bayesianas Estruturadas

Com as redes estruturadas, são estimadas as probabilidades condicionais a partir da base de dados, em que os fatores condicionantes são os valores dos pais de cada nó. Sabendo disso, como discutido na seção 3.5, utilizou-se a teoria freqüentista para estimar tais probabilidades.

Como exemplo de uma TPC, tem-se a Tabela 4.7 que mostra as probabilidades condicionais da variável “TtoRepair” dado seu pai “PERDASOT” da Figura 4.11.

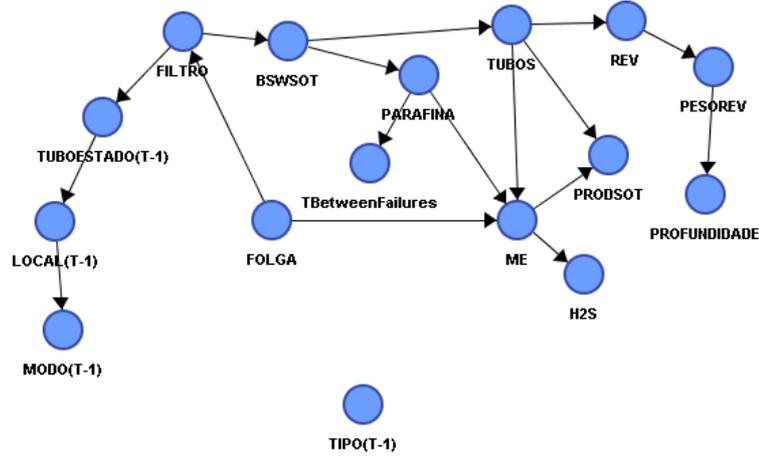


Figura 4.10: BBN do tubo para o TBF

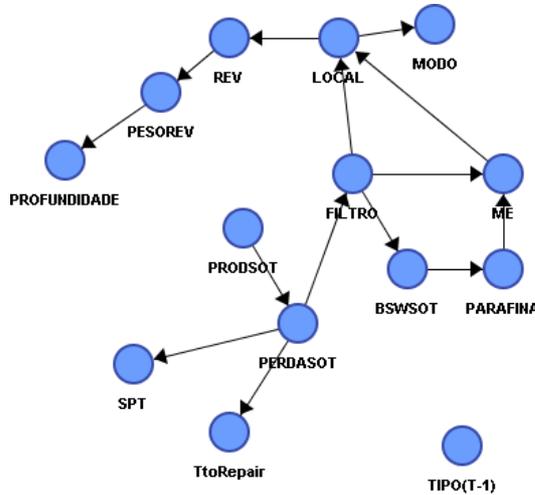


Figura 4.11: BBN do tubo para o TTR

4.4.5 Estruturação da Cadeia Markoviana do Sistema

Com os parâmetros estimados, pode-se partir para a estruturação da cadeia de Markov. Uma cadeia para cada tipo de método de elevação (BM e BCP) é estruturada. A Figura 4.12 mostra a estrutura adotada nos dois sistemas.

Cada nó é constituído pelo *status* operacional do conjunto dos três equipamentos (bomba-B, haste-H e tubo-T) onde o índice *O* significa operacional e o índice *F* falho. Além disso, é suposto um sistema em série, onde se ao menos um equipamento falhar o sistema se torna indisponível, e também a causa comum de falha inverossímil, o que neste, leva a não consideração de falhas simultâneas de equipamentos. Ver Tabela 4.8 para uma melhor visualização da caracterização dos nós da rede.

Por exemplo: o nó 1 caracteriza o sistema como indisponível, uma vez que retrata o

Tabela 4.7: TPC do tempo até o reparo dado o nível de perda para a BBN do tubo

PERDASOT	TtoRepair									
	≤ 1	≤ 1,92	≤ 3,84	≤ 5,76	≤ 7,68	≤ 9,6	≤ 11,52	≤ 13,44	≤ 15,36	>15,36
≤ 1	0,9775	0,0034	0,0092	0,0147	0,0370	0,0605	0,0523	0,0533	0,1333	0,1702
≤ 2	0,0000	0,0407	0,0552	0,1186	0,1992	0,2491	0,3464	0,3200	0,4667	0,4681
≤ 3,375	0,0008	0,0712	0,0936	0,1736	0,2382	0,2883	0,3072	0,3733	0,1333	0,2766
≤ 6,25	0,0041	0,1831	0,3129	0,3802	0,3573	0,3132	0,2745	0,2533	0,2444	0,0638
≤ 9,125	0,0033	0,1390	0,1925	0,1858	0,1068	0,0569	0,0131	0,0000	0,0222	0,0213
≤ 12	0,0011	0,1254	0,0790	0,0538	0,0308	0,0142	0,0065	0,0000	0,0000	0,0000
≤ 14,875	0,0008	0,0475	0,0445	0,0220	0,0082	0,0071	0,0000	0,0000	0,0000	0,0000
≤ 17,75	0,0014	0,0475	0,0522	0,0171	0,0021	0,0036	0,0000	0,0000	0,0000	0,0000
≤ 20,625	0,0016	0,0780	0,0330	0,0110	0,0041	0,0036	0,0000	0,0000	0,0000	0,0000
≤ 23,5	0,0014	0,0542	0,0230	0,0086	0,0062	0,0000	0,0000	0,0000	0,0000	0,0000
≤ 26,375	0,0016	0,0271	0,0245	0,0012	0,0021	0,0000	0,0000	0,0000	0,0000	0,0000
≤ 29,25	0,0011	0,0271	0,0176	0,0000	0,0021	0,0000	0,0000	0,0000	0,0000	0,0000
≤ 32,125	0,0008	0,0102	0,0100	0,0037	0,0021	0,0036	0,0000	0,0000	0,0000	0,0000
≤ 35	0,0003	0,0102	0,0084	0,0012	0,0021	0,0000	0,0000	0,0000	0,0000	0,0000
≤ 37,875	0,0008	0,0034	0,0084	0,0024	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
≤ 40,75	0,0011	0,0203	0,0054	0,0012	0,0021	0,0000	0,0000	0,0000	0,0000	0,0000
≤ 43,625	0,0005	0,0237	0,0054	0,0037	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
≤ 46,5	0,0003	0,0305	0,0054	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
≤ 49,375	0,0003	0,0237	0,0061	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
≤ 52,25	0,0008	0,0170	0,0061	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
≤ 55,125	0,0000	0,0068	0,0046	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
>55,125	0,0005	0,0102	0,0031	0,0012	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

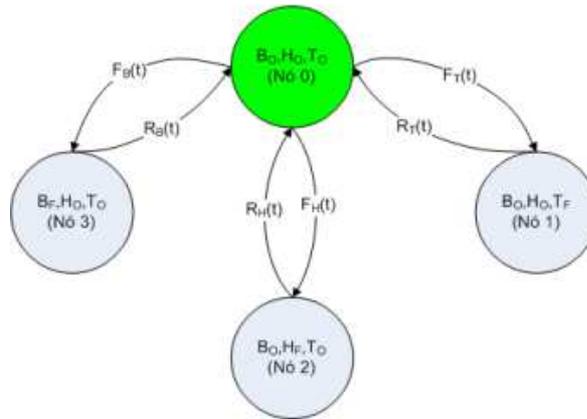


Figura 4.12: Cadeia de Markov para o sistema de subsuperfície de um poço

estado de um conjunto de componentes em série onde a bomba e a haste estão operacionais (B_O e H_O) e o tubo está falho (T_F). O mesmo raciocínio deve ser aplicado para os outros nós da rede.

