Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Investigação e Avaliação Experimental de Técnicas de Re-teste Seletivo para Teste de Regressão baseado em Especificação

Francisco Gomes de Oliveira Neto

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linhas de Pesquisa: Engenharia de Software

Prof[a]. PhD. Patrícia Duarte de Lima Machado
(Orientadora)

Campina Grande – Paraíba – Brasil

O48i Oliveira Neto, Francisco Gomes de
Investigacao e avaliacao experimental de tecnicas de re-teste seletivo para teste de regressao baseado em especificacao / Francisco Gomes de Oliveira Neto. - Campina Grande, 2011.
220 f. : il.

Dissertacao (Mestrado em Ciencia da Computacao) - Universidade Federal de Campina Grande, Centro de Engenharia Eletrica e Informatica.

1. Teste Caixa-Preta 2. Teste de Regressao 3. Avaliacao Experimental 4. Dissertacao I. Machado, Patricia Duarte de Lima, Dra. II. Universidade Federal de Campina Grande - Campina Grande (PB) III. Título

CDU 004.415.532.2(043)

"UMA AVALIAÇÃO E INVESTIGAÇÃO EXPERIMENTAL DE TÉCNICAS DE RE-TESTE SELETIVO PARA TESTE DE REGRESSÃO BASEADO EM ESPECIFICAÇÃO"

FRANCISCO GOMES DE OLIVEIRA NETO

DISSERTAÇÃO APROVADA EM 13.12.2010

PATRICIA DUARTE DE LIMA MACHADO, Ph.D
Orientador(a)

JORGE CESAR ABRANTES DE FIGUEIREDO, D.Sc
Examinador(a)

JULIANO MANABU IYODA, Ph.D
Examinador(a)

CAMPINA GRANDE - PB

Agradecimentos

Estas palavras são destinadas às pessoas que participaram de meus momentos de “iluminação” e “escuridão”, durante a realização deste trabalho. Estas palavras podem ser complementadas apenas com o sentimento de gratidão das quais foram inspiradas.

Agradeço primeiramente a **Deus**, ser que mais amo, que me deu o privilégio da vida e a benção da saúde.

A minha amada **avó Ana**, que sempre me abençoou com sabedoria e conselhos que levarei comigo durante a minha existência.

A **meus pais**, Lucia e Luís, que sempre estiveram ao meu lado, fornecendo, amor, apoio, sabedoria e conforto.

A **meus irmãos** Adam Bruno e Iris Emmanuelle, que contribuem bastante para meu crescimento pessoal, sempre estando disponíveis nas horas mais necessárias, mostrando o verdadeiro significado do amor fraternal.

A **Emanuela Cartaxo**, meu modelo e inspiração profissional, que contribui infinitamente para minha formação pessoal, profissional e acadêmica. Obrigado Manu, pelas violentas discussões profissionais, pelas intrigantes fofocas pessoais, pelas confortáveis caronas eventuais, e acima de tudo, por me ajudar a enxergar o meu potencial. Enfim, obrigado por **tudo**!

À **Professora Patrícia Machado**, pela paciência, confiança e orientação, bastante importantes para minha motivação, empenho e desempenho.

A **João Felipe Ouriques**, inicialmente, pela ajuda na implementação das técnicas deste trabalho, e especialmente, por todo o apoio e companheirismo que tornam o cotidiano profissional, em um momento agradável e prazeroso.

A minha grande amiga, **Andréa Mendonça**, pelo seu carinho em ouvir minhas palavras de angústia e incertezas, apenas para me retornar palavras de amparo, amizade e motivação. Além disto, sou muito grato pela sua ajuda com as minhas dúvidas a respeito do processo de

estudos experimentais.

A **Elloá Guedes**, uma das minhas primeiras amigas, pela inspiração em ser determinado, competente, responsável, dentre outras várias características que almejo ter e que são tão inerentes a você. Muito obrigado também pela sua ajuda na fundamentação em análise estatística, e na execução do experimento.

Aos **meus melhores amigos** Arthur Yuri, Rafaella Italiano, Paloma Freire e Roberta Guedes. Por participarem de forma tão maravilhosa em minha vida, me oferecendo "mãos" e "ombros" sempre que precisei. Meus verdadeiros anjos que me fazem enxergar o valor da amizade e da confiança.

À toda a **equipe do Grupo de Métodos Formais** da Universidade Federal de Campina Grande pelo ambiente e sentimento de trabalho em equipe.

À **equipe** do Projeto do **INES** na UFPE, pelas reuniões e trabalhos que contribuíram para minha formação na área de teste de software.

Aos **professores e funcionários da COPIN** pela grande ajuda, paciência e competência.

Ao professor **PhD. Hasan Ural**, por disponibilizar a Dissertação de Xie Bo, necessária para a implementação de uma das técnicas analisadas neste trabalho.

À **CAPES** e ao **INES**, pelo apoio e suporte financeiro fornecidos a este trabalho.

Resumo

Técnicas de re-teste seletivo, para teste de regressão baseado em especificação, podem ser utilizadas para aumentar a confiabilidade de sistemas computacionais cujas funcionalidades ou requisitos foram modificados ao longo do tempo. Existem diversas técnicas definidas na literatura, porém, ao contrário das técnicas propostas para o contexto do código, ainda não há muito conhecimento acerca dos benefícios, limitações e características, das técnicas propostas para o contexto de especificação, a não ser por expectativas. Este trabalho apresenta uma investigação experimental acerca de cinco técnicas de re-teste seletivo baseado em especificação sob cinco aspectos: inclusão, precisão, eficiência, potencial de redução e densidade de faltas. Estes critérios são amplamente utilizados na avaliação de técnicas de teste de regressão. As técnicas analisadas utilizam a abordagem de Teste Baseado em Modelos para realizar a geração e seleção automática dos casos de teste de regressão. Além das técnicas presentes na literatura, aqui analisadas, foi proposta uma nova técnica (*Weighted Similarity Approach for Regression Testing - WSA-RT*), capaz de reduzir os custos do processo e aumentar a capacidade de detecção de faltas de regressão. A partir dos resultados do experimento conseguimos identificar as vantagens e desvantagens de cada técnica, assim como, apresentar os aspectos de aplicabilidade dessas técnicas, a partir de uma análise de generalidade. Foram seguidas todas as etapas de um processo experimental, portanto, as conclusões obtidas a respeito do desempenho das técnicas analisadas são estatisticamente significativas.

Palavras-chaves: Teste de Regressão, Técnicas de Seleção de Casos de Teste, Teste Baseado em Modelos.

Abstract

Specification-based selective regression testing (selective retesting) techniques can be used to increase the reliability of computer systems which functionalities and/or requirements have been modified. Several technique have been proposed, however, unlike the code-based techniques, there isn't much knowledge about the benefits, limitations and characteristics from specification-based techniques, except for expectations. This work presents an experimental investigation of five specification-based selective retesting techniques, analyzed under five properties: inclusiveness, precision, efficiency, reduction potential and fault density. These properties are widely used when evaluating selective retesting techniques. The analyzed techniques use a Model-Based Testing approach, where test cases are automatically generated and selected. Besides the techniques presented in the literature, analyzed here, we propose a new technique (Weighted Similarity Approach for Regression Testing - WSA-RT), able to reduce costs for software testing and increase fault detection. From the results of the experiment we were able to identify the advantages and disadvantages of each technique, as well as describing aspects of applicability of these techniques by performing an analysis of generality. We followed all the steps of a process for an experimental study, therefore, the obtained conclusions concerning the performance of the analyzed techniques, are statistically significant.

Keywords: Regression Testing, Test Case Selection Techniques, Model-based Testing.

Sumário

Acrônimos

ANOVA	Análise de Variância
GED	Grafo Estático de Dependência
GFC	Grafo de Fluxo de Controle
IDE	*Integrated Development Environment*
LTS-BT	*Labeled Transition System Based-Testing*
MEFE	Máquinas de Estados Finitas Estendidas
STR	Sistema de Transições Rotuladas
TBM	Teste Baseado em Modelos
TGF	*Trivial Graph Format*
UML	*Unified Modeling Language*
WSA	*Weighted Similarity Approach*
WSA-RT	*Weighted Similarity Approach for Regression Testing*
XMI	*XML Metadata Interchange*
XML	*Extensible Markup Language*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Um dos principais objetivos da engenharia de software é desenvolver aplicações de alta qualidade [Gimenes et al. 1999]. Esta característica, unida ao alto custo de realização das atividades de testes [Beizer 1990], motivou o desenvolvimento de diversas técnicas com a finalidade de simplificar e diminuir o custo de realização de testes em um software. Atualmente, o alvo é melhorar a eficiência e completude de tais técnicas aumentando a qualidade das aplicações de software.

A realização adequada de testes em um elemento[1] novo ou modificado de um sistema, provê características importantes, para aumentar a confiança neste elemento. No entanto, quando estes elementos interagem, ou possuem alta dependência com outros elementos do sistema, esta confiança fica novamente comprometida [Binder 1999]. Mudanças em um destes elementos podem causar falhas em sistemas já funcionais, seja considerando, individualmente, os componentes que estruturam o sistema, ou falhas resultantes de interações e efeitos colaterais indesejados. Dessa forma, após as modificações, é recomendado que sejam realizadas atividades de Teste de Regressão [Agrawal et al. 1993].

Teste de regressão é uma atividade realizada em elementos modificados de um sistema, com o objetivo de capturar faltas de regressão. Exemplos de atividades que justifiquem a realização de Teste de Regressão são: modificações para o lançamento de uma nova versão do software; os refinamentos arquiteturais, ou no código; as modificações nas funcionalidades do sistema; a integração de componentes (novos ou modificados) em um sistema; e a

[1] Neste documento o termo "elemento modificado" é interpretado como um programa, um trecho de código, um componente do sistema ou uma funcionalidade, que passou por algum processo de modificação.

manutenção do sistema. Estas atividades podem gerar faltas de regressão, que, por sua vez, são faltas que foram inseridas devido a um processo de modificação [Binder 1999].

O teste de regressão é bastante caro, pois, a sua realização necessita de um histórico de testes, ou seja, dos casos de teste das versões anteriores do software, pois, identificando os casos de testes que passavam e começaram a falhar após as modificações, é possível identificar as faltas de regressão. Em cenários onde o software possui diversos componentes, estruturas complexas, e passou por diversas modificações, a quantidade de casos de testes de regressão aumenta consideravelmente, de forma que o processo de execução destes casos de teste de regressão se torna muito caro, e muitas vezes inviável.

O alto custo de execução e a dificuldade de gerenciar os recursos e artefatos durante as atividades de teste de regressão caracterizam um dos principais problemas deste tipo de testes. A partir deste problema, algumas atividades foram desenvolvidas para a redução destes custos e dificuldades (descritas no Capítulo 2 na seção 2.3) [Harrold and Orso 2008]. Uma destas atividades é o re-teste seletivo [Rothermel and Harrold 1996], que é caracterizada pela utilização de técnicas para reduzir o tamanho da suíte de teste de regressão.

No contexto de teste de regressão, é necessário obter um subconjunto da suíte de testes de regressão reduzindo a quantidade de casos de teste de regressão para execução. Diante disto, é importante que o subconjunto obtido durante este processo contemple o maior número possível de casos de teste que revelem faltas de regressão.

A realização de teste de regressão e re-teste seletivo, ocorre em dois contextos: o contexto de código e o contexto de especificação. O primeiro, denominado Teste de Regressão baseado em Código, utiliza o código da aplicação como principal diretriz para realização dos testes. Por sua vez, o Teste de Regressão baseado em Especificação, utiliza elementos da especificação para guiar a atividade de teste de regressão; dentre estes elementos se destacam os artefatos da especificação do sistema, como diagramas UML (*Unified Modeling Language*) e documentos de requisitos.

Dentre estes dois contextos, o código é o mais abordado [Korel et al. 2002]. No entanto, a utilização de código no contexto de teste de regressão possui três problemas [Chen et al. 2007]. O primeiro problema ocorre, pois os testes são aplicáveis ao código, de forma que, ao realizar o teste de regressão em um subsistema, ou em um nível maior de abstração o código não é gerenciável, considerando a perspectiva de dificultar o entendi-

mento do sistema, ou a rastreabilidade de um conjunto de requisitos e funcionalidades que poderiam ser reutilizáveis. O segundo problema está relacionado com as dificuldades de entendimento de testes, já que estes são desenvolvidos em função do código. O terceiro é a dependência do teste com o código, de forma que atividades de refatoramento do código (e.g. renomear métodos, classes) influenciam em alterações na suíte de testes de regressão.

Além de auxiliar na solução destes problemas, a abordagem baseada em especificação gera a expectativa de priorização mais precisa de casos de teste relativos à modificação de uma funcionalidade, visto que os casos de teste podem ser re-gerados a partir de um modelo. O teste de regressão baseado em especificação, por sua vez, também possui desvantagens. Dentre elas se destaca, a dificuldade em gerar casos de teste executáveis, uma vez que os casos de teste são descritos em função de elementos de especificação, que geralmente, possuem um alto nível de abstração.

É possível utilizar teste de regressão baseado em especificação para a realização de teste funcional, em especial quando há a disponibilidade de modelos. Diante disso, é viável utilizar Testes Baseados em Modelos (TBM) na atividade de teste de regressão [Korel et al. 2002]. TBM é uma técnica de teste de software que se beneficia da utilização do modelo de um software para a geração e execução dos casos de teste. Dessa forma, é possível avaliar a conformidade entre o produto desenvolvido e sua especificação. Esta abordagem é caracterizada pela comparação de comportamentos do modelo do sistema (saídas esperadas) com as saídas do sistema propriamente dito. Em outras palavras, TBM é uma abordagem que faz uso do modelo abstrato do sistema para testar uma implementação concreta.

TBM difere das demais práticas de testes pelo nível de automação que proporciona. A automação atinge o design dos casos de teste, possibilitando, algumas vezes, atingir também a execução. Dessa forma, pode ser criado um modelo comportamental esperado do sistema, em contraposição com o processo de escrita manual dos casos de teste. Então, a partir deste modelo, as ferramentas de TBM geram testes de forma automática. Portanto, uma característica de TBM é a explicitação deste modelo. A fase de design do processo de teste é baseada em um modelo do comportamento esperado do produto, mesmo no teste manual onde este modelo é uma representação informal (muitas vezes mental).

Muitas pesquisas têm sido desenvolvidas com o intuito de definir técnicas efetivas

de derivação de casos de teste a partir da especificação de requisitos das aplicações [Beizer 1990, Offutt and Abdurazik 1999, Briand and Labiche 2001, Barbosa et al. 2004, Nogueira et al. 2007]. Para isto, é necessário que a especificação dos requisitos do software seja formalmente definida [Tretmans and Brinksman 2002], de modo a caracterizar com exatidão o comportamento do sistema. Considerando que esta especificação (por exemplo, diagramas UML) está disponível, é possível utilizar a técnica de TBM para automaticamente derivar os casos de teste.

Com os elementos de especificação bem definidos, é possível obter, através de TBM, uma suíte de testes, e utilizando re-teste seletivo é possível reduzir esta suíte, podendo diminuir, portanto, os custos em um contexto de teste de regressão. A dificuldade está em relacionar os elementos do modelo que causam as faltas de regressão, com os casos de testes capazes de capturar estas faltas. A maioria dos trabalhos realizados em teste de regressão se concentram no teste baseado em código. Portanto, os trabalhos desenvolvidos para teste de regressão baseado em especificação caracterizam uma contribuição significativa para a área [Rothermel and Harrold 1996], uma vez que constitui um grande complemento para o trabalho realizado com as abordagens baseadas em código [Korel et al. 2002].