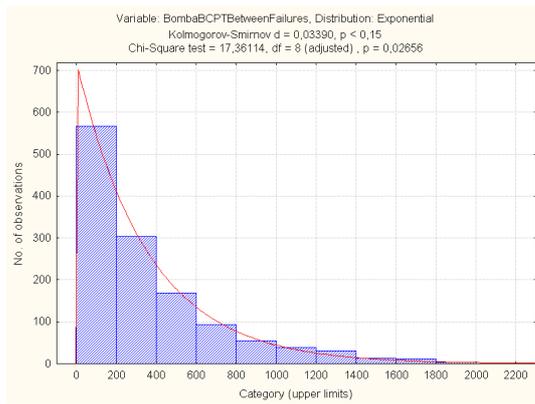
Tabela 4.8: Caracterização dos nós da cadeia de Markov

Nó	Sistema	Bomba	Haste	Tubo
0	Disponível	Operacional	Operacional	Operacional
1	Indisponível	Operacional	Operacional	Falho
2	Indisponível	Operacional	Falho	Operacional
3	Indisponível	Falho	Operacional	Operacional

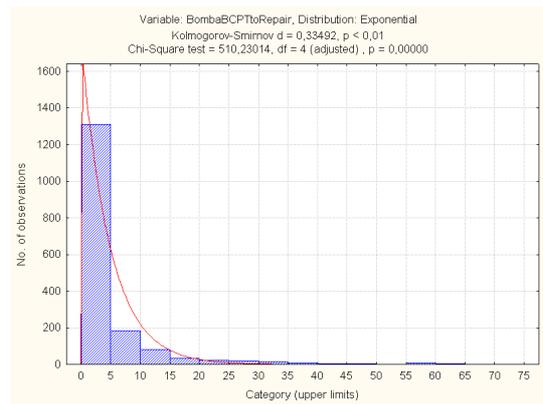
Para a aplicação dos algoritmos propostos, é necessário ter as distribuições dos tem-

pos entre falhas e de reparo para cada equipamento que caracterizam os nós da rede, o que são representadas por $F_i(t)$ e $R_i(t)$, respectivamente. Como já explicado, os $F_i(t)$ e $R_i(t)$ são obtidos através dos nós de saída das redes bayesianas para cada equipamento. As Figuras 4.13 à 4.16 mostram as distribuições não-paramétricas dos TBF's e TTR's para cada equipamento que caracterizam os nós da cadeia de Markov. Pode-se verificar nessas figuras, por meio dos testes de aderência a uma distribuição exponencial (*Kolmogorov Smirnov* e *Chi-Square*) com parâmetro encontrado via estimador de máximo-verossimilhança e para um *p-value* crítico igual a 0,05, que não é correto admitir as taxas de transição constantes entre todos os estados, portanto, um processo homogêneo de Markov é um modelo inadequado para analisar o sistema. Dessa forma, é necessário utilizar o processo não-homogêneo, o que é calculado pelo Algoritmo 1.

Esses testes de aderência comumente fazem uso de um ponto crítico e confrontam uma hipótese nula com uma hipótese alternativa a respeito de alguma medida. Se a estatística do teste (*p-value*) for inferior ao ponto crítico, diz-se que pode-se aceitar a hipótese alternativa, que nesse caso é um não ajustamento a uma distribuição exponencial. Como esta, por sua vez, tem sido usada como uma distribuição para modelar a renovação, diz-se assim não ser adequado assumir o modelo homogêneo de Markov. Para maiores detalhes a respeito dos testes *Kolmogorov Smirnov* e *Chi-Square*, ver Rubinstein (1981) ou Wackerly, Mendenhall e Scheaffer (1996).

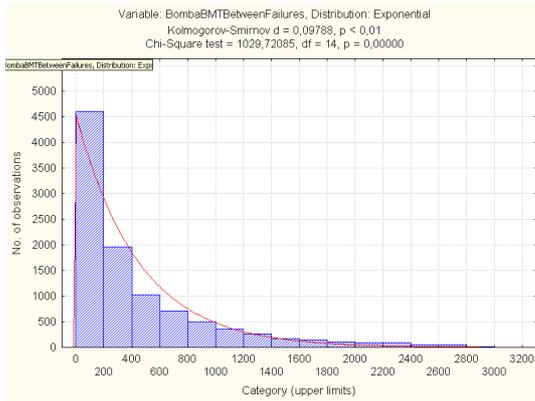


(a) Distribuição do TBF da bomba BCP

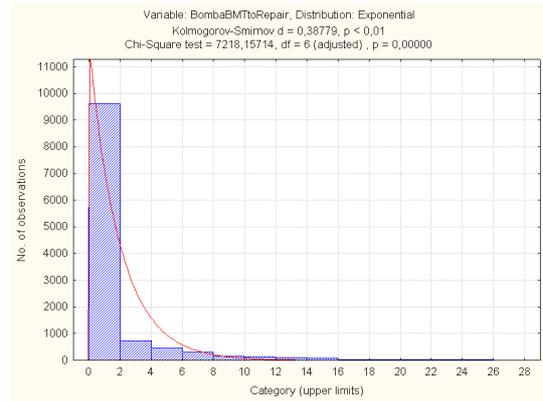


(b) Distribuição do TTR da bomba BCP

Figura 4.13: Distribuição de probabilidade para o TBF e o TTR da bomba BCP

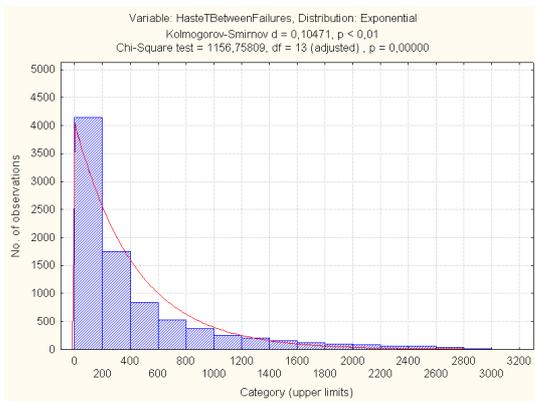


(a) Distribuição do TBF da bomba BM

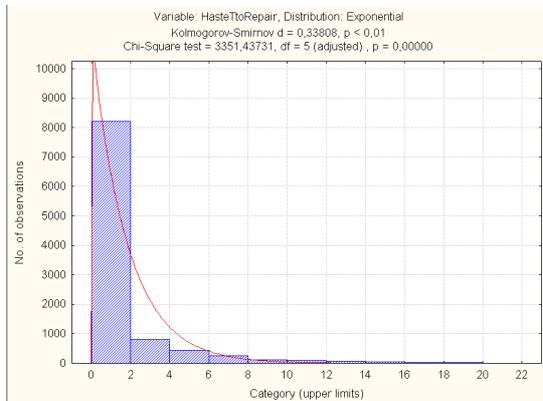


(b) Distribuição do TTR da bomba BM

Figura 4.14: Distribuição de probabilidade para o TBF e o TTR da bomba BM

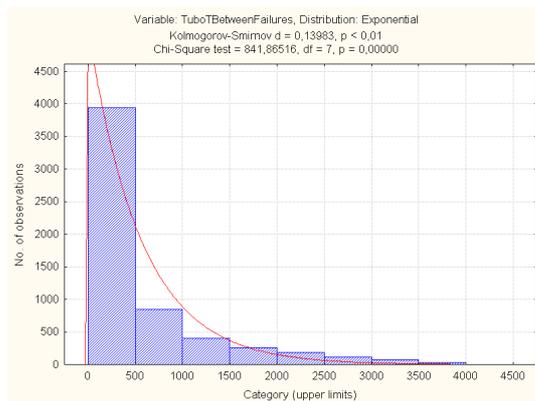


(a) Distribuição do TBF da haste

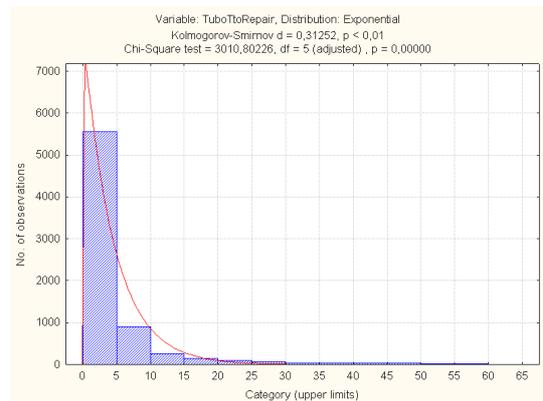


(b) Distribuição do TTR da haste

Figura 4.15: Distribuição de probabilidade para o TBF e o TTR da haste



(a) Distribuição do TBF do tubo



(b) Distribuição do TTR do tubo

Figura 4.16: Distribuição de probabilidade para o TBF e o TTR do tubo

4.4.6 Análise de Disponibilidade com a Cadeia de Markov

Com a cadeia de Markov estruturada e definidas as distribuições dos tempos, pode-se calcular a disponibilidade do sistema ao longo do tempo e os seus limites de confiança. A disponibilidade é calculada utilizando os algoritmos implementados no capítulo 3 e os intervalos são estimados com múltiplas execuções dos algoritmos 1 e 2 para diferentes sementes da função *Random*. Os resultados do comportamento da disponibilidade dos sistemas são mostrados nas Figuras 4.17 e 4.18. O comportamento diferenciado para o primeiro período da Figura 4.17, em relação aos dois algoritmos, é decorrente de uma variação amostral no momento da geração dos números aleatórios.

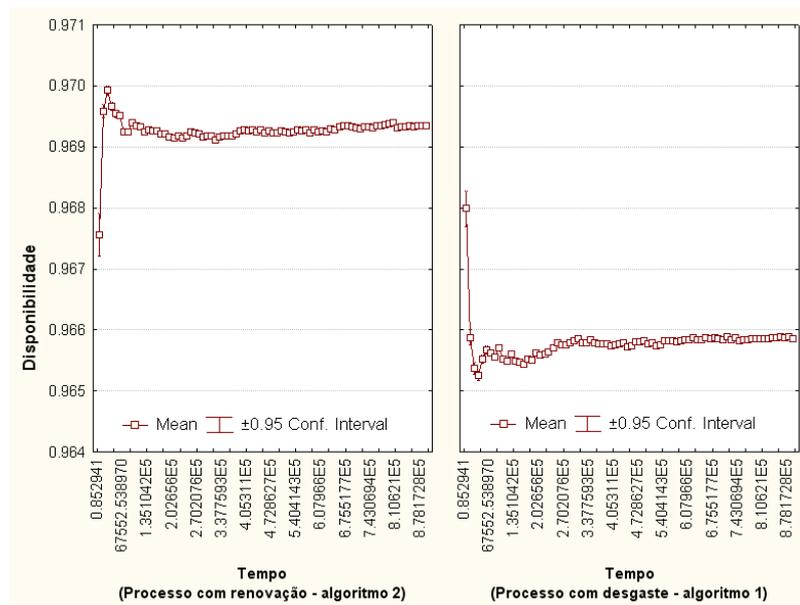


Figura 4.17: Disponibilidade para o sistema poço com método de elevação centrífugo

Pode-se observar nessas figuras que o tempo necessário de vida para o sistema atingir o estado estacionário é relativamente elevado (visualmente, algo em torno de $2,5 \cdot 10^5$ dias para o sistema BCP e $5 \cdot 10^5$ dias para o sistema BM), fugindo da realidade operacional de funcionamento de um poço. Nesse caso, é imprescindível que se analise os sistemas dentro de um período mais curto, considerando apenas seus estados transientes. Para tanto, será considerado apenas um prazo de vida útil médio de um poço estimado por especialista: 20 anos.

Observando esse período de transiência, nota-se visualmente por meio das Figuras 4.19 e 4.20 que a disponibilidade dos sistemas utilizando os Algoritmos 1 e 2 confundem-se. Além disso, é observável que o nível de disponibilidade encontra-se em patamar muito

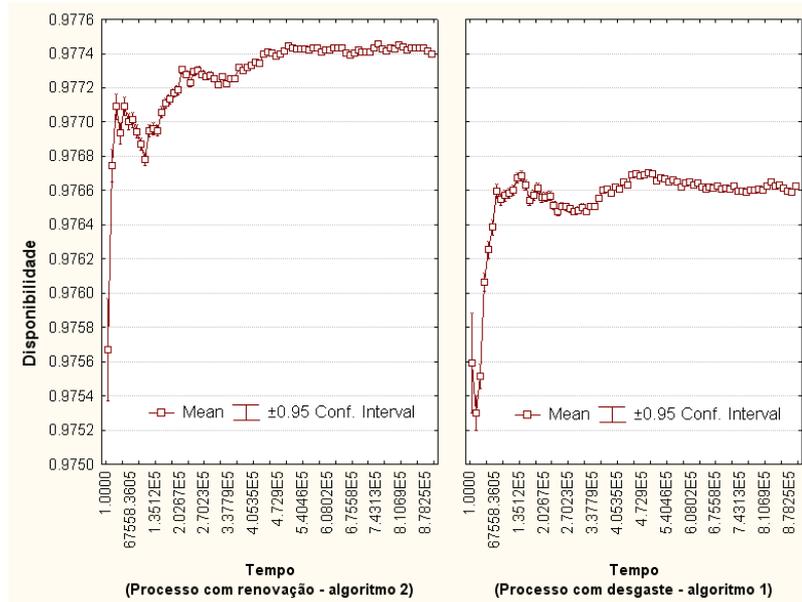


Figura 4.18: Disponibilidade para o sistema poço com método de elevação mecânico

elevado. Este fato ocorre devido ao fato do tempo entre falhas encontrar-se em patamar muito superior ao do tempo de reparo, assim o tempo de permanência em estados indisponíveis do sistema é desprezível se comparado ao tempo de permanência no estado operacional, o que eleva a disponibilidade do poço.

A afirmação anteriormente realizada quanto ao alto nível de disponibilidade não seria verdadeira se em alguns instantes da simulação não houvesse renovação dos componentes do sistema, o que é realizado no passo 3(a) do Algoritmo 1 e no Algoritmo 2.

Para se ter uma idéia da disponibilidade se não houvesse nenhuma política de renovação dos equipamentos, considerando apenas reparos mínimos, a disponibilidade dos sistemas BCP e BM utilizando o Algoritmo 1 sem o passo 3(a) se comporta como o exposto nas Figuras 4.21 e 4.22. É importante visualizar esse tipo de comportamento de modo a ter-se uma medida do nível de degradação dos sistemas.

Além da disponibilidade dos sistemas, como já visto no capítulo anterior, pode-se obter algumas outras métricas úteis na análise do sistema, ver Tabelas 4.9 e 4.10.

Os resultados dos indicadores apresentados nessas tabelas trazem informações de priorização na correção dos problemas. Como exemplo, pode-se atuar prioritariamente no componente falho que caracteriza o nó 3 da Tabela 4.9 (bomba BCP), uma vez que há um superior número de visitas a este estado, quando comparados com os outros dois estados indisponíveis do sistema, e também um elevado tempo médio de manutenção para esse

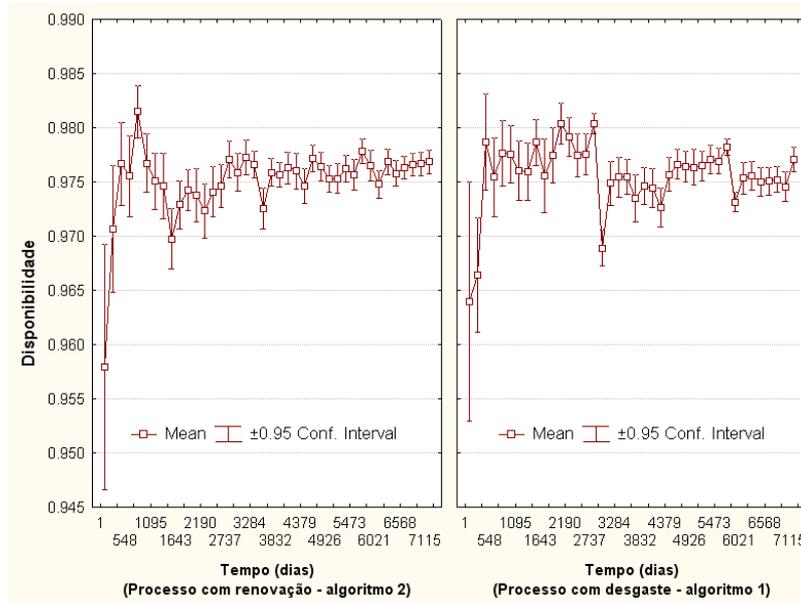


Figura 4.19: Disponibilidade para o sistema poço com método de elevação mecânico (20 anos)