Exemplos desta contribuição na união dos dois contextos de teste de regressão, seria a utilização de técnicas baseadas em código para realização de teste de regressão em novas versões que contemplassem, apenas, modificações estruturais no código, por exemplo, refatoramento de classes, ou módulos do sistema; enquanto que técnicas baseadas em especificação fossem aplicadas em mudanças a níveis de funcionalidades ou requisitos do sistema. O teste de regressão baseado em código não é adequado para testar a modificação de funcionalidades, devido à dificuldade em identificar, precisamente, quais os casos de teste que devem ser executados para cobrir as respectivas funcionalidades modificadas, já que ao nível de abstração do código, é possível obter muitos casos de teste cobrindo uma ou várias funcionalidades. Por sua vez, o teste de regressão baseado em especificação não é adequado para testar os refatoramentos de código, pela lacuna presente entre os níveis de abstração do código e da especificação, que dificulta o mapeamento das funcionalidades que possam ser afetadas pelo mapeamento.

1.1 Problema e Solução Proposta

Este trabalho procura solucionar o seguinte problema: As propostas para TBM em Teste de Regressão baseado em Especificação ainda são poucas e não há muito conhecimento acerca de seus benefícios, limitações e características[2], a não ser por expectativas [Korel et al. 2002]. Dentre possíveis limitações de uma técnica estão, a viabilidade de aplicação da técnica, ou seja, dependências de ferramentas para a execução da técnica; capacidade de lidar apenas com mudanças específicas, e.g. apenas adição de funcionalidades; dentre outras. Como possíveis benefícios de uma técnica, podem ser citados: O aumento na capacidade de captura de faltas de regressão, ou precisão desta captura; pouco custo de execução da técnica, ou seleção de um subconjunto muito pequeno causando a redução significativa dos custos de execução dos casos de teste de regressão.

Estes aspectos podem ser obtidos através da realização de um Estudo Experimental. Este tipo de estudo é fundamental para analisar estes pontos de forma mais precisa e controlada [Wohlin et al. 2000], possibilitando a definição do estado da arte e desafios a serem superados pelas técnicas de re-teste seletivo. Dessa forma, para realizar esta atividade é necessário contemplar o processo de experimentos.

Além da realização do estudo experimental, este trabalho propõe uma nova técnica de re-teste seletivo baseado em especificação. Esta técnica é adaptada a partir da técnica *Weighted Similarity Approach* (WSA), proposta por Bertolino et al. [Bertolino et al. 2008], para a seleção automática de casos de teste em TBM. A motivação desta adaptação é propor uma técnica de re-teste seletivo baseado em especificação capaz de selecionar um subconjunto da suíte de testes de regressão, observando as similaridades entre os casos de teste de versões diferentes do software e o seu respectivo perfil de uso. A técnica proposta, também é analisada no estudo experimental, com o objetivo de obter suas características, benefícios e limitações, assim como, comparar o seu desempenho com o de outras técnicas propostas na literatura.

[2]Características, neste contexto, pode ser interpretado como aspectos particulares de uma técnica, e.g. capacidade de capturar de faltas de regressão.

1.2 Objetivos

O objetivo geral deste trabalho é: "*a investigação e análise experimental de diferentes técnicas de re-teste seletivo em teste de regressão baseado em especificação*" Como resultado deste trabalho, são obtidas as características, vantagens e limitações de cada técnica de re-teste seletivo escolhida, sob um mesmo conjunto de propriedades utilizadas para caracterizar técnicas de seleção de casos de teste. A atividade que possibilitará a obtenção deste resultado é um estudo experimental.

Mais especificamente, as metas definidas para atingir este objetivo contemplam:

- Investigar e implementar técnicas representativas na área de teste de regressão baseada em especificação, tais como, a Análise de Dependência baseada em Máquinas de Estado Finitas Estendidas [Korel et al. 2002, Chen et al. 2007], e a Seleção baseada em *Clustering* [Laski and Szermer 1992];

- Adaptar e implementar a técnica WSA para o contexto de teste de regressão;

- Definir um modelo de estudo experimental que contemple aspectos significativos do teste de regressão e das técnicas de re-teste seletivo, como por exemplo, a quantidade de casos de teste que capturam faltas de regressão, e os custo de execução das técnicas, respectivamente.

- Conduzir e analisar os resultados do estudo experimental com o objetivo de obter informações a respeito das limitações, vantagens e características de cada técnica utilizada no experimento. Estes aspectos devem ser observados a partir das hipóteses nulas e das hipóteses alternativas definidas durante o planejamento do experimento [Wohlin et al. 2000].

O objetivo geral é alcançado através dos resultados do estudo experimental, enquanto que as ferramentas principais para atingir os objetivos específicos (metas) definidos acima, são as técnicas de re-teste seletivo, uma vez que estas técnicas, junto com a execução do estudo experimental, são necessários para a obtenção dos resultados. A partir destes objetivos e metas, a próxima seção contempla a relevância do trabalho.

1.3 Avaliação

Neste trabalho, são propostos: uma técnica de re-teste seletivo baseado em especificação, e um estudo experimental de técnicas de re-teste seletivo baseado em especificação. Inicialmente, a técnica proposta é avaliada através do estudo experimental. Neste estudo, a técnica proposta é executada diversas vezes, em um contexto experimental.

Os dados coletados refletem as propriedades das técnicas investigadas na literatura de teste de regressão e TBM [Rothermel and Harrold 1996, Rothermel and Harrold 1997, Graves et al. 1998, Cartaxo et al. 2009, Mahdian et al. 2009]. Estes resultados são então comparados com os das demais técnicas em um estudo experimental. Esta comparação provê uma perspectiva sobre o desempenho da seleção realizada pela técnica proposta.

Por sua vez, um estudo experimental é avaliado a partir de seus aspectos de validade. Dessa forma, nosso estudo experimental passa por uma avaliação de validade que considera 4 aspectos: validade de conclusão, validade interna, validade externa e validade de construção. Além disto, são identificadas as ameaças à validade do estudo, e determinamos como controlar estas ameaças.

O plano experimental, dado o controle e avaliação da validade, podem ser repetidos e/ou adaptados para outros contextos (TBM, teste de regressão baseado em código, teste de integração, teste funcional, dentre outros), auxiliando a pesquisa e o desenvolvimento de técnicas de seleção de casos de testes. Dessa forma, demais pesquisadores da comunidade podem reproduzir os resultados, com o objetivo de verificar as conclusões obtidas, ou então, adaptar o estudo, procurando lidar com as ameaças e níveis de controle que não são contemplados neste estudo. Para este estudo, esses elementos (precisão, ameaças à validade e controle) são discutidos durante o processo do experimento.

1.4 Contribuições do Trabalho

A quantidade de estudos experimentais, realizados na Ciência da Computação, ainda não é suficiente para consolidar a quantidade de métodos e tecnologias propostas na área [Feitelson and Russell 2006]. Diante disto, a realização de um estudo experimental para investigar e avaliar um conjunto de técnicas caracteriza uma das contribuições acadêmicas

deste trabalho.

Sob o contexto de teste de regressão, diversas técnicas de re-teste seletivo foram propostas na literatura [Korel et al. 2002]. No entanto, estudos que possibilitem destacar as vantagens, limitações, viabilidade e características de cada técnica ainda são insuficientes [Graves et al. 2001], pois estes abordam poucas técnicas (geralmente, os estudos comparam 2 técnicas), ou caracterizam os resultados apenas para classes de técnicas (e.g. técnicas de minimização [Graves et al. 2001], redução de suítes de teste [Ma et al. 2005], dentre outros) deixando lacunas que prejudicam o esclarecimento do desempenho das técnicas. Estas lacunas estão relacionadas com a aplicabilidade das técnicas em, por exemplo, diferentes processos de desenvolvimento.

Portanto, é necessário observar aspectos como: a redução de custos ao utilizar a técnica, a confiabilidade no processo de seleção realizado, dentre outros. Através das variáveis investigadas, é possível observar a capacidade de capturar faltas de regressão (inclusão, precisão e densidade de faltas), reduzir os custos de execução das técnicas (eficiência) ou o custo de realização do teste de regressão (potencial de redução). Diante disto, os dados obtidos de cada técnica podem contribuir para a escolha da técnica mais adequada para os cenários de recursos em que um projeto esteja inserido, facilitando a realização de um processo de teste de regressão, e diminuindo, portanto, os custos do processo de teste de um software.

Por exemplo, é possível observar que as técnicas com um alto desempenho de eficiência e potencial de redução, são aplicáveis em processos com uma alta restrição de recursos, pois, é adequada a utilização de uma técnica rápida e capaz de reduzir bastante a suíte de regressão. Por sua vez, um software crítico, como aplicações de sistemas de tempo real, ou ciências médicas, é recomendado utilizar a técnica com os melhores desempenhos de inclusão e densidade de faltas, pois a cobertura de faltas de regressão é mais importante, neste cenário.

Considerando este contexto, e na tentativa de melhorar a compreensão sobre as vantagens e as limitações de técnicas de re-teste seletivo baseado em especificação propostas na literatura, o presente trabalho apresenta as seguintes contribuições:

- **A utilização de TBM em teste de regressão:** Ao utilizar TBM junto ao teste de regressão há uma contribuição para a redução de custos da etapa de teste, pois a utilização de TBM possibilita a geração e seleção automática da suíte de testes;

- **A técnica *Weighted Similarity Aprroach for Regression Testing*:** A técnica desenvolvida neste trabalho apresenta uma perspectiva de seleção de casos de teste de regressão considerando as modificações realizadas no software e as informações de uso. Os resultados desta técnica no estudo experimental revelam as vantagens de sua utilização, dentre elas, observamos que 50% dos casos de teste selecionados capturam faltas de regressão, mesmo com uma redução de 70% na quantidade de casos de teste da suíte de regressão. A partir destes resultados, verificamos que a técnica apresenta uma grande redução de custos, e uma boa cobertura de faltas de regressão;

- **Realização de um estudo experimental sobre re-teste seletivo baseado em especificação:** Os elementos utilizados para o planejamento, execução e avaliação dos resultados, do estudo experimental, deste trabalho, podem ser estruturados em um *framework*. Dessa forma, as técnicas de re-teste seletivo, propostas após a realização deste trabalho, podem ser avaliadas, e comparadas com os resultados obtidos neste trabalho;

- **Implementação das técnicas e do experimento:** As técnicas e o estudo experimental deste trabalho estão implementados na ferramenta LTS-BT (*Labeled Transitions System - Based Testing*) [Cartaxo et al. 2008, Oliveira Neto and Machado 2008]. Dessa forma, fornecemos o suporte ferramental para a utilização destas técnicas, e a realização de outros estudos experimentais;

- **Resultados empíricos sobre o desempenho das técnicas:** A análise e comparação de desempenho das técnicas a partir de propriedades que caracterizam o desempenho das técnicas fornecem informações que facilitam a escolha de técnicas para diferentes cenários de desenvolvimento (e.g. restrições de custo, disponibilidade de tempo, poucos recursos humanos, níveis de experiência dos testadores, dentre outros). Estas propriedades são: inclusão, precisão, eficiência, potencial de redução e densidade de faltas (descritas no Capítulo 2, Seção 2.4.1).

1.5 Considerações Finais do Capítulo

Neste capítulo foram apresentados os aspectos de introdução do trabalho realizado. Foram discutidos a contextualização e motivação do trabalho, assim como o problema e a solução proposta, os objetivos e as contribuições do trabalho. Os próximos capítulos contemplam a fundamentação teórica, a metodologia e os resultados obtidos neste trabalho.

O Capítulo 2 apresenta a fundamentação teórica do trabalho, e.g. teste de regressão e estudos experimentais em engenharia de software. As técnicas de re-teste seletivo da literatura analisadas neste trabalho são descritas no Capítulo 3, enquanto que a técnica proposta neste trabalho é apresentada no Capítulo 4. Todas as informações apresentadas nestes capítulos, são importantes para a melhor compreensão das etapas de definição, planejamento e instrumentação do estudo experimental.

Uma vez especificada toda a fundamentação teórica, são apresentados os aspectos da metodologia deste trabalho, contemplados pelos Capítulos 5 e 6 que descrevem a definição e planejamento, e a instrumentação do estudo experimental, respectivamente. Os resultados, análise e conclusões são apresentados no Capítulo 7. Por fim, os aspectos de considerações finais deste trabalho, como por exemplo, os trabalhos relacionados e futuros, são discutidos no Capítulo 8.

Capítulo 2

Fundamentação Teórica

Este trabalho é desenvolvido na área de Teste de Software. O teste de software é uma das fases do processo de engenharia de software que possui o objetivo de encontrar defeitos no software para que estes sejam corrigidos antes da entrega final do produto. Dessa forma, esta atividade é realizada para aferir a qualidade do software. O teste de software é uma atividade com um alto custo, devido ao processo de correção de erros e manutenção do software [Beizer 1990]. Portanto, métodos e técnicas sistemáticas de teste são desenvolvidos com o objetivo de diminuir este custo relacionado ao teste de software [McGregor and Sykes 2001].

A essência do teste de software é determinar o conjunto de casos de teste para o software a ser testado. Estes casos de teste são constituídos de entradas (condições iniciais e passos do sistema), e saídas (pós-condições e resultados esperados do sistema) [Jorgensen 1995]. O conjunto de casos de teste definido para uma dada aplicação é denominado suíte de teste.

Neste capítulo será apresentada a Fundamentação Teórica referente ao trabalho desenvolvido. Esta fundamentação contempla aspectos do teste de software como: teste baseado em modelos, abordagem baseada em valores, teste de regressão, re-teste seletivo, e as propriedades observadas em técnicas de re-teste seletivo. Além destes tópicos de fundamentação, serão apresentados alguns princípios de experimentação em engenharia de software. Por fim, serão apresentados alguns fundamentos de análise estatística necessários para melhor compreender a análise conduzida e os respectivos resultados obtidos neste estudo experimental.

2.1 Teste Baseado em Modelo

Teste baseado em modelo (TBM) é uma abordagem *black-box* para geração automática de teste de software a partir de modelos da aplicação [El-Far 2001]. Dessa forma, os casos de testes são executados para avaliar a correspondência entre o modelo e a aplicação. Para isto, é necessário que a especificação da aplicação a ser testada esteja descrita através de um modelo, caracterizando, portanto, o seu comportamento [Beizer 1995, Dalal et al. 1999].