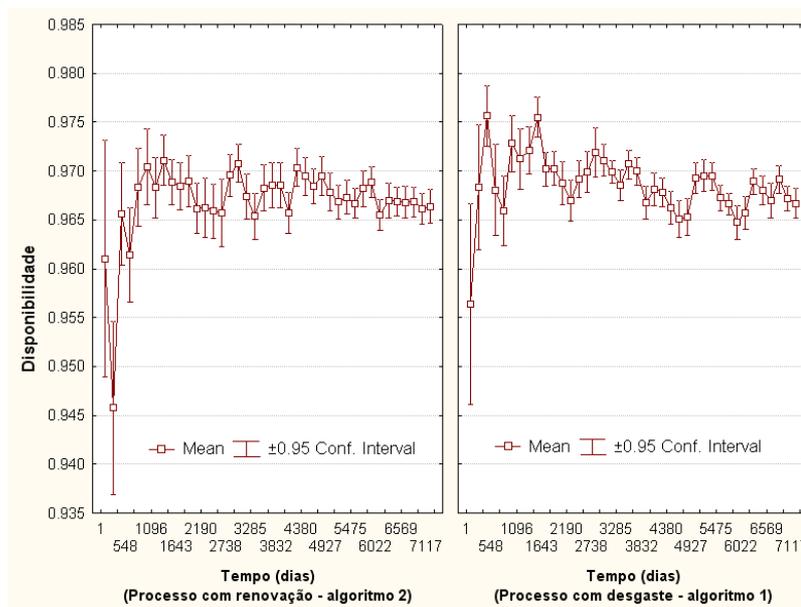


Figura 4.20: Disponibilidade para o sistema poço com método de elevação centrífugo (20 anos)

tipo de equipamento (aproximadamente 62% superior ao segundo pior tempo médio de reparo – 0,0176/0,0109).

Com isso, de uma forma geral, pode-se sugerir um aprofundamento no estudo na forma como as bombas estão operando e no procedimento de manutenção das mesmas. E, com os resultados da disponibilidades, elaborar diversos cenários para suportar à decisão econômica de continuar produzindo ou não ou de realizar novos investimentos.

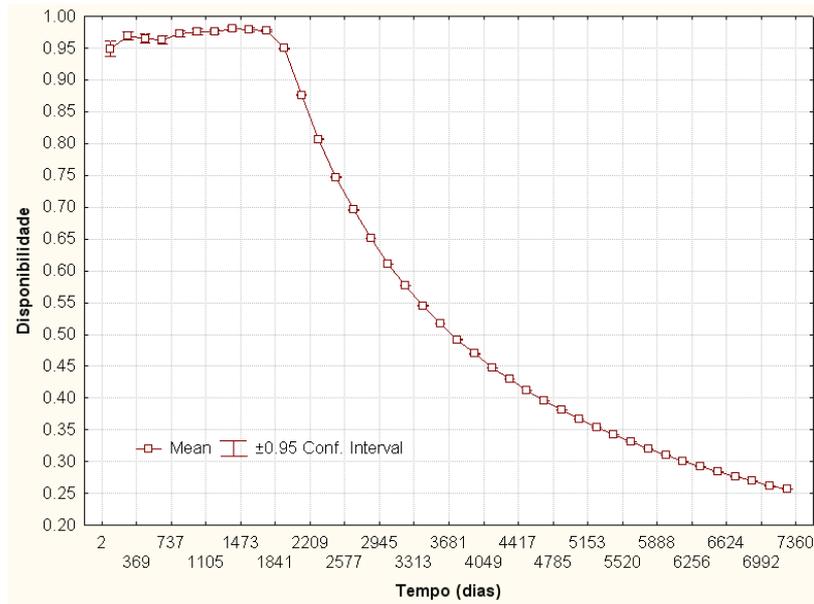


Figura 4.21: Disponibilidade para o sistema poço com método de elevação mecânico completamente sem renovação (20 anos)

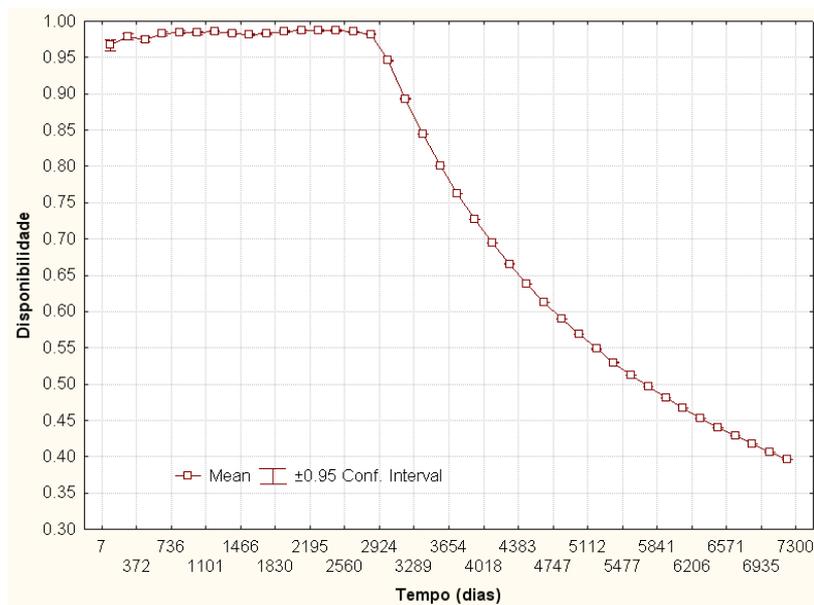


Figura 4.22: Disponibilidade para o sistema poço com método de elevação centrífugo completamente sem renovação (20 anos)

4.4.7 Atualização do Estudo de Disponibilidade via Evidências nas Variáveis das Redes Bayesianas dos Equipamentos do Sistema

Para realizar a atualização da disponibilidade do sistema, evidências nos nós das BBN's são imputadas e estima-se a distribuição de probabilidade da variável alvo de cada BBN,

Tabela 4.9: Indicadores para a análise de disponibilidade do sistema BCP

Algoritmo 1			Algoritmo 2		
Nó	Número médio de visitas	Fração média do tempo total de permanência	Nó	Número médio de visitas	Fração média do tempo total de permanência
0	8.097	0,9659	0	7.445	0,9694
1	2.159	0,0109	1	2.158	0,0108
2	2.695	0,0057	2	2.672	0,0057
3	3.241	0,0176	3	2.613	0,0141

Tabela 4.10: Indicadores para a análise de disponibilidade do sistema BM

Algoritmo 1			Algoritmo 2		
Nó	Número médio de visitas	Fração média do tempo total de permanência	Nó	Número médio de visitas	Fração média do tempo total de permanência
0	7.755	0,9766	0	7.433	0,9774
1	2.144	0,0108	1	2.139	0,0107
2	2.657	0,0057	2	2.646	0,0057
3	2.953	0,0069	3	2.646	0,0062

que no presente caso são os tempos entre falhas e de reparo. Assim, para cada BBN de cada equipamento (falha e reparo) é evidenciada a variável BSWSOT (percentual de H₂O e sólidos no poço) no seu nível mais alto (> 94,4%) e também o tipo de método de elevação, quando possível.