Na Figura 2.1 são representados os principais artefatos e atividades da abordagem TBM. A seguir, apresentamos a descrição das principais atividades relacionadas ao teste baseado em modelos TBM [El-Far 2001]:

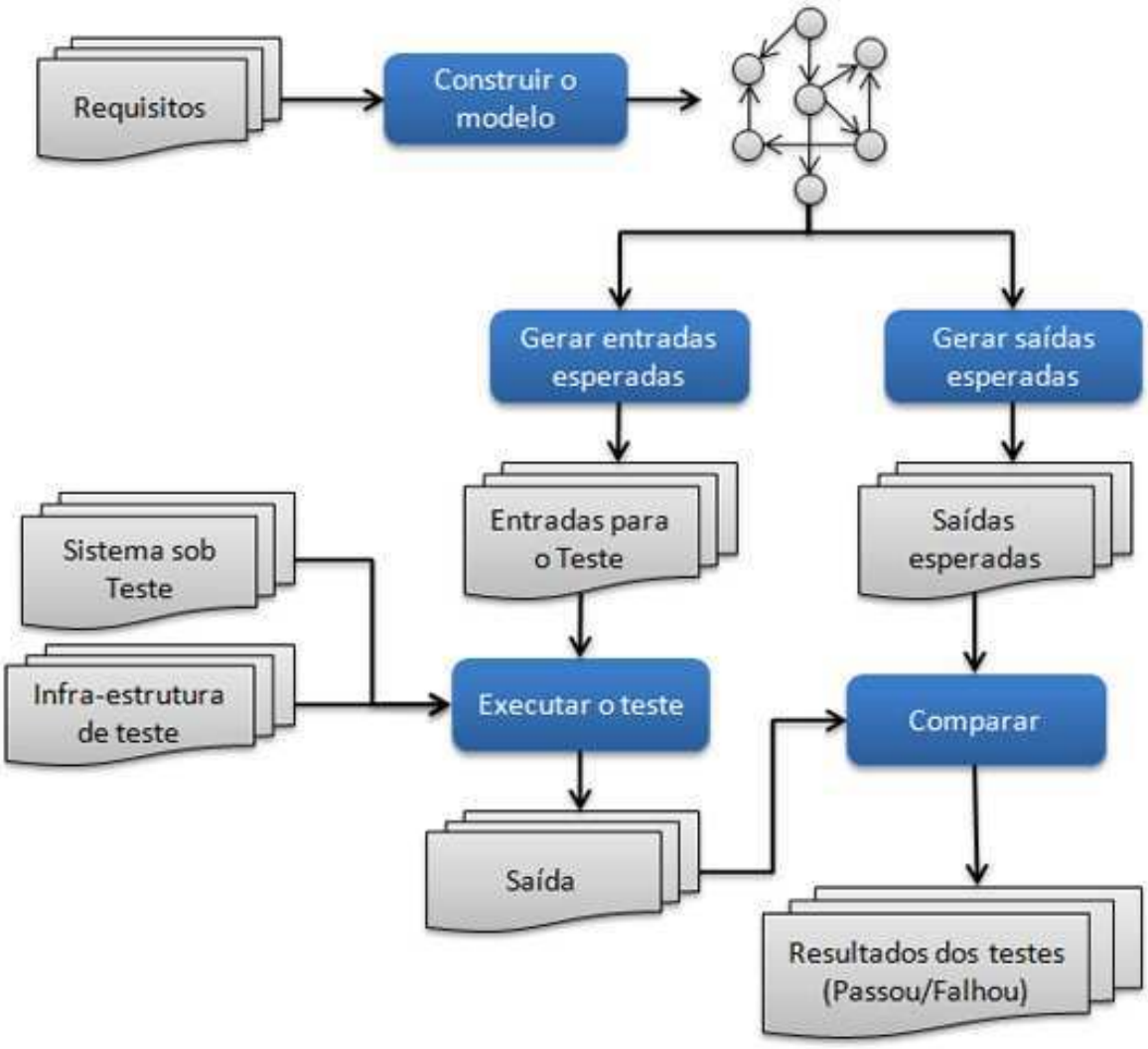

Figura 2.1: Atividades e artefatos do Teste baseado em Modelos.

- **Construir o modelo:** Construção do modelo formal a partir dos requisitos da aplicação;

- **Gerar entradas e saídas esperadas:** A geração de entradas, e a geração de saídas esperadas são duas atividades realizadas em paralelo. A geração das entradas de testes parte do modelo. Essas entradas são passos que servirão para exercitar a aplicação que está sendo testada. Semelhantemente, as saídas esperadas do teste são geradas a partir do modelo formal, e indicam o comportamento esperado do sistema.

- **Executar testes:** Execução da aplicação com as entradas geradas, produzindo saídas;

- **Comparar saídas com saídas esperadas:** As saídas esperadas geradas a partir do modelo são comparadas com as saídas da aplicação que está sendo testada.

O processo de TBM se inicia com os requisitos. Dessa forma, o processo de teste pode ser iniciado assim que os requisitos da aplicação estiverem definidos. Com os requisitos definidos, o próximo passo é a construção de um modelo que retrate de forma íntegra o comportamento requisitado. A especificação dos casos de teste inclui entradas e saídas esperadas. A partir da verificação entre saídas esperadas, e as saídas obtidas, é possível avaliar a presença de defeitos na aplicação. Dessa forma, podemos observar que a utilização de TBM pode nos fornece as seguintes vantagens [El-Far 2001]:

- **Comunicação entre desenvolvedores e testadores:** O modelo do comportamento da aplicação pode ser utilizado como a base de comunicação entre testadores e desenvolvedores;

- **Geração automática de testes:** A geração de casos de teste pode ser facilmente automatizada a partir do modelo do comportamento da aplicação;

No entanto, a utilização de TBM tem a desvantagem de requerer conhecimento da notação do modelo. Dessa forma, o testador deve ser familiar com a notação que será utilizada, o que culmina na necessidade de investimento em treinamentos, além do tempo necessário para a obtenção do modelo [El-Far 2001]. Outra desvantagem é a dependência da existência do modelo, bem como a relação entre a qualidade dos casos de testes obtidos e a qualidade do modelo.

2.2 Abordagem baseada em Valores

Muitos trabalhos da pesquisa e prática de engenharia de software, são baseados numa configuração neutra de valores. Na abordagem neutra de valores, todos os casos de uso, requisitos, ou funcionalidades, são considerados como iguais. A engenharia de software baseada em valor [Boehm 2006], é inspirada na idéia de que a qualidade não deve ser considerada uma meta, em si, na ausência de uma economia favorável [Boehm 1981]. Portanto, é necessário estruturar quais são os "valores" desejáveis para o software.

Os valores são os recursos utilizados para diferenciar a importância de elementos como casos de uso, requisitos, ou funcionalidades no produto desenvolvido. Sob esta perspectiva, diversos valores relacionados ao software têm sido incorporados em processos de desenvolvimento bem sucedidos [Bertolino et al. 2008].

O principal objetivo desta abordagem é destacar as considerações e o conhecimento necessários para que as decisões de engenharia de software otimizem os valores do produto. Porém, definir, claramente, o que estes valores representam não é simples, pois eles podem estar relacionados com diversos aspectos como: qualidade, disponibilidade, confiabilidade, valores de probabilidade, recursos de tempo, dentre outros [Bertolino et al. 2008].

Para a engenharia de software, os valores podem ser baseados nas características do produto, como por exemplo um conjunto de funcionalidades críticas para a execução do sistema, ou a quantidade de faltas encontradas no sistema. Apesar de representar um aspecto subjetivo do produto, é recomendado que o valor seja expresso através de elementos quantitativos. Diante disto, estes elementos quantitativos podem ser utilizados para identificar elementos do produto (trechos de código, casos de teste, requisitos) que apresentam um maior valor.

Exemplos de trabalhos que utilizam uma abordagem baseada em valores no contexto de teste são as técnicas propostas por Musa para teste de confiabilidade [Musa 1998], e Bertolino et al. para a seleção de casos de teste em TBM [Bertolino et al. 2008]. Neste último trabalho, a abordagem baseada em valores é utilizada para estabelecer os critérios para a seleção de casos de teste, baseada no perfil de uso da aplicação. Diante disto, em situações onde todos os casos de teste não podem ser executados, é possível observar e executar os casos de teste que cobrem os elementos de maior valor do produto.

2.3 Teste de Regressão

Teste de regressão é uma atividade realizada em uma versão nova ou modificada de um software, com o objetivo de identificar faltas inseridas durante estas modificações. Esta atividade tenta revalidar as antigas funcionalidades herdadas pela versão anterior. Dentre as versões de um componente ou sistema, podemos definir a versão base e a versão delta. A versão base é a versão do componente/sistema que passou em uma suíte de teste. Uma versão delta de um componente/sistema é uma versão modificada que não passou em uma suíte de teste de regressão. Uma *delta build* é uma configuração executável do Sistema sob Teste (SSB) que contém todos os componentes base e delta [Binder 1999].

Uma suíte de testes de regressão é composta por diversos casos de teste de regressão. Casos de teste de regressão são definidos como casos de testes que passaram na versão base e é esperado que passem na *delta build*. Quando encontramos um caso de teste que passava na versão base e não passa mais na versão delta, localizamos uma "falta de regressão".

Teste de regressão é uma atividade que pode ser utilizada para aumentar a confiança e qualidade de um sistema, em um cenário onde constantes mudanças são realizadas [Korel et al. 2002]. Ainda que muito importante, esta atividade também é muito cara, pois exige uma grande quantidade de tempo e recursos computacionais, em especial quando lidamos com sistemas de software grande (i.e., um sistema com muitos subsistemas e componentes integrados). Durante o teste de regressão, casos de teste desenvolvidos anteriormente são implantados para revalidar um sistema modificado, bem como novos casos de teste são freqüentemente gerados [Agrawal et al. 1993].

Diante disto, alguns relatórios indicam que a etapa de teste de regressão chega a consumir cerca de 80% dos recursos da etapa de Teste e até 50% dos recursos da etapa de manutenção. Dessa forma, diversos pesquisadores têm desenvolvidos trabalhos com o objetivo de reduzir este custo [Harrold and Orso 2008]. Estes trabalhos consistem em técnicas para lidar com as diversas atividades realizadas em um teste de regressão, que possibilitem a redução destes custos. Algumas destas atividades são ilustradas na Figura 2.2

As atividades (representadas pelas setas na Figura 2.2) caracterizam também os principais problemas abordados na área de teste de regressão [Harrold and Orso 2008]. Através de técnicas propostas para a realização destas atividades, os custos do teste de regressão po-

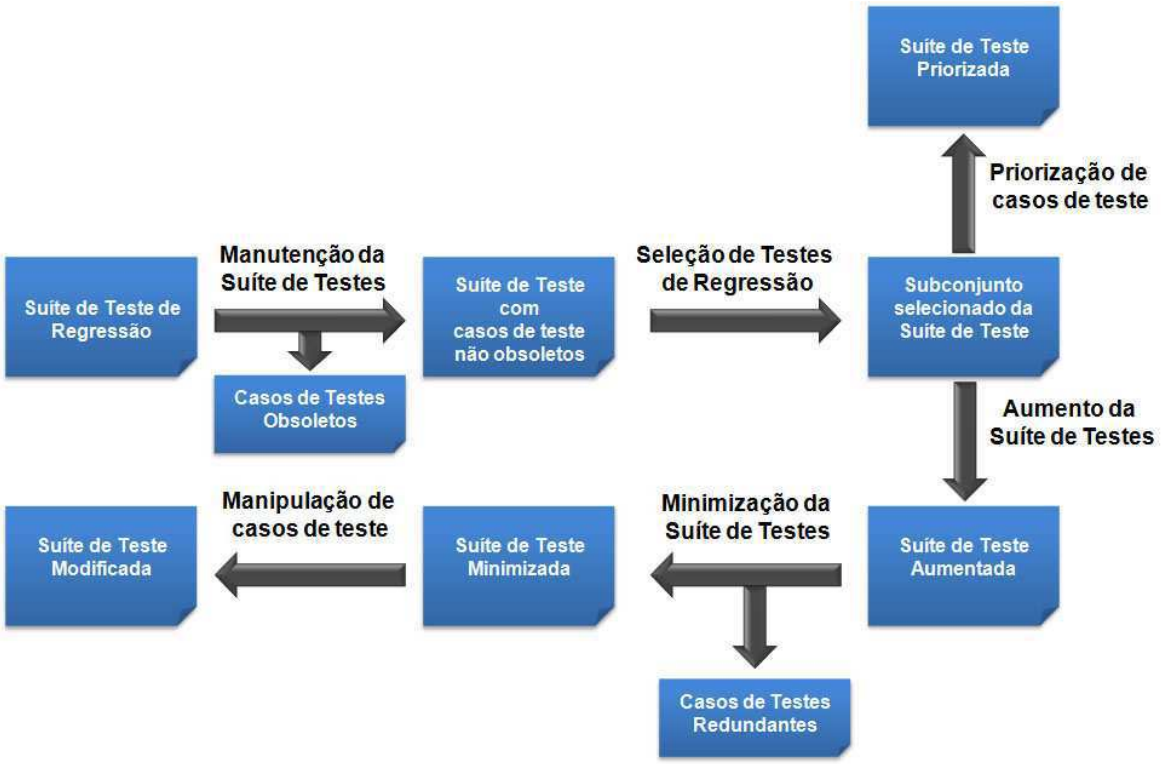

Figura 2.2: Atividades de um Teste de Regressão [Adaptada de Harrold e Orso]

dem ser reduzidos. Considerando um cenário de teste de regressão convencional, com um software C, uma suíte de testes T para esta versão do software, e uma versão modificada C', obtida de C, as atividades da Figura 2.2 são caracterizadas da seguinte forma:

- **Manutenção da Suíte de Testes:** Esta atividade contempla a manutenção nos casos de teste da suíte de testes T. O objetivo desta manutenção é identificar os casos de teste obsoletos, ou seja, casos de teste que não são mais executáveis na nova versão do software (C'). Um caso de teste obsoleto que falha não revela faltas de regressão, pois a falha não é causada pela modificação em si, e sim pelo fato do caso de teste não ser mais executável na versão correspondete do sistema. Um cenário onde casos de testes obsoletos são, frequentemente, encontrados ocorre quando há modificações na interface gráfica do software.

- **Seleção de Teste de Regressão:** Esta atividade é realizada no subconjunto obtido como resultado da atividade de Manutenção da Suíte de Testes. Mesmo com a remoção dos casos de teste obsoletos, o tamanho da suíte de testes de regressão pode inviabilizar

a execução de todos os casos de teste devido à indisponibilidade de recursos. por exemplo. Dessa forma, é necessário selecionar um subconjunto de casos de teste, a partir da suíte de teste de regressão, utilizando técnicas de re-teste seletivo.

- **Priorização da Suíte de Testes:** Mesmo após o processo de seleção dos casos de testes, a suíte de teste pode ser organizada com o objetivo de priorizar os casos de teste. Esta organização pode ser definida de acordo com critérios, como cobertura de requisitos ou estimativas de faltas de regressão [Harrold and Orso 2008], com o objetivo de diminuir o tempo de execução do teste de regressão. A priorização possui um objetivo distinto da selação de casos de teste. A priorização procura ordenar a suíte de testes, de forma que a ordem em que os casos de teste são executados afeta o processo de detecção de faltas. A seleção, por sua vez, procura reduzir a quantidade de casos de teste que devem ser executados.

- **Aumento da Suíte de Teste:** O objetivo desta atividade é adicionar casos de teste na suíte de teste de regressão, para contemplar algum critério de cobertura que não foi atingido pela suíte disponível (esta suíte pode ser referente à versão base, ou a suíte resultante das atividades anteriores).

- **Minimização da Suíte de Testes**: O objetivo da minimização é remover as redundâncias presentes na suíte de teste de regressão. Estas redundâncias podem ser resultados da atividade de aumento na suíte de testes. Esta atividade considera como redundantes, os casos de teste que exercitam o mesmo comportamento na aplicação.

- **Manipulação de casos de teste:** Esta atividade é realizada com o objetivo de manipular o caso de teste para reduzir os recursos necessários para a sua criação. Dessa forma, são verificados aspectos que viabilizem o reuso dos casos de teste, como por exemplo, realizar a geração dos casos de testes em diversos níveis de granularidade, para acelerar a execução e reuso de apenas algumas partes dos casos de teste.

Uma etapa de Teste de Regressão pode contemplar apenas algumas ou nenhuma destas atividades [Harrold and Orso 2008]. O nosso trabalho, se encaixa na atividade de Seleção de Testes de Regressão. Esta atividade também é conhecida como re-teste seletivo, e é geralmente realizada através de técnicas automáticas que selecionam subconjuntos da suíte

de teste de regressão, para diminuir os custos. O re-teste seletivo é discutido na subseção a seguir.

2.4 Re-teste Seletivo

Re-teste seletivo (do termo *selective retesting*, ou ainda *selective regression testing*), tem como principal objetivo a redução dos custos de testar um programa modificado. Este objetivo é alcançado através do reuso de testes existentes e da identificação de trechos modificados do programa, ou de sua especificação, que devem ser testados. Para descrever o processo, consideremos a mesma nomenclatura adotada na descrição do teste de regressão, onde: C um componente, sistema ou até mesmo especificação, C' uma versão modificada de C, e T uma suíte de teste para C. O processo de re-teste seletivo é descrito da seguinte forma [Rothermel and Harrold 1996, Graves et al. 2001]:

1. Selecionar T', um subconjunto de casos de teste de T a serem executados na versão modificada, C'.

2. Testar C' com T', estabelecendo a corretude desta versão modificada com respeito a este subconjunto.

3. Se necessário, criar T'', um conjunto de novos casos de testes estruturais ou funcionais. Esta atividade é realizada quando a confiança na versão modificada após executar T' não foi atingida.

4. Testar a versão modificada, C', com este novo conjunto de casos de teste, estabelecendo a corretude de C' com respeito a T''.

5. Criar T''', uma nova suíte de teste e relatório de execução de testes para a versão modificada, a partir de T, T' e T'', onde estes três conjuntos de casos de teste serão o histórico de testes.

Quando comparada com a técnica de *retest-all* que executa todos os testes da suíte de regressão, re-teste seletivo é mais econômico se o custo de realizar a seleção do subconjunto T' da suíte é menor que executar os casos de teste que não serão incluídos nesta suíte

[Leung and White 1991]. Dentro deste processo, diversos problemas da área são identificados [Rothermel and Harrold 1997]. O primeiro deles é o Problema de Seleção de Testes de Regressão, caracterizado pelo problema de selecionar T', um subconjunto representativo de T para testar a versão modificada, C'. A maioria das técnicas de re-teste seletivo, são propostas com o objetivo de solucionar este problema [Graves et al. 2001] selecionando um subconjunto representativo a partir da suíte de testes de regressão.

Diante disso, o nosso trabalho aborda este problema verificando, através das técnicas escolhidas, as propriedades que refletem a representatividade do subconjunto selecionado. O segundo problema é referente à Identificação de Cobertura, caracterizado pela identificação adequada de porções da versão modificada, C', que necessitam de testes adicionais. O problema da Execução da Suíte de Teste é caracterizado pelo problema de executar os testes e verificar os respectivos resultados eficientemente. Por último, temos o problema da Manutenção da Suíte de Testes, caracterizado pelo problema de atualizar e armazenar as informações referentes aos testes.

Dentre estes problemas, as técnicas invcestigadas neste trabalho foram propostas com o objetivo de solucionar o problema de Seleção de Testes de Regressão. As técnicas utilizadas no estudo experimental deste trabalho selecionam subconjuntos da suíte de teste de regressão, para diminuir os custos do teste de regressão.

Cada técnica realiza um processo de seleção distinto, que pode, ou não, considerar aspectos como: critérios de cobertura (e.g. modificações, transições do modelo ou requisitos), cenários de uso do software (e.g. fluxos da aplicação mais prováveis de execução pelo usuário), dentre outros. Dessa forma, através da investigação experimental, verificamos a representatividade de cada subconjunto T' escolhido por cada técnica. Esta representatividade é verificada através das propriedades das suítes de teste de regressão, descritas na próxima subseção deste capítulo.

2.4.1 Propriedades das Técnicas de Re-teste Seletivo

Com o objetivo de resolver o problema de Seleção de Teste de Regressão, diversas técnicas de re-teste seletivo foram desenvolvidas. Apesar da grande quantidade, a escolha de uma técnica pode afetar de forma significativa a redução dos custos do processo de teste de regressão [Graves et al. 2001]. Diante disso, alguns estudos foram realizados com o objetivo

de obter informações a respeito de cada técnica sob um conjunto de propriedades comuns.

Rothermel e Harrold [Rothermel and Harrold 1996] desenvolveram um *framework* para a realização de análises referentes às técnicas de re-teste seletivo em teste de regressão. Esse *framework* foi utilizado por vários autores [Rothermel and Harrold 1997, Graves et al. 1998, Graves et al. 2001, Do et al. 2005, Mahdian et al. 2009] para analisar diversas técnicas para teste de regressão, e define um conjunto de propriedades utilizado para analisar o desempenho das técnicas de re-teste seletivo. Estas propriedades são a inclusão, precisão, eficiência e generalidade.

Uma vez que as técnicas analisadas neste estudo experimental são técnicas baseadas em elementos da especificação e utilizam uma abordagem de TBM, alguns elementos da análise de técnicas em TBM também foram utilizados para avaliar as técnicas deste experimento. Foram utilizadas duas propriedades das técnicas de TBM neste estudo experimental: o potencial de redução das técnicas, e a densidade de faltas da suíte selecionada. Com o objetivo de ilustrar a obtenção dos dados a respeito destas propriedades, será utilizado o cenário da Figura 2.3.

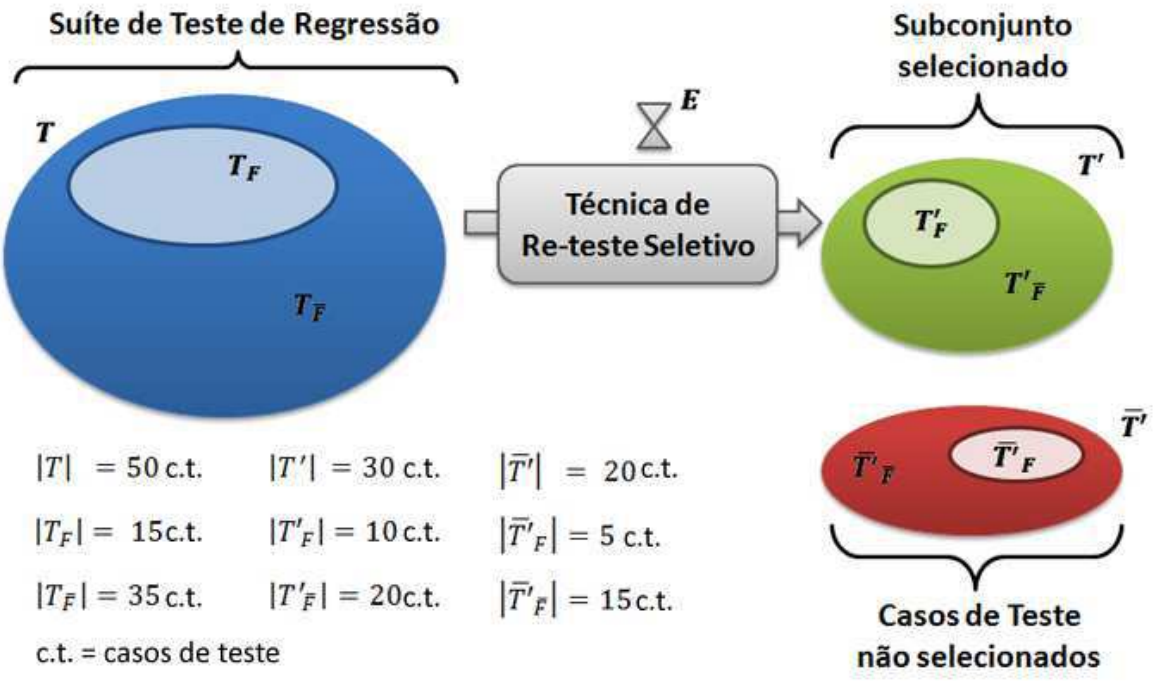

Figura 2.3: Conjuntos analisados nas técnicas de re-teste seletivo.

Na Figura 2.3, são ilustrados os seguintes conjuntos de casos de teste:

- T: Suíte de testes de regressão, submetida à técnica de re-teste seletivo.
- T': Suíte com os casos de testes selecionados pela técnica de re-teste seletivo.
- $\overline{T'}$: Suíte com os casos de testes que não foram selecionados pela técnica de re-teste seletivo.
- T_F, T'_F e $\overline{T'}_F$: Conjunto de casos de teste que capturam faltas de regressão, na suíte de regressão, na suíte selecionada, e na suíte não selecionada, respectivamente.
- $T_{\overline{F}}$, $T'_{\overline{F}}$ e $\overline{T'}_{\overline{F}}$: Conjunto de casos de teste que não capturam faltas de regressão, na suíte de regressão, na suíte selecionada, e na suíte não selecionada, respectivamente.
- E: Marcador de tempo de execução da técnica de re-teste seletivo.

Neste cenário, consideremos que a suíte de teste de regressão T com 50 casos de teste é submetida à uma técnica de re-teste seletivo. Destes 50 casos de testes em T, temos que 15 casos de teste capturam faltas de regressão (T_F), enquanto que os 35 restantes não capturam faltas de regressão ($T_{\overline{F}}$). Esta suíte de regressão é então submetida a uma técnica de re-teste seletivo.

Após executar esta técnica, são selecionados 30 casos de teste, que formam a suíte T'. Consideremos ainda que, no subconjunto selecionado, temos 10 casos de teste que revelam falta de regressão T'_F, enquanto que os 20 restantes de T' não capturam faltas de regressão ($T'_{\overline{F}}$). Além de observar os casos de teste selecionados, as propriedades verificam também os casos de teste não selecionados pela técnica.

Diante disto, vamos considerar $\overline{T'}$ o conjunto de casos de teste não selecionados pela técnica de re-teste seletivo. Portanto, temos que $\overline{T'}$ é formado por 20 casos de teste ($50-30 = 20$), e que dentre estes casos de teste, 5 capturam faltas de regressão ($\overline{T'}_F$), e os 15 restantes não capturam faltas de regressão ($\overline{T'}_{\overline{F}}$).

A partir dos conjuntos da Figura 2.3, e do cenário descrito acima, podemos observar os seguintes aspectos a respeito de T, T' e $\overline{T'}$:

- $T' \cup \overline{T'} = T$;
- $T' \cap \overline{T'} = \emptyset$;
- $T_F \cup T_{\overline{F}} = T$;
- $T'_F \cup T'_{\overline{F}} = T'$;
- $\overline{T'}_F \cup \overline{T'}_{\overline{F}} = \overline{T'}$;
- $T'_F \cup \overline{T'}_F = T_F$.

Estes conjuntos serão utilizados para explicar e ilustrar a obtenção da inclusão, precisão, potencial de redução e densidade de faltas das técnicas. As demais propriedades (eficiência e generalidade), neste experimento, são determinados a partir de outros elementos (descritos no Capítulo 5). Os aspectos de cada propriedade serão detalhados nas subseções a seguir.

Inclusão

A inclusão de uma técnica é a quantidade de casos de teste, na suíte reduzida, que revelam pelo menos uma falta de regressão. Esta quantidade é expressa em porcentagem, e está relacionada com a quantidade total de casos de teste, da suíte de regressão, que revelam pelo menos uma falta de regressão. Ou seja, a inclusão é o percentual de casos de teste selecionados que revelam faltas em relação ao número total de casos de teste (da suíte de regressão) que revelam faltas. Podemos definir a inclusão $I(T_x)$ de uma técnica T_x ($x = 1, 2, 3, \ldots$) da seguinte forma:

$$I(T_x) = \begin{cases} 100 \times \frac{|T'_F|}{|T_F|} & \text{se} \quad |T_F| \neq 0 \\ 100\% & \text{se} \quad |T_F| = 0. \end{cases} \tag{2.1}$$

Utilizando o exemplo ilustrado na Figura 2.3, temos:

$$\begin{aligned} I(T_x) &= 100 \times \frac{|T'_F|}{|T_F|} \\ &= 100 \times \frac{10}{15} \\ &\approxeq 67\%. \end{aligned}$$

Portanto, para o exemplo ilustrado, temos que a técnica de re-teste seletivo possui uma inclusão de 67%, ou seja, a suíte selecionada pela técnica é capaz de capturar 67% do total de faltas de regressão na suíte de regressão. As técnicas denominadas *seguras*, são as técnicas que possuem 100% de inclusão. Este nível de inclusão é atingido através de algoritmos que analisam o modelo, as modificações e as dependências destas modificações.

A inclusão de uma técnica de re-teste seletivo é utilizada para indicar se a suíte selecionada apresenta uma boa cobertura dos elementos modificados do sistema. As faltas de

regressão estão relacionadas com as modificações, seja diretamente ou através de efeitos colaterais durante a integração de componentes modificados [Binder 1999]. Diante disto, as técnicas com uma alta inclusão são recomendadas para obter uma maior cobertura de faltas de regressão no sistema, mesmo com a redução no tamanho da suíte de regressão.

Precisão

A precisão, ao contrário da inclusão, está relacionada com os casos de teste que não capturam faltas de regressão. Uma técnica é denominada precisa quando ela não seleciona os casos de teste que não capturam faltas de regressão. Diante disto, e a partir do exemplo da Figura 2.3, definimos a precisão $P(T_x)$ de uma técnica como:

$$P(T_x) = \begin{cases} 100 \times \frac{|\overline{T'}_{\overline{F}}|}{|T_{\overline{F}}|} & \text{se} \quad |T_F| \neq 0 \\ 100\% & \text{se} \quad x = 0 \end{cases} \tag{2.2}$$

Através do exemplo da Figura 2.3, podemos calcular a seguinte precisão para a técnica utilizada:

$$\begin{aligned} P(T_x) &= 100 \times \frac{|\overline{T'}_{\overline{F}}|}{|T_{\overline{F}}|} \\ &= 100 \times \frac{15}{35} \\ &\approx 42,8\%. \end{aligned}$$

Portanto, podemos observar que a técnica não seleciona 42% dos casos de teste que não capturam faltas de regressão. Uma técnica ideal deve ser segura e possuir 100% de precisão, ou seja, selecionar *apenas* os casos de teste que capturam as faltas de regressão. Algumas técnicas apresentam 100% de inclusão, no entanto, os pesquisadores em teste de regressão argumentam que 100% de precisão não é uma meta atingível [Rothermel and Harrold 1997, Graves et al. 2001], pois não há um mecanismo para identificar os casos de teste que não revelam faltas de regressão.

É importante observar a precisão das técnicas, pois uma técnica capaz de não selecionar casos de teste que não revelem faltas de regressão, é uma técnica que melhor utiliza os

recursos alocados para a execução e análise dos casos de teste de regressão. Em um teste de regressão, um dos principais objetivos é capturar as faltas de regressão, portanto, não é recomendado executar casos de teste que não capturam faltas de regressão.