Como exemplo, é mostrada a atualização da distribuição da bomba BCP dada a evidência no nó BSWSOT (ver Figuras 4.23 e 4.24). Como esperado, uma alta evidência no nível de estruturas sólidas no interior do poço faz o MTBF reduzir. E, o mesmo é constatado para os outros equipamentos do sistema BCP e o comportamento inverso para o tempo médio até a falha, conforme o esperado.

Para verificar o impacto dessa atualização na disponibilidade do sistema de subsuperfície do poços, é evidenciada na Figura 4.25 a disponibilidade do sistema até o período estacionário com o método de elevação BCP. Percebe-se que há uma redução na disponibilidade do sistema, como era esperado. Como a atualização serve apenas para exemplificar a metodologia proposta, acredita-se que o objetivo de mostrar a variação na disponibilidade do sistema dada a evidência é atingido. Ao passo que novas evidências forem sendo incorporadas às BBN's, novos comportamentos para a disponibilidade serão observados.

O mesmo é feito para o sistema com método de elevação mecânico, tendo obtido o

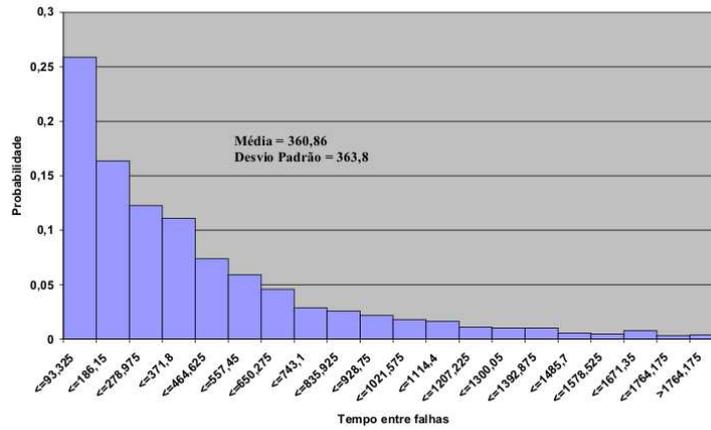


Figura 4.23: Distribuição de probabilidade do TBF da bomba BCP antes da evidência

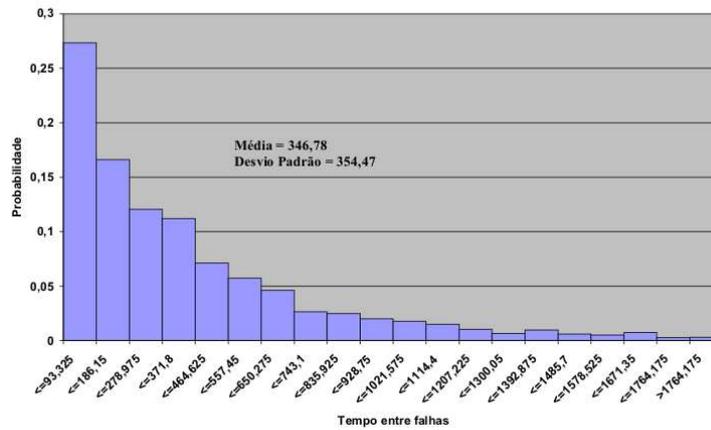


Figura 4.24: Distribuição de probabilidade do TBF da bomba BCP após evidência

mesmo comportamento esperado, como mostrado na Figura 4.26.

Essa fase da aplicação da metodologia é importante, pois permite obter estimativas atualizadas da disponibilidade do sistema de subsuperfície do poço à medida que dados sobre as atuais condições de operação e de seus componentes são obtidos. Essa é uma ferramenta para ser usada essencialmente no auxílio da análise da operação e produção do dia-a-dia de um poço.

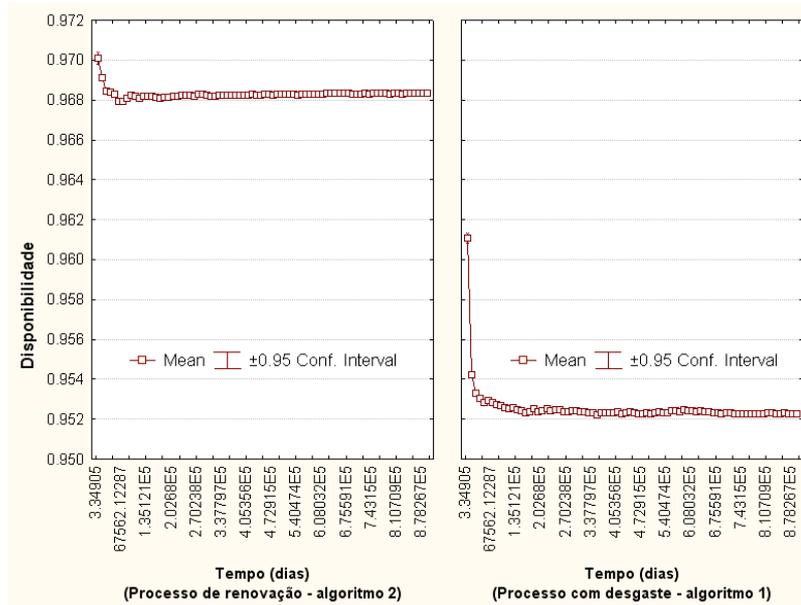


Figura 4.25: Disponibilidade para o sistema poço com método de elevação centrífugo após evidência

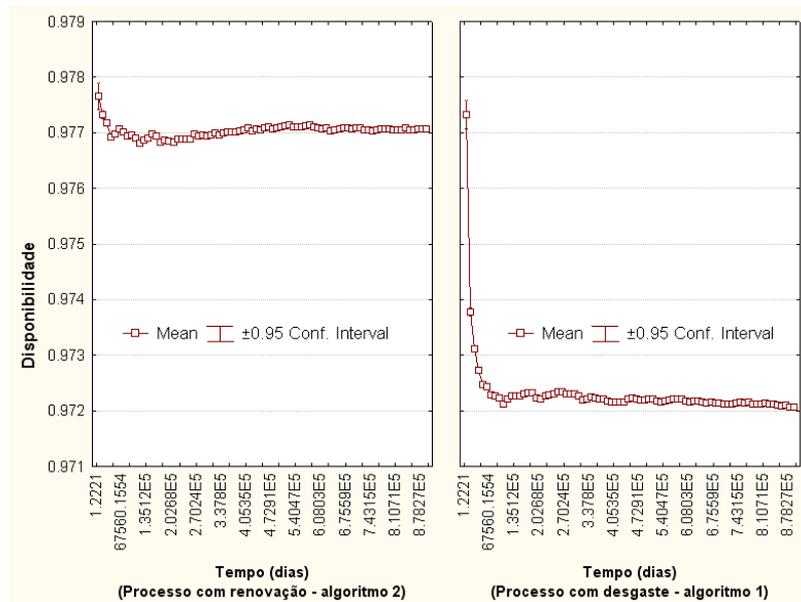


Figura 4.26: Disponibilidade para o sistema poço com método de elevação mecânica após evidência

5 CONCLUSÕES E TRABALHOS FUTUROS

Na primeira seção deste capítulo será realizada uma sucinta reflexão sobre os principais resultados deste trabalho e numa seção posterior, propostas de futuras pesquisas serão sugeridas.

5.1 Conclusões

Para melhor entendimento da opinião do autor sobre os assuntos aqui expostos, será realizada uma discussão por grandes temas abordados nesta dissertação, iniciando pelo aprendizado das redes bayesianas. Mas, antes, será realizado um confronto entre os objetivos traçados e o realizado no trabalho.

Quanto ao objetivo de realizar uma revisão literária de extração de conhecimento de base de dados via aprendizagem das redes bayesianas, foi realizada uma revisão bibliográfica no capítulo 1 que terá os principais pontos discutidos na seção seguinte.

Em relação à revisão dos conceitos utilizados na análise de disponibilidade, foram, no capítulo 2.1, mostrados os conceitos referentes aos processos estocásticos markovianos que são necessários para o entendimento da avaliação de disponibilidade de sistemas. Além disso, as principais distribuições de probabilidade utilizadas para modelar os tempos de transição da cadeia de Markov foram mostradas. Assim, teve-se uma idéia de como seria a utilização de um modelo paramétrico para a análise de disponibilidade de sistemas.

A forma de interação entre a cadeia de Markov e as redes bayesianas foi mostrada no capítulo 3, a qual ocorre por meio da distribuição de probabilidade não-paramétrica dos tempos entre falhas e de reparo. Adicionalmente, os passos metodológicos discutidos no mesmo capítulo, além de viabilizar a implementação em casos de interesse prático, esclarecem como a atualização das distribuições dos tempos provocada por evidências nas variáveis monitoradas ocorre. Mais detalhes sobre os principais pontos da metodologia serão discutidos nas duas seções após o aprendizado das redes bayesianas.

Comentários sobre os algoritmos elaborados no capítulo 3 serão realizados na seção 5.1.2.

Já no capítulo anterior, foi realizado um exemplo de aplicação com o intuito de atingir o último objetivo anunciado. Críticas e comentários à base de dados utilizada no exemplo de aplicação serão realizados na seção 5.1.4.

5.1.1 O Aprendizado das Redes Bayesianas

Como já detalhado, o aprendizado das redes bayesianas se dá em duas etapas: o aprendizado da estrutura e o aprendizado dos parâmetros da rede. Quanto a este, acredita-se que pela simplicidade conceitual não restam grandes dúvidas. O ponto principal é quanto ao primeiro tipo.

No aprendizado automático da topologia das redes bayesianas, no momento da revisão bibliográfica, observou-se uma grande quantidade de algoritmos, cada um com as suas peculiaridades na realização desse feito, mas o que se observou foi a falta de uma teoria, calcada em teoremas e demonstrações a esse respeito. Apesar da impossibilidade, pelo menos atual, de se determinar via dados a relação de causalidade entre variáveis, a idéia do uso das propriedades de dependência e independência das redes bayesianas na definição das redes markovianas equivalentes, para o autor, tem se mostrado uma boa maneira de representar um modelo inicial de causalidade, uma vez que para se retirar essa informação do especialista levaria muito tempo. Além disso, a quantidade excessiva de possíveis relações poderiam levar a construção de um errôneo modelo por parte do especialista.

É importante destacar que o que se busca na estruturação automática da topologia das redes bayesianas são as relações de dependência e independência “relevantes” e uma representação primeira das relações causais das variáveis monitoradas da base de dados, para uma posterior confirmação e correção do especialista. A importância disso, é que o especialista apenas focará sua análise nas relações que apresentarem relações significativas e nas que, por algum viés da base de dados, não foram explicitadas.

5.1.2 Algoritmos de Estimação da disponibilidade de Sistemas

Outro importante ponto, é quanto aos algoritmos dos modelos probabilísticos dinâmicos utilizados na análise de disponibilidade de sistemas.

Pontos a Considerar

Com os algoritmos propostos, espera-se facilitar e deixar a análise de disponibilidade mais robusta, sem a necessidade de realizar parametrizações nas distribuições do tempo. Esse é um ponto bastante importante, pois possibilita aos menos afinados com os testes de hipóteses, apenas verificando se o sistema encontra-se em processo de desgaste, realizar

uma análise de disponibilidade apurada.

Uma outra importante consideração a fazer é quanto aos passos 5 e 4 dos Algoritmos 1 e 2, respectivamente. Quando se faz $U_j = \infty \forall S_i \in P_A$ em que $E_A(c_j) = E_{S_i}(c_j)$, está se possibilitando trabalhar com tempos de ordens de grandezas diferentes. Um exemplo disso seria se o tempo entre falhas de um equipamento que caracteriza um nó estivesse na ordem de grandeza 10^3 e o tempo de reparo 10^1 . Se uma comparação fosse realizada para ver se o sistema iria para um estado indisponível ou continuaria no estado disponível, muito provavelmente ele iria continuar no estado disponível. Assuma agora que não há a possibilidade topológica do sistema estar no estado disponível e continuar no estado disponível. Se a opção $U_j = \infty \forall S_i \in P_A$ em que $E_A(c_j) = E_{S_i}(c_j)$ não fosse considerada, demoraria muito para o sistema ir para o estado indisponível, o que seria computacionalmente impraticável a simulação desse tipo de sistema. Assim, com a construção desses passos, os tempos para estados impossíveis (os que não foram definidos na cadeia de Markov) recebem infinito, tornando a simulação mais rápida.

Pontos a Ponderar

É verdade que, por outro lado, com a simulação discreta de eventos, perde-se em velocidade de processamento, uma vez que o tempo para convergência em média das probabilidades estacionárias é mais elevado. Além disso, a necessidade de avaliar intervalos de confiança para a métrica desejada, pelo menos na fase transiente, eleva ainda mais esse tempo.

5.1.3 A Metodologia Proposta

Considerando-se os passos metodológicos mostrados no capítulo 3, tem-se alguns aspectos que merecem uma reflexão.

Pontos a Considerar

De um forma geral, a metodologia sugerida, que faz uso de um modelo híbrido constituído pelos processos de Markov e pelas redes bayesianas, possibilitou analisar um sistema real de interesse prático de uma forma dinâmica. Nesse modelo, as mudanças nos parâmetros são realizadas via evidências durante a operação e as variações nas distribuições do

tempos de transição da cadeia de Markov são realizadas por meio da propagação dessas evidências ao longo das redes bayesianas até as variáveis alvo. A percepção dessas variações por parte do modelo, por sua vez, são utilizados para obter, através do processo de Markov, uma estimativa atualizada da disponibilidade utilizando os algoritmos propostos para o cálculo das probabilidades de estado.

Se uma hierarquização dos benefícios trazidos pela metodologia proposta for realizada a possibilidade de avaliar em tempo real a disponibilidade do sistema será a mais notória e colocada como a mais importante. Isso se deve, pois, de uma forma mista, ou por opinião de especialistas e/ou por intermédio de ocorrências registradas na base de dados, a disponibilidade do sistema pode ser atualizada de uma maneira notadamente simples, apenas utilizando a regra de Bayes repetidas vezes.

Além disso, buscou-se explorar ao máximo as informações possíveis de serem extraídas da base de dados, deixando para o especialista a parte que realmente necessitasse da percepção e cognição humana.

Pontos a Ponderar

Como todo modelo, simplificações realizadas impossibilitam a percepção de sutilezas que ocorrem no “mundo real”. Especificamente nesse caso, tais sutilezas dizem respeito principalmente a incapacidade do modelo de estar captando, em tempo de simulação, variações de tecnologia ou de *modus operandi* que levem a uma mudança na topologia das relações causais das variáveis monitoradas. Ou seja, se houver alguma variação mais significativa no sistema que de algum modo altere o nível de influência das relações inter-variáveis, o modelo será insensível a este tipo de modificação, pois é assumido que a topologia da rede bayesiana é invariável ao longo do tempo.

O que se propõe para solucionar o posto no parágrafo anterior, é que em vez da simples utilização do passo 7 da metodologia proposta, onde se fala do processo de atualização da análise de disponibilidade via evidências nas redes bayesianas, os passos anteriores da metodologia, iniciando da etapa 3, sejam refeitos. Assim, em vez de se ter apenas a atualização das redes bayesianas, deve-se realizar novamente a estimação da topologia das redes bayesianas e seus respectivos parâmetros.

Outro ponto bastante custoso para o modelo, é a necessidade do registro do histórico do sistema em análise numa base dados relativamente grande. É assim, pois para estimação

da topologia das redes bayesianas é necessária uma grande quantidade de dados relacionados para a entrada nos algoritmos de aprendizado de topologia das redes bayesianas.

5.1.4 O Estudo de Caso

Pontos a Considerar

Observou-se no setor de produção de petróleo um promissor campo de aplicação para a metodologia proposta. As paradas para manutenção nesse setor apresentam enormes perdas e o monitoramento das variáveis de operação de um poço são realizadas rotineiramente.