Eficiência

A eficiência está relacionada com a quantidade de tempo que a técnica necessita para executar. Geralmente, a eficiência é observada comparando o custo de execução da técnica com o custo da técnica de *retest-all*, onde todos os casos de teste da suíte de regressão são executados. A eficiência pode ser medida através de modelos de custos, determinados pelo investigador do experimento [Graves et al. 1998, Graves et al. 2001]. Alguns destes modelos consideram o tempo de execução dos casos de teste, o tempo da análise dos resultados dos casos de testes executados, assim como, o próprio tempo utilizado pela técnica para selecionar os casos de teste.

Neste estudo experimental, uma vez que os casos de teste são abstratos, ou seja, não são executáveis automaticamente, não há a disponibilidade de tempo de executar todas as suítes selecionadas para cada execução do estudo experimental. Portanto, será considerado apenas o tempo de execução da técnica de re-teste seletivo. Dessa forma, as técnicas que necessitarem de menos tempo para realizar a seleção dos casos de teste de regressão são as que possuem uma melhor eficiência.

Potencial de Redução

O potencial de redução de uma técnica representa a porcentagem de redução da suíte de testes de regressão, ou seja, a quantidade de casos de teste de regressão que não foram selecionados. Diante disto, e a partir do exemplo da Figura 2.3, podemos definir o potencial de redução $R(T_x)$ da técnica da seguinte forma:

$$R(T_x) = 100 \times \frac{|\overline{T'}|}{|T|}. \tag{2.3}$$

Utilizando os dados do próprio exemplo, podemos calcular o seguinte potencial de redução para a técnica da Figura 2.3:

$$\begin{aligned} R(T_x) &= 100 \times \frac{|\overline{T'}|}{|T|} \\ &= 100 \times \frac{20}{50} \\ &= 40\%. \end{aligned}$$

Dessa forma, podemos observar que a técnica do exemplo apresenta um potencial de redução de 40%, ou seja, ela é capaz de remover cerca de 40% da suíte de testes de regressão. O potencial de redução é observado em situações onde a suíte de testes é muito grande, e todos os casos de teste não podem ser executados devido à falta de recursos. Esta situação é comum durante o teste de regressão devido ao tamanho da suíte de testes de regressão.

Densidade de Faltas

A densidade de faltas de uma técnica é utilizada para observar a porcentagem de casos de teste, na suíte selecionada, que capturam pelo menos uma falta de regressão. Uma técnica apresenta uma alta densidade de faltas quando a suíte que esta seleciona possui uma grande proporção de casos de teste que capturam faltas de regressão. Portanto, definimos a densidade de faltas $D(T_x)$ da seguinte forma:

$$D(T_x) = 100 \times \frac{|T'_F|}{|T'|}. \tag{2.4}$$

Utilizando o exemplo da Figura 2.3, podemos obter a seguinte densidade de faltas:

$$\begin{aligned} D(T_x) &= 100 \times \frac{|T'_F|}{|T'|} \\ &= 100 \times \frac{10}{30} \\ &= 34\%. \end{aligned}$$

Portanto, 34% dos casos de teste da suíte selecionada, pela técnica do exemplo, capturam faltas de regressão. Esta propriedade é utilizada para relacionar os casos de teste que capturam faltas de regressão com o tamanho da suíte selecionada pela técnica. A principal

diferença entre a densidade de faltas e a inclusão está no denominador das equações 2.4 e 2.1, respectivamente. Enquanto que a inclusão relaciona T'_F com o total de casos de teste que capturam faltas de regressão (T_F), a densidade de faltas observa T'_F com a quantidade de casos de teste selecionados (T').

É importante observar a densidade de faltas, pois algumas técnicas apresentam uma alta inclusão, no entanto, não são capazes de reduzir, significativamente, o tamanho da suíte de testes de regressão. Ou seja, algumas técnicas apresentam uma alta cobertura dos casos de teste que capturam faltas de regressão, porém esta cobertura é observada devido à baixa redução da suíte de regressão. Dessa forma, a cobertura de faltas de regressão permanece grande, mas não é obtida uma redução no custo do teste de regressão, pois ainda são selecionados muitos casos de teste para a execução. Diante disto, desejamos observar as técnicas que mantêm uma boa cobertura dos casos de teste que capturam faltas de regressão, mesmo com uma grande, ou pequena, redução da suíte de testes de regressão.

Generalidade

A generalidade de uma técnica de re-teste seletivo está relacionada com a sua habilidade de funcionar em diversas situações. Estas situações variam de acordo com o contexto do teste de regressão (código ou especificação) e/ou a natureza e freqüência das modificações realizadas [Rothermel and Harrold 1996]. Por exemplo, no contexto de código, a generalidade poderia estar relacionada com a habilidade de lidar com blocos condicionais, laços, ou outros aspectos do código-fonte da aplicação. Para analisar a generalidade de uma técnica, devem ser definidos, inicialmente, os critérios que a caracterizam.

Uma vez que estes critérios são estabelecidos, o investigador deve observar quais critérios são contemplados pelas técnicas. Exemplos de critérios são as limitações para a execução da técnica (e.g. requisitos mínimos de hardware), ou dependências ferramentais ou operacionais (e.g. quantidade de testadores, ou desenvolvedores, dentre outros). Estes critérios podem ser estruturados para pesar de forma positiva ou negativa na generalidade da técnica.

A análise de generalidade geralmente é baseada em aspectos qualitativos, dificultando a realização de uma análise baseada em recursos estatísticos, como média e desvio-padrão [Rothermel and Harrold 1996]. No entanto, a generalidade da técnica é uma propriedade amplamente observada em estudos experimentais de técnicas de re-teste seletivo, pois fornecem

uma perspectiva a respeito da aplicabilidade da técnica, por exemplo, na indústria ou em empresas.

2.5 Fundamentos de Experimentação em Engenharia de Software

Engenharia de Software é uma área multi-disciplinar. Ela aborda aspectos técnicos (banco de dados, sistemas operacionais, linguagens de programação), bem como aspectos humanos (papéis e responsabilidades em um processo de desenvolvimento, técnicas de motivação e gerenciamento de equipe). Diante disso, desenvolver pesquisas na área de engenharia de software exige a definição de um método científico compatível com os objetivos da pesquisa [Wohlin et al. 2000]. Um dos métodos científicos é o método empírico. O método empírico utiliza estudos empíricos para avaliar modelos propostos. Um estudo empírico pode ser realizado através de um *survey*, estudo de caso, ou um experimento.

O *survey* é um estudo empírico realizado com o objetivo de avaliar e entender uma população, da qual uma amostra foi retirada [Babbie 1990]. Esta avaliação ocorre, geralmente, através de formulários, questionários e entrevistas. Diante disso, um *survey* foca em objetivos como explicar, explorar e descrever uma população.

O estudo de caso, por sua vez, é utilizado para investigar um fenômeno em um intervalo de tempo específico. Este fenômeno, pode ser a adoção de uma tecnologia, ou metodologia da Engenharia de Software. Dessa forma, os estudos de caso são adequados para uma avaliação industrial [Yin 1994]. Estudos de caso podem ser utilizados para um estudo comparativo, no entanto, o resultado é proveniente de uma observação das variáveis usadas na pesquisa em um intervalo de tempo específico, dificultando a obtenção de informação mais precisa e completa, sob uma perspectiva do comportamento geral das variáveis.

Um estudo experimental é realizado com o objetivo de observar uma relação causa-efeito entre aspectos da teoria (Figura 2.4). Um exemplo de causa, seria uma linguagem de programação e o objetivo seria observar o efeito do uso desta linguagem no desenvolvimento de um software. Portanto, estes elementos da teoria (causa e efeito) são estruturados em elementos observáveis, ou seja, elementos do estudo experimental.

A causa que desejamos observar, é interpretada através de variáveis independentes, fa-

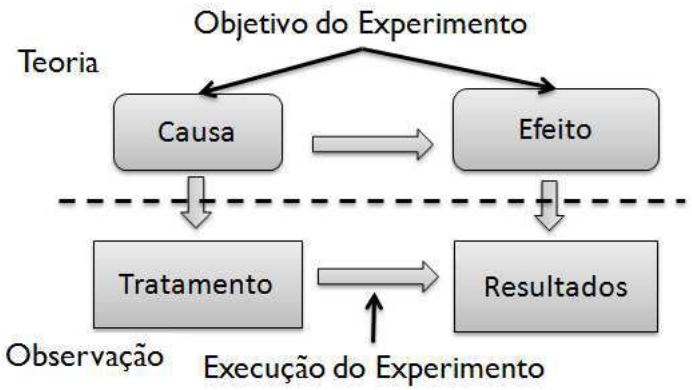

Figura 2.4: Princípios de um experimento.

tores e níveis do experimento. O efeito, por sua vez, é interpretado a partir de variáveis dependentes do estudo experimental. O objetivo é observar a relação causa-efeito através da execução do experimento.

Experimentos são realizados para obter informações precisas acerca de um conjunto de variáveis. Para obter estas informações, é necessário um alto nível de controle sobre o ambiente e a execução do experimento. Além do controle, outra característica do experimento é a capacidade de repetição. Um experimento deve ser realizado de forma que ele possa ser repetido, e produzir os mesmos resultados.

Para melhor estruturar os elementos do experimento, Wohlin et. al [Wohlin et al. 2000] descreve um processo para realização de estudos experimentais. Este processo estrutura os elementos significativos do experimento possibilitando o foco no controle e nas variáveis importantes para a obtenção dos resultados.

O processo de um experimento é dividido em cinco etapas: a definição, planejamento, operação do experimento, análise e interpretação dos resultados, e a apresentação dos resultados. Por fim, o estudo experimental é avaliado através de sua validade. São estabelecidos os métodos de tratamento das ameaças à validade do experimento, de acordo com o tipo da validade analisada (conclusão, construção, interna e externa). Cada uma destas etapas do estudo experimental assim como os aspectos de avaliação de validade serão descritos nas subseções a seguir.

2.5.1 Definição do Experimento

Nesta etapa, são definidos elementos que caracterizam o estudo experimental. Estes elementos são: o objetivo, a hipótese informal, e os objetos de estudo. Estes elementos podem ser definidos através do seguinte *template*:

Analisar ***<objetos de estudo>***
com o propósito de ***<propósito>***
com respeito ao ***<foco de qualidade>***
do ponto de vista da ***<perspectiva>***
no contexto de ***<contexto>***

Neste *template* os seguintes elementos são especificados:

- **Objetos de estudo:** Os objetos de estudo[1] são as entidades estudadas através do estudo experimental. Na Engenharia de Software, exemplos comuns de objetos de estudos são: processos de desenvolvimento, modelos de confiabilidade, técnicas para melhorar o processo de desenvolvimento, dentre outros [Wohlin et al. 2000].

- **Propósito:** O propósito define qual o objetivo ao realizar o estudo experimental. Exemplos destes objetivos são: investigação, comparação, caracterização de uma curva de aprendizado, etc.

- **Foco da qualidade:** Esta característica reflete os efeitos sob estudo, ou seja, as características observadas pelo estudo experimental. Durante o estudo experimental, os elementos do foco de qualidade são referenciados como as variáveis dependentes do experimento.

- **Perspectiva:** Este elemento define o ponto de vista do qual os resultados do experimento serão interpretados. Exemplos de perspectiva, em engenharia de software são: desenvolvedores, gerentes, testadores, arquitetos, dentre outros.

[1] Não devem ser confundidos com os objetos do experimento. Objetos de estudo, e objetos do experimento são duas coisas distintas [Wohlin et al. 2000].

- **Contexto:** O contexto define o ambiente no qual o estudo experimental é executado. Este ambiente é caracterizado através das pessoas envolvidas (sujeitos do experimento), e dos artefatos utilizados (objetos). Em engenharia de software, os contextos podem ser caracterizados pelas etapas do ciclo de vida de um software (e.g. projeto, desenvolvimento, manutenção, testes, dentre outros).

Após definir os elementos, o contexto deve ser especificado, ou seja, é necessário definir quais os sujeitos e objetos utilizados no estudo experimental. Para os sujeitos, é necessário especificar a quantidade, as prioridades, o nível de experiência, e as atividades realizadas. Os objetos são definidos em função da quantidade, tamanho, complexidade, e domínio da aplicação (e.g. domínio de especificação, código, aplicações móveis, etc.) [Wohlin et al. 2000].

Uma vez que os elementos que caracterizam o experimento são definidos, é necessário planejar as atividades e como os recursos do experimento (objetos e sujeitos) serão utilizados. Estes aspectos são especificados durante a etapa de planejamento, descrita na próxima subseção.

2.5.2 Planejamento do Experimento

A partir dos elementos definidos para o estudo experimental, é realizado o planejamento das atividades e recursos do experimento. Wohlin et al. [Wohlin et al. 2000] sugere um planejamento para o estudo experimental que contempla as principais características do estudo, como a instrumentação, a execução e a análise estatística realizadas. Este planejamento é descrito nesta subseção.

O primeiro aspecto planejado é a seleção do contexto, mais especificamente, a seleção dos sujeitos que irão participar no experimento. A seleção do contexto é definida a partir de parâmetros que indicam o modo em que a pesquisa será realizada para manter o realismo, bem como, a viabilidade do experimento.

Após esta seleção, é necessário caracterizar as variáveis dependentes e independentes do experimento. As variáveis independentes são as variáveis que serão controladas e modificadas durante a execução do experimento. Por sua vez, as variáveis dependentes são obtidas como resultados da execução do experimento, ou seja, são as variáveis que se deseja observar.

Uma, ou mais, das variáveis independentes caracterizam o(s) fator(es) do estudo experimental. Durante a execução do experimento, os fatores são permutados com o objetivo de verificar o efeito desta permutação nas variáveis dependentes. Os fatores são caracterizados pelos respectivos tratamentos (ou níveis). Os níveis e fatores devem ser definidos a partir dos objetos de estudo, pois são eles que devem ser observados através das variáveis dependentes.

As variáveis dependentes são aquelas que desejamos observar. É investigado o efeito das variáveis independentes (fatores) nas variáveis dependentes durante a execução do estudo experimental. As hipóteses nulas e alternativas analisadas são estruturadas a partir das veriáveis dependentes. Estas hipóteses são submetidas aos recursos estatísticos (e.g. testes estatísticos, análise de intervalos de confiança, dentre outros), para fornecer os resultados da execução.

Para definir os recursos estatísticos utilizados, assim como estabelecer os aspectos de execução como as replicações ou a variação nos fatores e níveis, são definidos os projetos (*design*) experimentais. O projeto experimental é estabelecido a partir das características do experimento, como a quantidade de objetos, sujeitos, fatores e níveis [Wohlin et al. 2000, Jain 1991].

Além destes elementos, durante o planejamento são definidos aspectos da instrumentação do experimento, como a definição de ferramentas utilizadas e entidades implementadas. Os objetos utilizados também devem ser especificados nesta etapa. A última etapa do planejamento é definir como será analisada a validade do experimento. Diante disto, as ameaças à validade devem ser apontadas e os métodos de como estas ameaças serão controladas devem ser também especificados.