Essa prática de monitoramento gera registros de variáveis em banco de dados que são de vital importância para a operação de um poço, sendo todas elas armazenadas de forma relacional (os valores de todas as variáveis são registrados a cada ocorrência), viabilizando o exemplo de aplicação.

Apesar desses pontos a considerar, foi observada uma inadequada medição e registro dos dados que servem de suporte para a análise de confiabilidade e disponibilidade dos equipamentos, o que será discutido a seguir.

Pontos a Ponderar

A maneira como são dispostos os dados de manutenção apresentam agregações que, do ponto de vista da análise de confiabilidade e disponibilidade, levam o analista a realizar análises mais genéricas. Como exemplo disso, tem-se os tempos de reparo, em que não é diferenciado o tempo de reparo ativo (tempo efetivamente gasto para realizar o reparo do equipamento) do reparo inativo (tempo gasto nas operações precedentes ao reparo em si, o qual pode ser exemplificado como o tempo de espera pela sonda desde a sua solicitação).

Além desse ponto, não é realizada uma distinção clara entre as paradas realizadas devido à quebra dos equipamentos e as paradas para estimulação do poço (procedimentos realizados para elevar a vazão do poço).

Além de tudo isso, há uma confusão quanto ao conceito de mecanismo de falha, modo de falha e efeito de falha. A base de dados utilizada registra o mecanismo de falha (nomeado como “causa”) e o efeito de falha (nomeado como “modo”), deixando de lado o registro do modo de falha. Esse fato impossibilita a realização de uma análise mais

realista conforme os trabalhos publicados nessa área, os quais consideram o tempo entre falhas segregados por modo de falha.

Tendo em vista essas dificuldades para implementar a metodologia devido à inadequação da base de dados, sugerem-se algumas modificações:

- Além do campo de data de solicitação da sonda, acrescentar outro campo para registro da data de chegada da sonda no poço para avaliar a logística de alocação do recurso sonda. Além disso, adicionar outro campo referente a data de início da operação de manutenção em si, para se avaliar o tempo de preparação da sonda para o início da operação de manutenção;
- Esclarecer o campo TIPO, deixando claro se houve substituição dos equipamentos no momento da restauração do poço;
- Reavaliar os conceitos de mecanismos de falha, modo de falha e efeito de falha para preencher os campos MODO e CAUSA da base de dados. Para mais detalhes, ver Rausand e Oien (1996). É importante destacar que o conceito modo de falha deve ficar bem claro e registrado na base de dados, pois as falhas devem ser analisadas mais detalhadamente através dos modos de falha.
- Outro ponto que merece destaque e que foi mencionado rapidamente nos capítulos 3 e 4 é o conceito de *socket*. É interessante avaliar a viabilidade econômica de se construir uma base de dados utilizando o histórico das bombas em si e não do *socket* onde o equipamento é instalado no poço. Tecnicamente esse detalhamento é bastante necessário, pois se sabe que as bombas após reparos podem ser realocadas para outros poços menos produtivos e, além disso, deseja-se analisar mais precisamente o comportamento das bombas, visto que, por elas estarem em série com outros equipamentos, e se falharem, o poço se torna indisponível, são de vital importância para o funcionamento de um poço.

5.2 Sugestões para Trabalhos Futuros

Como sugestão para trabalhos futuros, propõe-se alguns avanços nesta pesquisa que são listados a seguir:

- Elaboração de uma metodologia de estruturação de base de dados de confiabilidade, a fim de se obter variáveis significativas para o estudo da análise de disponibilidade. Levar em consideração, por exemplo: a distinção entre tempo de reparo ativo e inativo de um equipamento, a clareza na distinção do motivo da parada de um equipamento, separando os casos de intervenção devido à falha e devido à estimulação, no caso de um poço, entre outros.
- A aplicação da metodologia desenvolvida neste trabalho em outros tipos de sistemas que se utilizam de base de dados para acompanhamento de seus comportamentos e que têm o tempo como variável alvo monitorada, por exemplo, o mercado de ações;
- Adicionar à metodologia critérios para se avaliar a necessidade de reconfiguração da topologia das redes bayesianas após evidências nas variáveis monitoradas;
- Verificar a possibilidade de adaptação nos algoritmos propostos para absorver a idéia do modelo semi-markoviano exposta Becker, Camarinopoulos e Zioutas (2000) que considera o tempo de permanência no presente estado como um fator influenciante nas probabilidades de transição;
- Adaptação nos algoritmos propostos para incluir a idéia dos modelos de renovação generalizados mostrada em Yanez, Joglar e Modarres (2002) em que os equipamentos após o reparo não só retornam à operação tão bons quanto novos ou tão ruins quanto velhos, mas além dessas possibilidades outras três situações tornam-se possíveis: melhor do que velho mas pior do que novo, melhor do que novo ou pior do que velho. O que se pretende com essa sugestão é a inserção de uma medida de eficácia das equipes de manutenção nos modelos.

REFERÊNCIAS BIBLIOGRÁFICAS

ASCHER, H.; FEINGOLD, H. *Repairable systems reliability modeling, inference misconceptions and their causes*. New York, USA: Marcel Dekker, 1984.

BAI, C. et al. Software failure prediction based on a markov bayesian network model. *The Journal of Systems and Software*, n. 74, p. 275–282, 2005.

BECKER, G.; CAMARINOPOULOS, L.; ZIOUTAS, G. A semi-markovian model allowing for inhomogenities with respect to process time. *Reliability Engineering and system safety*, v. 70, p. 41–48, 2000.

BUKOWSKI, J. V. Using markov models to compute probability of failed dangerous when repair times are not exponentially distributed. In: *Proceedings of the Annual Reliability and Maintainability Symposium*. [S.l.: s.n.], 2006.

CHENG, J. et al. Learning bayesian networks from data: An information–theory based approach. *Artificial Intelligence*, v. 137, p. 43–90, 2002.

CHOW, C. K.; LIU, C. N. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, v. 14, n. 3, p. 462–467, 1968.

COCHRAN, J. K.; MURUGAN, A.; KRISHNAMURTHY, V. Generic markov models for availability estimation and failure characterization in petroleum refineries. *Computers & Operations Research*, v. 28, p. 1–12, 2001.

COWELL, R. When learning bayesian networks from data, using conditional independence tests is equivalent to a local scoring metric. In: *Proc. 17th International Conference on Uncertainty in Artificial Intelligence*. Seattle, WA: UAI, 2001.

- DAMPSTER, A. P.; LAIRD, M. N.; RUBIN, D. B. Maximum likelihood from incomplete data via an EM algorithm. *Journal of the Royal Statistical Society (Series B)*, v. 39, n. 1, p. 1–38, 1977.
- DEGROOT, M. H.; SCHERVISH, M. J. *Probability and Statistics*. 3. ed. USA: Addison Wesley, 2002.
- DROGUETT, E. L. *Modelos de Markov em Confiabilidade*. Novembro 2002. Apostila da disciplina Engenharia de Confiabilidade do Departamento de Engenharia de Produção. Disponível em: <<http://www.risctec.org>>.
- DROGUETT, E. L. et al. Probability assessment of offshore oil multilateral wells construction process in Brazil. In: *International Conference on Probabilistic Safety Assessment and Management*. New Orleans, USA: Probabilistic Safety Assessment and Management, 2006.
- GLOBO, O. *Petróleo fecha em forte queda diante de atentados frustrados*. 2006. Website. Disponível em: <<http://oglobo.globo.com/petroleo/materias/2006/08/10/285216785.asp>>. Acesso em: 10 ago. 2006.
- GRUNWALD, P. A tutorial introduction to the minimum description length principle. In: GRUNWALD, P.; MYUNG, I.; PITT, M. (Ed.). *Advances in Minimum Description Length: Theory and Applications*. [S.l.]: MIT Press, 2005.
- HAHN, G. J.; SHAPIRO, S. S. *Statistical models in engineering*. New York: Wiley, 1967.
- HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. San Francisco, USA: Morgan Kaufmann Publishers, 2001.
- HECKERMAN, D.; GEIGER, D.; CHICKERING, D. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, v. 20, n. 3, p. 197–243, 1995.
- HOUAISS, I. A. *Dicionário Houaiss da Língua Portuguesa*. Brasil: Editora Objetiva Ltda, 2001. Versão 1.0.

- HOWARD, R. A. *Dynamic probabilistic systems*. New York, USA: Wiley, 1971. (Series in decision and control, v. 1).
- HRUDCHKA, E. R. J. *Imputação Bayesiana no Contexto da Mineração de Dados*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, Outubro 2003.
- JOUFFE, L. Nouvelle classe de méthodes d'apprentissage de réseaux bayésiens. *Journées francophones d'Extraction et de Gestion des Connaissances (EGC)*, janvier 2002.
- KORB, K. B.; NICHOLSON, A. E. *Bayesian Artificial Intelligence*. New York, USA: Chapman & Hall / CRC, 2004. (Series in Computer Science and Data Analysis).
- LAFRAIA, J. R. B. *Manual de confiabilidade, manutenibilidade e disponibilidade*. Rio de Janeiro: Qualitymark: Petrobras, 2001.
- LAI, C. et al. A model for availability analysis of distributed software/hardware systems. *Information and Software Technology*, n. 44, p. 343–350, 2002.
- LAM, W.; BACCHUS, F. Learning bayesian belief networks: An approach based on MDL principle. *Computational Intelligence*, v. 10, n. 4, p. 269–293, 1994.
- LEE, S. M.; ABBOTT, P. A. Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers. *Journal of Biomedical Informatics*, v. 36, p. 389–399, 2003.
- LEWIS, E. E. *Introduction to Reliability Engineering*. New York: John Wiley & Sons, 1987.
- MENEZES, R. da C. S. *Uma metodologia para avaliação da confiabilidade humana em atividades de substituição de cadeias de isoladores em linhas de transmissão*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, Recife, Junho 2005.
- MILANO, L. *Real-time sensor data validation*. France, April 1994.
- MILANO, L. *Inférence bayésienne à partir d'une loi de Weibull*. France, 1995.

- MUNTEANU, P.; BENDOU, M. The EQ framework for learning equivalence classes of bayesian networks. *IEEE International Conference on Data Mining (IEEE ICDM)*, 2001.
- NETO, A. A. *Finanças Corporativas e Valor*. São Paulo, Brasil: Editora Atlas, 2003.
- OUHBI, B.; LIMNIOS, N. Reliability estimation of semi-markov systems: a case study. *Reliability Engineering and System Safety*, n. 58, p. 201–204, 1997.
- PEARL, J. *Causality: Models, Reasoning, and Inference*. UK: Cambridge University, 2000.
- PETROBRAS. *Produção de Óleo, LGN e Gás Natural*. 2006. Website. Disponível em: <http://www2.petrobras.com.br/ri/port/DestaquesOperacionais/ExploracaoProducao/pdf/ProducaoOleoGas_E&P_jun2006_Port.pdf>. Acesso em: 10 ago. 2006.
- RAUSAND, M.; OIEN, K. The basic concepts of failures analysis. *Reliability Engineering and System Safety*, v. 53, p. 73–83, 1996.
- RISSANEN, J. Modeling by shortest data description. *Automatica*, v. 14, n. 5, p. 465–471, September 1978.
- ROBINSON, R. W. Counting unlabeled acyclic digraphs. In: LITTLE, C. H. C. (Ed.). *Combinatorial Mathematics V*. Australia: Springer-Verlag, 1977. (Lecture Notes in Mathematics, v. 622), p. 239–273.
- ROSS, S. M. *Simulation*. 3. ed. USA: Academic Press, 2002.
- RUBINSTEIN, R. Y. *Simulation and the Monte Carlo method*. USA: John Wiley & Sons, 1981.
- SIGURDSSON, J.; WALLS, L.; QUIGLEY, J. Bayesian belief nets for managing expert judgment and modeling reliability. *Quality and Reliability Engineering International*, n. 17, p. 181–190, 2001.
- SILVA, A. R. da. *Análise da Disponibilidade de Malhas de Completação Inteligente em Poços de Petróleo*. Dissertação (Mestrado) — Universidade Federal de Pernambuco – PPGEP, 2004.

- SILVA, W. T. da; LADEIRA, M. Mineração de dados em redes bayesianas. In: PORTO, I. J. (Ed.). *Anais do XXII Congresso Brasileiro de Computação*. Florianópolis, Brasil: SBC, 2002. v. 2, p. 235–286.
- SMYTH, P. Belief networks, hidden markov models, and markov random fields: A unifying view. *Pattern Recognition Letters*, n. 18, p. 1261–1268, 2005.
- THOMAS, J. E. *Fundamentos de Engenharia do Petróleo*. Rio de Janeiro, Brasil: PETROBRAS, 2000.
- VERMA, T. S.; PEARL, J. Equivalence and synthesis of causal models. In: *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*. San Mateo, CA, USA: Morgan Kaufmann, 1990. p. 220–227.
- WACKERLY, D. D.; MENDENHALL, W.; SCHEAFFER, R. L. *Mathematical Statistics with Applications*. 5th. ed. USA: Duxbury Press, 1996.
- WALLACE, C. S.; BOULTON, D. M. An information measure for classification. *the Computer Journal*, v. 11, p. 185–194, 1968.
- WEBERA, P.; JOUFFE, L. Complex system reliability modelling with dynamic object oriented bayesian networks (doobn). *Reliability Engineering and System Safety*, n. 91, p. 149–162, 2006.
- YANEZ, M.; JOGLAR, F.; MODARRES, M. Generalized renewal process for analysis of repairable systems with limited failure experience. *Reliability Engineering and System Safety*, v. 77, p. 167–180, 2002.

Livros Grátis

(<http://www.livrosgratis.com.br>)

Milhares de Livros para Download:

[Baixar livros de Administração](#)

[Baixar livros de Agronomia](#)

[Baixar livros de Arquitetura](#)

[Baixar livros de Artes](#)

[Baixar livros de Astronomia](#)

[Baixar livros de Biologia Geral](#)

[Baixar livros de Ciência da Computação](#)

[Baixar livros de Ciência da Informação](#)

[Baixar livros de Ciência Política](#)

[Baixar livros de Ciências da Saúde](#)

[Baixar livros de Comunicação](#)

[Baixar livros do Conselho Nacional de Educação - CNE](#)

[Baixar livros de Defesa civil](#)

[Baixar livros de Direito](#)

[Baixar livros de Direitos humanos](#)

[Baixar livros de Economia](#)

[Baixar livros de Economia Doméstica](#)

[Baixar livros de Educação](#)

[Baixar livros de Educação - Trânsito](#)

[Baixar livros de Educação Física](#)

[Baixar livros de Engenharia Aeroespacial](#)

[Baixar livros de Farmácia](#)

[Baixar livros de Filosofia](#)

[Baixar livros de Física](#)

[Baixar livros de Geociências](#)

[Baixar livros de Geografia](#)

[Baixar livros de História](#)

[Baixar livros de Línguas](#)

[Baixar livros de Literatura](#)
[Baixar livros de Literatura de Cordel](#)
[Baixar livros de Literatura Infantil](#)
[Baixar livros de Matemática](#)
[Baixar livros de Medicina](#)
[Baixar livros de Medicina Veterinária](#)
[Baixar livros de Meio Ambiente](#)
[Baixar livros de Meteorologia](#)
[Baixar Monografias e TCC](#)
[Baixar livros Multidisciplinar](#)
[Baixar livros de Música](#)
[Baixar livros de Psicologia](#)
[Baixar livros de Química](#)
[Baixar livros de Saúde Coletiva](#)
[Baixar livros de Serviço Social](#)
[Baixar livros de Sociologia](#)
[Baixar livros de Teologia](#)
[Baixar livros de Trabalho](#)
[Baixar livros de Turismo](#)