É comum que outras ameaças sejam encontradas durante as demais etapas do experimento. No entanto, é necessário, durante a apresentação dos resultados, reportar as novas ameaças e como estas foram controladas. Também é recomendado, quando possível, relacionar estas novas ameaças com aquelas observadas no planejamento. Os aspectos de validade do estudo experimental é apresentado na próxima subseção [Wohlin et al. 2000].

2.5.3 Avaliação de validade

A avaliação de validade é realizada para verificar a validade dos resultados e conclusões obtidas pelo estudo experimental. Esta característica deve ser estabelecida durante o plane-

jamento evitando que a avaliação de validade seja estruturada a partir dos resultados obtidos. Uma boa validade é definida a partir de dois aspectos: a relação entre os resultados com a amostra utilizada no experimento, e a capacidade de generalização dos resultados.

Em algumas situações, a generalização dos resultados não é atingida, por limitações do cenário em que o experimento é realizado (e.g. quantidade de objetos, sujeitos, representatividade da amostra disponível, dentre outros). Dessa forma, a validade é analisada dentro do escopo limitado pela generalização. Ou seja, não é necessário comparar a validade do experimento, com a generalização obtida [Wohlin et al. 2000]. Um exemplo é a realização de experimentos dentro de empresas, em que a comparação entre um processo de desenvolvimento de software é avaliada com relação ao ambiente da empresa, e não à população (todas as técnicas de desenvolvimento de software).

A avaliação de validade pode ser caracterizada de acordo com a validade analisada. As validades podem ser de: conclusão, construção, externa e interna. As ameaças a cada tipo de validade deve ser identificada e controlada para não comprometer a validade dos resultados obtidos, e portanto, a validade do estudo experimental realizado.

A validade de conclusão se preocupa em manter a validade entre cada tratamento (nível) do experimento e os respectivos resultados obtidos. Ou seja, é necessário manter uma relação estatisticamente significativa entre a execução do tratamento e os resultados. Dessa forma, as conclusões não são conseqüências de erros experimentais, como medições inadequadas, ou erros na instrumentação, e.g. a implementação ou o ambiente de execução do experimento.

A validade de construção se preocupa com o relacionamento entre os elementos da teoria, e os elementos da observação estabelecidos na definição do experimento. Em outras palavras, é necessário que a construção do experimento (a transição entre os elementos da teoria e os elementos operacionais, e.g. a implementação das técnicas) sejam válidos.

A validade externa se preocupa com a generalização do experimento. Diante disto, esta validade está diretamente relacionada com o design experimental, os sujeitos, os objetos e a instrumentação do experimento.

A validade interna se preocupa com a relação entre os tratamentos e os resultados obtidos. É necessário que os resultados obtidos sejam conseqüências da execução do tratamento em si, e não de um outro fator, ou aspecto operacional que não foi controlado. O principal aspecto que deve ser observado para evitar ameaças à validade interna, é o controle do

experimento. É necessário controlar a captura e armazenamento dos dados, a atuação dos sujeitos, e o objeto utilizado na execução, para evitar que erros em algum destes elementos, i.e. ações ou execuções inesperadas, afetem o dado observado (neste caso, uma ou mais variáveis dependentes).

Algumas vezes aspectos do planejamento devem ser modificados, devido às mudanças encontradas durante a execução, ou a etapa de análise. Este processo não segue um modelo cascata, e é possível revisar os aspectos de planejamento de acordo com o desenvolvimento das demais etapas [Wohlin et al. 2000]. Uma vez que os elementos do experimento foram planejados, é possível iniciar a etapa de operação. Esta etapa é descrita na próxima subseção.

2.5.4 Etapa operacional

Durante a etapa operacional é realizada a instrumentação e a execução do estudo experimental. A instrumentação contempla as atividades de implementação dos elementos da teoria. Através desta implementação, é possível executar o experimento e coletar os dados desta execução. As atividades de instrumentação contemplam a instalação do suporte ferramental ao experimento (e.g. ambiente de desenvolvimento, ferramentas para coleta e/ou análise dos dados, dentre outros), a implementação dos objetos de estudos, e a aplicação dos tratamentos aos sujeitos.

É essencial que os sujeitos possuam as informações necessárias para sua participação no experimento. Além disto, a motivação e disposição destes sujeitos podem afetar os resultados do estudo experimental. É importante fornecer a documentação adequada para o envolvimento dos sujeitos no estudo experimental, por exemplo, formulários, ferramentas, tutoriais, dentre outros [Wohlin et al. 2000]. A participação dos sujeitos pode caracterizar uma ameaça à validade interna quando não é devidamente controlada.

Uma vez que a instrumentação é concluída, é possível iniciar a execução do experimento. Ou seja, os tratamentos são executados, e os dados referentes à esta execução são coletados para obter o efeito dos diversos tratamentos nas variáveis dependentes. Durante esta execução, aspectos como não-determinismo e paralelismos devem ser controlados para que os resultados dos tratamentos sejam devidamente coletados, evitando erros de medição. Uma vez que a execução é concluída os dados coletados são submetidos à etapa de análise.

2.5.5 Etapa de análise

As conclusões do estudo experimental são obtidas a partir da análise realizada nos dados coletados durante a execução. Esta análise deve ser realizada de acordo com o projeto experimental especificado durante o planejamento. Os principais objetos da análise são as hipóteses nulas e alternativas. Geralmente, estas hipóteses são estruturadas com o objetivo de rejeitar as hipóteses nulas, através de testes estatísticos como a Análise de Variância (ANOVA), ou teste t de Student [Siegel and Junior 1988, Jain 1991, Wohlin et al. 2000]. Os resultados destes testes são interpretados com o objetivo de obter as conclusões do estudo experimental.

É importante realizar a análise de acordo com o propósito do estudo experimental, estabelecido durante a definição do experimento (e.g. comparação, investigação, dentre outros). Ainda sob esta perspectiva, é necessário utilizar os recursos estatísticos adequados. A utilização inadequada de um teste estatístico caracteriza uma forte ameaça à validade do experimento. Dessa forma, é essencial realizar as verificações das premissas necessárias (e.g. testes de normalidade, análise de resíduos) para aplicar os testes e recursos estatísticos.

Durante a análise, é necessário observar os aspectos de generalização das conclusões obtidas. Deve ser investigado se os dados obtidos são suficientes para representar a população da qual a amostra foi retirada. Além disto, os recursos estatísticos utilizados podem fornecer uma perspectiva da representatividade dos dados, por exemplo, através dos testes paramétricos. Todos os aspectos obtidos durante a análise (e.g. testes, gráficos, conclusões, modelos) devem ser estruturados na etapa de apresentação do experimento.

2.5.6 Etapa de apresentação

O objetivo da etapa de apresentação é organizar os elementos do estudo experimental em um relatório. Exemplos de elementos apresentados no relatório são: a metodologia (i.e. a descrição de como as etapas foram realizadas no experimento), os recursos estatísticos utilizados, gráficos apresentando os comportamentos observados nos dados, as conclusões obtidas, dentre outros.

Alguns elementos podem não ser incluídos no relatório, por motivos de espaço no relatório, propriedade intelectual (e.g. dados coletados, descrição das ferramentas utilizadas),

ou princípios éticos (identidade e documentação referente à participação dos sujeitos). No entanto, é importante fornecer a descrição destes elementos para outros investigadores avaliando, reproduzindo ou melhorando o experimento.

2.6 Fundamentos em análise estatística

Diversos métodos e técnicas são propostas na área de Engenharia de Software, tanto no contexto acadêmico quanto industrial. Um dos recursos utilizados para avaliar os dados obtidos através da utilização destes métodos e técnicas, é a análise estatística. Através de recursos estatísticos, os dados obtidos podem ser interpretados com o objetivo de gerar conclusões a respeito do desempenho da técnica ou método analisado.

Os recursos estatísticos são os elementos fundamentais da análise de um estudo experimental [Wohlin et al. 2000]. Estes recursos são utilizados para obter conclusões a partir das hipóteses levantadas durante a definição e o planejamento do expeirmento. O teste de hipóteses, ou o teste visual de intervalos de confiança, por exemplo, são recursos utilizados para rejeitar as hipóteses nulas do experimento.

Estes recursos estatístico são fundamentados em dois elementos descritivos dos dados: a média e a variância. A média é uma medida de tendência central que pode ser interpretada como uma estimativa a respeito da uma variável estocástica. Outros exemplos de medidas de tendência central são a mediana e a moda. A mediana representa o valor que divide o conjunto de dados pela metade, enquanto que a moda representa o observado com maior frequência no conjunto.

A variância, por sua vez, é um índice de dispersão dos dados. A partir da variância, é possível observar a variabilidade dos dados. A média e a variância são utilizadas para verificar o comportamento dos dados, possibilitando, por exemplo, a estimativa de valores através de modelos de regressão. Além disto, podemos investigar a distribuição dos dados para melhorar estes modelos de predição, e melhorar a comparação com outros fenômenos já analisados.

Uma das principais vantagens da utilização de recursos estatísticos é a significância e representatividade das conclusões obtidas [Jain 1991, Wohlin et al. 2000]. Diante disto, alguns elementos da análise estatística serão discutidos nesta seção, com o objetivo de fornecer uma

fundamentação a respeito da análise do estudo experimental realizado neste trabalho. Serão apresentados alguns conceitos acerca de testes visuais, e o teste de hipóteses, contemplando a diferença entre testes paramétricos e não-paramétricos, e alguns exemplos de cada tipo de teste.

2.6.1 Testes visuais

Algumas conclusões podem ser obtidas através da observação dos dados. Através da amostra obtida durante a execução do experimento, é possível gerar gráficos que exibem tendências e comportamentos acerca dos dados. Através destes gráficos, podem ser realizadas comparações de desempenhos, ou investigações acerca de dependência entre fatores analisados. Alguns exemplos de gráficos (Figura 2.5) utilizados nestes testes são: gráficos de dispersão, histogramas, e intervalos de confiança.

Os gráficos de dispersão são bons para aferir dependências entre tratamentos analisados. Neste gráfico, cada eixo representa um tratamento, ou variável, analisada, e cada dado é observado como um ponto no gráfico. Observando o espalhamento e concentração destes pontos (dados) é possível prever relações lineares entre tratamentos, assim como identificar valores atípicos (*outliers*) encontrados na amostra. Um exemplo de gráfico de dispersão é apresentado na Figura 2.5 (a).

O histograma, por sua vez, é utilizado para observar informações a respeito de um tratamento analisado. Neste gráfico, é possível observar a frequência (eixo-y) dos dados da amostra (eixo-x). Desta forma, são observados aspectos como os valores mais frequentes e a densidade da distribuição da amostra. O gráfico apresentado na Figura 2.5 (b) é um exemplo de histograma.

Enquanto que o gráfico de dispersão apresenta informações sobre os dados dos tratamentos, e o histograma apresenta a densidade da distribuição da amostra, o intervalo de confiança (IC) fornece informações sobre os dados da amostra e da população. Na realização de experimento, o objetivo é obter conclusões a respeito de uma população. Dessa forma, executamos o experimento a partir de amostras desta população. O IC representa um intervalo, construído através de um nível de confiança, para a estimativa da média populacional a partir dos dados da amostra.

Diante disto, os IC de duas ou mais variáveis (cada uma com sua respectiva amostra),

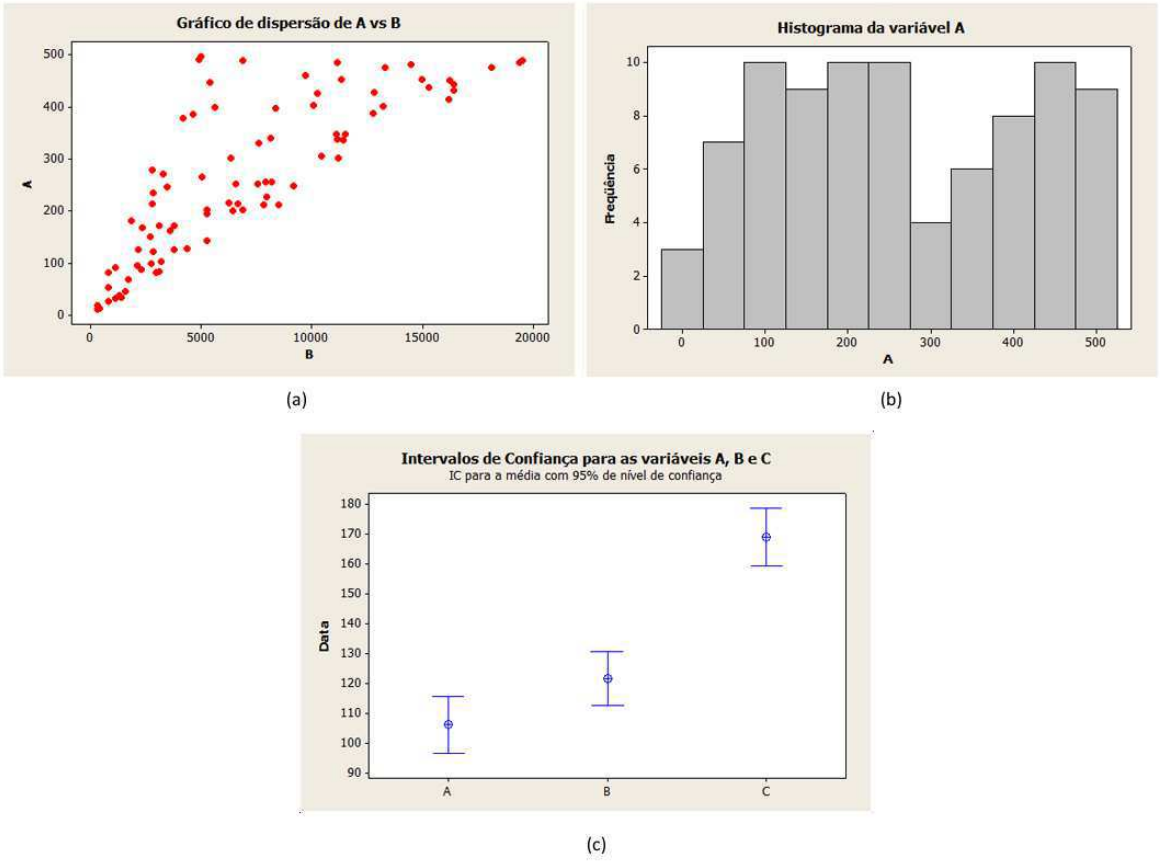

Figura 2.5: Exemplos de gráficos utilizados em testes visuais. (a) Gráfico de dispersão. (b) Histograma. (c) Intervalos de Confiança.

podem ser utilizados para verificar se estas variáveis são de uma mesma população. A partir do teste visual realizado com IC (e.g. Figura 2.5 (c)), é possível observar o desempenho relativo entre os níveis observados, de acordo com a disposição dos intervalos no gráfico.

Por outro lado, quando os intervalos dos tratamentos se sobrepõem, não é possível obter conclusões estatisticametne significativas a respeito do desempenho comparativo desejado. Portanto, outros recursos estatísticos devem ser utilizados, como por exemplo, os testes de hipótese.

2.6.2 Testes de hipótese

Os testes visuais são adequados para fornecer informações acerca de comparações, tendências e comportamentos dos tratamentos ou variáveis. No entanto, há situações em que o teste

visual não apresenta força estatística para fornecer conclusões significativas. Um exemplo comum, é a sobreposição de intervalos de confiança. Em algumas situações, os IC são muito semelhantes e não é possível obter conclusões precisas a respeito das hipóteses do experimento.

Nestas situações, é necessária a utilização de outros recursos estatísticos para obter conclusões, com maior confiança estatística, sobre as hipóteses. Um destes recursos são os testes de hipóteses. A partir dos dados e informações da amostra (e.g. distribuição e resíduos), é possível aplicar testes capazes de fornecer um resultado binário (rejeitar ou não rejeitar) a respeito das hipóteses. É importante lembrar que os recursos da análise estatística estão relacionados com aspectos de probabilidade, e que os resultados obtidos são dependentes de um nível de confiança, ou significância, aplicado no teste [Jain 1991].

Existem diversos testes de hipóteses na literatura de estatística, por exemplo o teste t de Student, ANOVA, Bonferroni, Tukey, Wilcoxon, dentre outros [Lilja 2000, Jain 1991, Kvam and Vidakovic 2007]. Cada teste é utilizado de acordo com a configuração do projeto experimental estabelecida (e.g. quantidade de fatores, níveis, tamanhos de amostras, dentre outros). É importante que as premissas dos testes sejam devidamente analisadas, pois a utilização de um teste errado, compromete significativamente a validade do estudo realizado [Wohlin et al. 2000, Jain 1991, Kvam and Vidakovic 2007].

Os testes de hipóteses podem ser paramétricos ou não-paramétricos. Os testes paramétricos são baseados em uma distribuição específica da amostra (e.g. distribuição normal padrão), verificada através dos testes para investigar distribuições (e.g. testes de Anderson-Darling, ou Kolmogorov-Smirnov) [Jain 1991, Kvam and Vidakovic 2007]. Portanto, são utilizados parâmetros como a média e desvio-padrão destas distribuições existentes para a realização dos cálculos e obtenção de resultados. Dessa forma, os testes paramétricos estão associados a um conjunto de suposições (*assumptions*) e premissas que devem ser respeitadas para que estes testes possam ser utilizados.

Os testes não-paramétricos, por sua vez, não se baseiam em informações como dsitribuições, ou média e variância de distribuições semelhantes a da amostra obtida. Dessa forma, os testes não-paramétricos não apresentam o mesmo rigor com relação às suposições e premissas necessárias para a realização do teste, quando comparados aos testes paramétricos.

Apesar das exigências para sua utilização, os testes paramétricos são capazes de gerar conclusões com significância estatística mais alta do que as obtidas através de um teste não-paramétrico. Alguns pesquisadores argumentam sobre a utilização de cada tipo de teste, pois apesar de possuir um poder estatístico maior, a análise de premissas de um teste paramétrico pode requerer muito esforço do pesquisador, inviabilizando sua utilização. Além disto, diversas destas premissas são sensíveis aos elementos dados (e.g. erros de medição, *outliers*, tamanho da amostra e da população, dentre outros), o que pode invalidar a própria análise de premissas realizada pelo investigador [Wohlin et al. 2000].

Para a escolha adequada do teste, deve ser observado, inicialmente, os projetos experimentais estabelecidos para o experimento. A partir desta informação, é realizada uma investigação para escolher o teste estatístico. Em algumas situações, é escolhido um teste paramétrico, e então, a análise de suas premissas é realizada. Quando as premissas não são respeitadas pelos dados da amostra, é necessário escolher um teste não-paramétrico correspondente ao teste paramétrico escolhido. Exemplos de testes paramétricos e não paramétricos são apresentados na Tabela 2.1.

Tabela 2.1: Exemplos de testes estatísticos para diversas configurações de projetos experimentais.

Projeto experimental	**Teste paramétrico**	**Teste não-paramétrico**
Comparação de médias	Tukey	Dunn
Um fator, dois níveis	teste t de Student	Mann-Whitney
Um fator, dois níveis (observações pareadas)	teste t de Student (pareado)	Wilcoxon
Um fator, mais de dois níveis	*One-way* ANOVA	Kruskal-Wallis
Mais de um fator	ANOVA Fatorial	Teste de Friedman

O resultado de um teste estatístico (paramétrico, ou não-paramétrico), é o p (também conhecido como p-valor ou valor p de um teste estatístico). Este valor é comparado com o nível de significância (α) do teste para determinar se é possível, sob o nível de confiança estabelecido, rejeitar a hipótese nula. A partir deste resultado, é possível caracterizar a decisão do teste, obtendo as conclusões do estudo experimental.

A estrutura das hipóteses (nulas e alternativas) assim como o nível de confiança estab-

elecidos e os testes estatísticos utilizados (testes visuais, testes paramétricos, ou testes não paramétricos) são definidos a partir do projeto experimental. Portanto, é necessário utilizar os elementos adequados para cada projeto experimental realizado, evitando as ameaças à validade de conclusão do estudo.

2.7 Considerações Finais do Capítulo

Os aspectos discutidos neste capítulo contemplam a fundamentação teórica utilizada neste trabalho. Foram apresentados alguns fundamentos em teste de software, estudos experimentais em Engenharia de Software, e análise estatística. Outros elementos da fundamentação teórica do trabalho são as técnicas de re-teste seletivo analisadas no experimento.

No próximo capítulo serão apresentadas as técnicas propostas na literatura que foram analisadas no estudo experimental, assim como, a técnica WSA, que inspirou a técnica proposta neste trabalho (WSA-RT). Esta última, por sua vez, é apresentada no Capítulo 4.

Capítulo 3

Técnicas de Re-teste Seletivo

Diversas técnicas de re-teste seletivo foram propostas na literatura, em especial técnicas baseadas no código [Korel et al. 2002]. O foco deste trabalho é o re-teste seletivo baseado em especificação com a abordagem de Teste baseado em Modelos. Portanto, as técnicas utilizadas neste trabalho executam em elementos do modelo da especificação. Este capítulo apresenta algumas técnicas da literatura de re-teste seletivo baseado em especificação e seleção de casos de teste em TBM. É importante lembrar que o objetivo destas técnicas é reduzir a quantidade de casos de teste em uma suíte.

Diante disto, o procedimento realizado por cada técnica no processo de seleção do subconjunto da suíte de teste é apresentado na respectiva seção da técnica. Foram contempladas as técnicas de seleção baseada em análise de dependência (Seção 3.1), seleção baseada em análise de riscos e diagrama de atividades (Seção 3.2); re-teste baseado em perfis (Seção 3.3); *Weighted Similarity Approach* (Seção 3.4); seleção baseada em *clusters* (Seção 3.5); e a seleção aleatória de casos de teste (Seção 3.6).

3.1 Técnica baseada em Análise de Dependência

A técnica de seleção baseada em análise de dependência, proposta por Korel et al. [Korel et al. 2002], realiza a seleção dos casos de teste de regressão observando os diversos tipos de modificações realizadas, assim como os elementos do modelo que possuem alguma dependência com a modificação. Esta técnica é amplamente referenciada em diversos trabalhos de re-teste seletivo baseado em especificação [Korel et al. 2005,

White et al. 2008, Korel and Koutsogiannakis 2009, Briand et al. 2009] por apresentar, além da própria técnica, diversos aspectos fundamentais da utilização de TBM em teste de regressão.

Chen et al. [Chen et al. 2007] apresentaram melhorias para a técnica, revisando alguns conceitos e propondo novas abordagens para a análise de dependência realizada pela técnica. Neste estudo experimental, foi utilizada a versão proposta por Chen et al. [Chen et al. 2007]. O modelo utilizado pela técnica é a máquina de estados finita estendida (MEFE). Os elementos deste modelo são descritos na subseção abaixo.

3.1.1 Máquinas de Estados Finitas Estendidas (MEFE)

Em uma MEFE, os estados e transições representam, respectivamente, os estados e passos da aplicação. Cada transição apresenta a informação de um evento e uma ação. Além disto, as transições podem conter informações a respeito das condições necessárias para a realização do evento. Uma transição é acionada quando o evento e a condição associada a este evento ocorrem durante a execução da aplicação.

Os elementos da transição (evento, ação e condição) podem estar associados às variáveis. Estas variáveis podem representar parâmetros dos eventos, ou sinalizadores para as condições (e.g. contadores para a verificação de senhas). A Figura 3.1 apresenta os elementos da transição da MEFE.

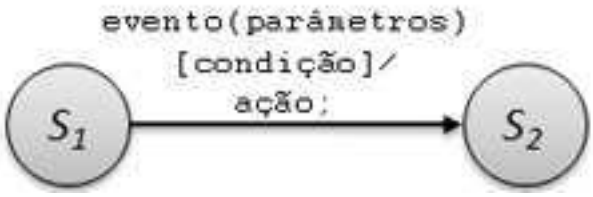

Figura 3.1: Elementos da transição de uma máquina de estados finita estendida.

Na Figura 3.2, apresentamos uma MEFE utilizada para descrever a técnica. A MEFE utilizada como exemplo representa o sistema de um caixa eletrônico, onde é possível realizar depósitos e saques de dinheiro. Este exemplo foi adaptado de Chen et al., e Korel et al. [Chen et al. 2007, Korel et al. 2002], pois é um exemplo que retrata bem a MEFE assim como apresenta os elementos de dependência observados pela técnica.

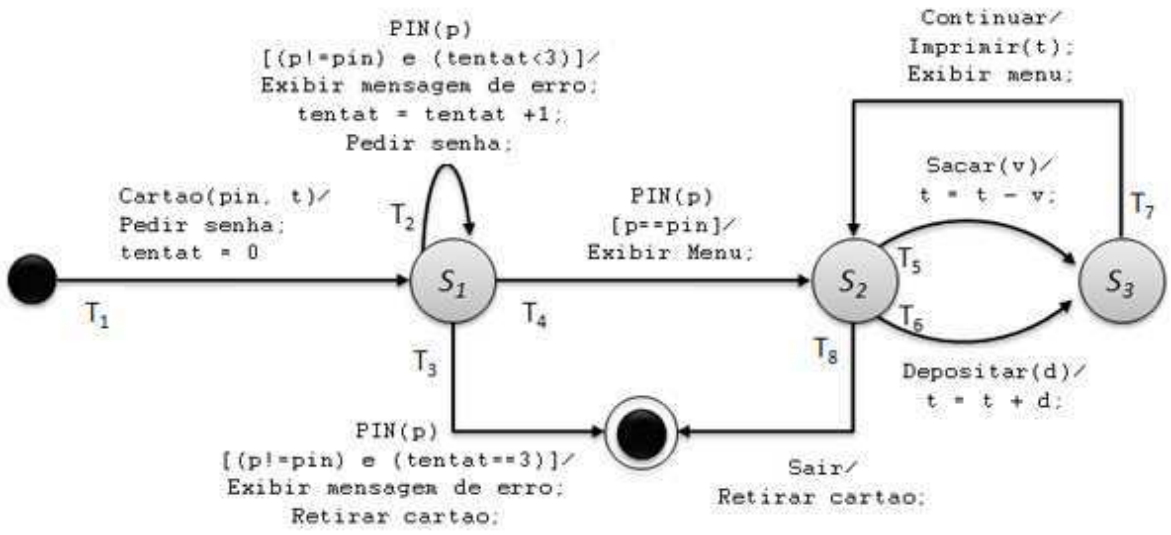

Figura 3.2: Máquina de estados finita estendida utilizada como exemplo da técnica [Adaptada de Chen et al. [Chen et al. 2007]].

3.1.2 Descrição da técnica

Para iniciar a execução, a técnica necessita da MEFE e de um arquivo contendo as modificações realizadas nesta MEFE. As modificações descritas neste arquivo devem apontar que transições ou estados foram modificados e qual o tipo da modificação, ou seja, se o estado ou transição foi: removido, adicionado ou modificado. Este último tipo de modificação é referente às alterações nos rótulos dos estados ou transições (e.g. alguma modificação de rótulo de botões na interface gráfica).

Uma vez que o modelo e as respectivas modificações são especificados, a técnica inicia a análise de dependência no modelo da versão base. A análise verifica todos os pares de transição para identificar as dependências de dados e controle. As dependências de dados são verificadas através de variáveis definidas e utilizadas nas transições do modelo. Por sua vez, as dependências de controle são identificadas através dos caminhos na MEFE, que iniciam no estado inicial e terminam no estado final.

É utilizado o conceito de pós-dominância [Korel et al. 2002, Bo 2005, Chen et al. 2007] entre as entidades do modelo para identificar as dependências de controle. As relações entre a definição e uso de uma variável são utilizadas para identificar as dependências de dados entre as transições. Diante disto, é necessário realizar diversos caminhamentos na máquina

de estados, procurando identificar, para cada transição, as respectivas dependências.

Estas informações de dependências são estruturadas em um grafo estático de dependências (GED) [Chen et al. 2007]. Neste grafo, cada vértice representa uma transição do modelo, e as transições que conectam os vértices deste grafo representam as dependências de dados e controle existente entre as transições da máquina de estados. O grafo estático de dependência obtido a partir do modelo da Figura 3.2 é apresentado na Figura 3.3. Este grafo apresenta as dependências encontradas seguindo a análise de dependências apresentada por Chen et al. [Chen et al. 2007].

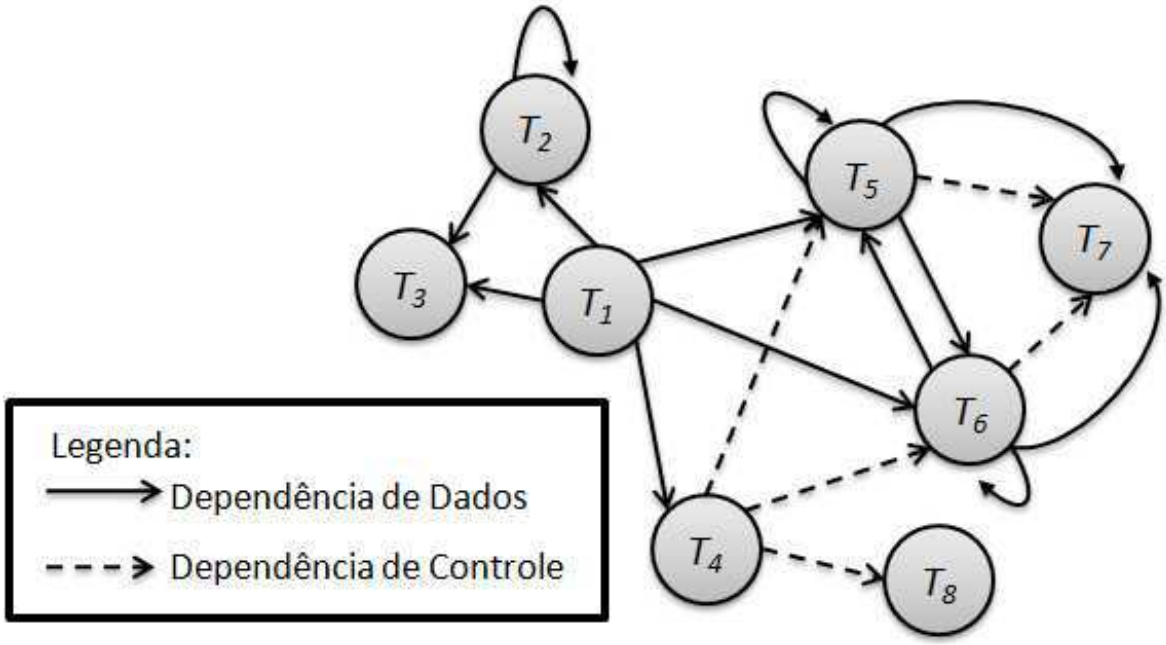

Figura 3.3: Grafo estático de dependências obtido a partir da máquina de estados do exemplo.

Após estruturar o GED do modelo da versão base, a técnica realiza as modificações no modelo para obter a máquina de estados da versão delta. Cada tipo de modificação é tratada de forma diferente para a obtenção do modelo modificado. Para ilustrar estas modificações, foram realizadas 3 modificações. São elas:

- Adição da transição T_9: Verificação de saldo na conta.
- Remoção da transição T_6: Depósito de dinheiro.
- Mudança na transição T_5: Adicionada uma condição no saque.

De acordo com a técnica, ao remover uma transição, é necessário criar um auto-laço no estado de origem da transição removida. Este auto-laço, indica que aquele estado pode apresentar uma falta por remoção. O modelo modificado pode ser visto na Figura 3.4.

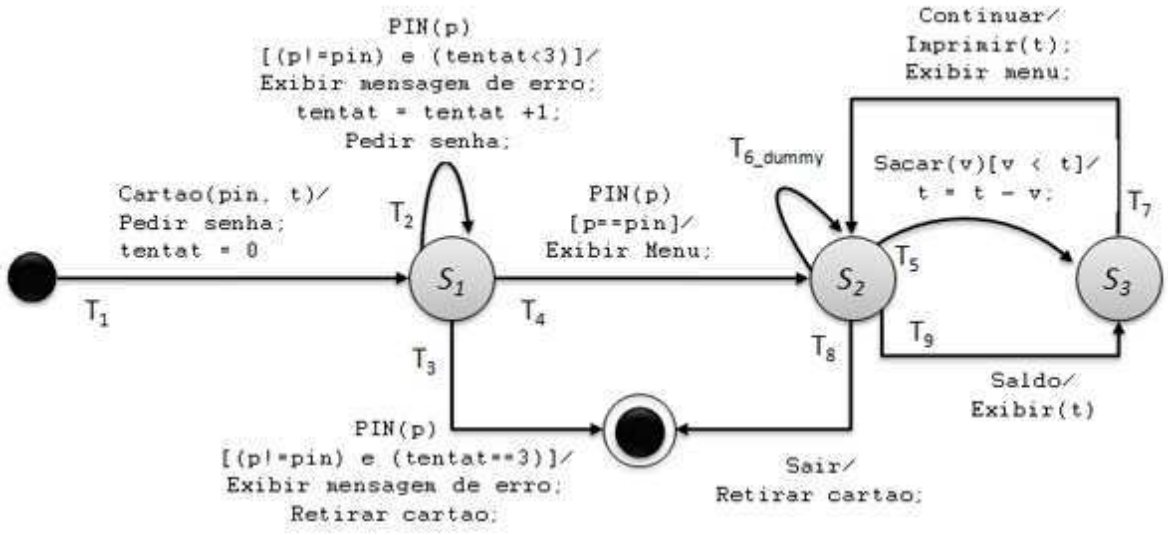

Figura 3.4: Máquina de estados com as modificações especificadas.

Após obter o modelo modificado, o próximo passo é obter o GED do modelo modificado. Portanto, uma nova análise de dependência é inicializada verificando todas as transições da máquina de estados da versão base, procurando identificar as dependências e entidades afetadas pelas modificações. O GED da versão base também é verificado para identificar as dependências por efeito colateral, ou seja, as dependências que foram removidas, ou adicionadas devido às modificações realizadas. O GED da versão delta do nosso exemplo, é apresentado na Figura 3.5

Uma vez que as MEFE e os GED estão especificados, é possível iniciar a seleção na suíte de testes de regressão. Os autores da técnica argumentam que a suíte de testes de regressão utilizada não influencia no resultado da técnica, no entanto, eles recomendam que sejam utilizados casos de teste que cobrem as funcionalidades já modificadas, por exemplo, os casos de teste que cobrem as transições adicionadas, ou que não cobrem as transições removidas durante as modificações. Também é possível utilizar uma suíte de testes de regressão gerada automaticamente a partir da máquina de estados da versão delta.

Para realizar a seleção, a técnica constrói os padrões de interações de cada caso de teste.

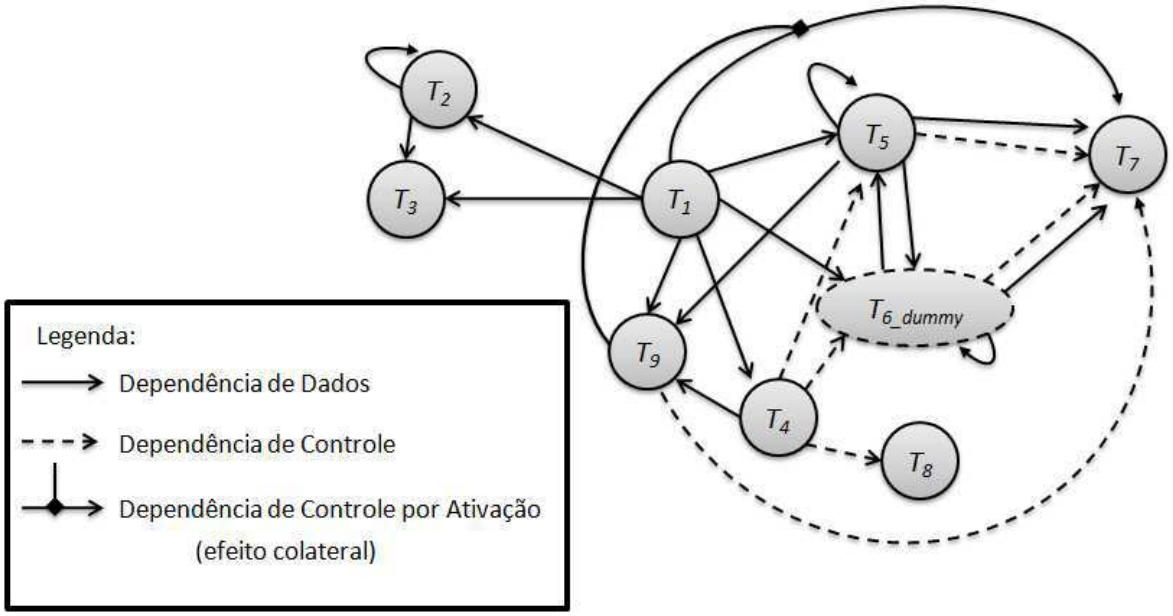

Figura 3.5: Grafo estático de dependências obtido a partir da máquina de estados com as modificações.

Os padrões de interação são subgrafos do GED modificado, e são obtidos a partir de algoritmos específicos para cada modificação realizada. Chen et al. estabelecem cerca de 15 padrões de interações para os tipos de modificações realizadas [Chen et al. 2007].

Para cada modificação, é selecionado um subconjunto da suíte de regressão, composto pelos casos de teste que exercitam a transição ou estado modificado. Então, cada caso de teste deste subconjunto é analisado, e são construídos até 3 padrões de interações para o caso de teste. Depois cada par de caso de teste deste subconjunto é verificado com o objetivo identificar os casos de teste que possuem os mesmos padrões de interação. Possuir os mesmos padrões de interação é representado pela técnica como redundância e portanto, dentre os casos de teste que possuem os mesmos padrões de interação, apenas um é escolhido para a suíte selecionada.

Após realizar a verificação de todos os subconjuntos de cada modificação, os casos de teste selecionados serão executados no teste de regressão. Foi observado, em estudos de caso reportados pelos autores da técnica, uma redução de cerca de 90% da suíte de testes de regressão ao utilizar esta técnica. No entanto, é importante considerar que os casos de teste gerados automaticamente em uma máquina de estados, geralmente, co-

brem um mesmo conjunto de transições, o que gera muita redundância na suíte de testes [Fraser and Wotawa 2007].

Além disto, há uma relação entre a quantidade de padrões de interação e o tamanho do GED, pois os padrões de interação são subgrafos do GED da versão delta, o que pode limitar a variedade de padrões de interações encontrados nos casos de teste. Dessa forma, é importante observar sob um contexto experimental, se em modelos maiores (i.e. com mais transições e estados) a técnica é capaz de atingir o mesmo desempenho observado pelos autores das técnicas.

3.2 Técnica baseada em Análise de Risco e Diagramas de Atividade

Chen et al., em seu trabalho de título "*Specification-based Regression Test Selection with Risk Analysis*", propuseram uma técnica de re-teste seletivo baseada em especificação [Chen et al. 2002]. Esta técnica realiza a seleção dos casos de teste de regressão através de uma análise de risco e diagrama de atividades. Dessa forma, são observados, durante a seleção, aspectos como as modificações realizadas no modelo, assim como, algumas informações de custo e risco fornecidas pelo testador.

Na Figura 3.6 são apresentados dois diagramas de atividades que podem ser utilizados como entrada da técnica. Estes diagramas representam o cenário de um caso de uso para a verificação de saldo bancário. A Figura 3.6 (a) apresenta o modelo da versão base, enquanto que a Figura 3.6 (b) apresenta o modelo da versão delta.

Os elementos do diagrama de atividades utilizados pela técnica são: os estados iniciais, os estados finais, as atividades, os nós de decisão, as transições e as suas eventuais guardas/condições. Cada atividade do diagrama representa um passo do caso de teste, ou um requisito do sistema, de acordo com a representação utilizada pelo testador. No nosso estudo experimental, é considerado que cada atividade representa um passo do fluxo de execução da aplicação, e os nós de decisão são utilizados para indicar a divisão do fluxo em fluxos principais e fluxos alternativos da aplicação.

Para descrever a técnica, os autores separam os casos de teste da suíte de regressão em duas categorias. Estas categorias são utilizadas para descrever o foco da seleção, ou seja,

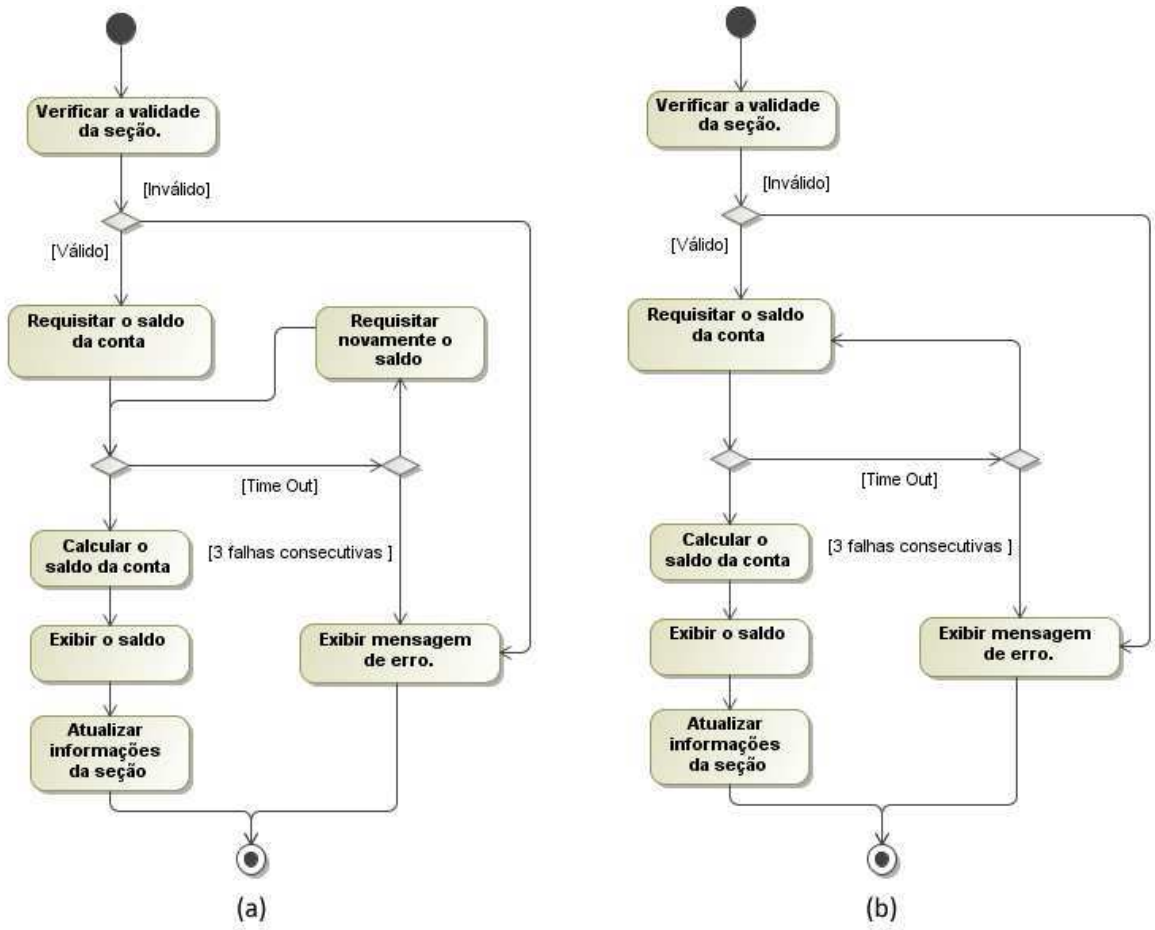

Figura 3.6: Exemplo de diagramas de atividades da versão base (a) e da versão delta (b), para um caso de uso de verificar saldo.

quais casos de teste um testador deseja selecionar para o subconjunto da suíte de regressão. Dessa forma, no trabalho, os casos de teste são caracterizados da seguinte forma:

- *Targeted Tests*: Estes casos de teste são caracterizados por exercitarem um ou mais elementos modificados do modelo.
- *Safety Tests*: Estes casos de teste cobrem os demais elementos desejados durante a seleção. Ou seja, elementos que não contemplam as modificações mas cobrem outros requisitos de interesse do testador (e.g. alto custo de execução, funcionalidades críticas, dentre outros).

A partir destas duas categorias, duas etapas de seleção são definidas para a técnica. A primeira etapa é caracterizada pela seleção dos *Targeted Tests*, enquanto que a segunda etapa,

é caracterizada pela seleção dos *Safety Tests*. As subseções a seguir apresentam a descrição de cada etapa.

3.2.1 Seleção dos *Targeted Tests*

Durante esta etapa da seleção, o objetivo é encontrar os casos de teste que exercitam os trechos modificados do modelo. Para atingir este objetivo, a técnica constrói uma matriz de rastreabilidade relacionando todas as entidades dos modelos (atividades e transições), com os casos de teste que as exercitam. Os autores descrevem a técnica utilizando, como modelo, um diagrama de atividades UML, e uma suíte de testes de regressão, gerada automaticamente a partir do modelo da versão base do software. Neste trabalho, um caso de teste é um caminho no diagrama de atividades que inicia em um estado inicial e termina em um estado final.

Após construir a matriz de rastreabilidade dos elementos do modelo, é realizado um caminhamento nos fluxos do modelo da versão base e delta para identificar os elementos modificados. Ao encontrar uma modificação no modelo da versão delta, a técnica observa qual a transição ou estado que foi modificado. Ao identificar a entidade modificada, a técnica procura na matriz de rastreabilidade os casos de testes afetados, ou seja, os casos de teste que exercitam esta entidade. Após identificados, os casos de teste afetados são selecionados para o subconjunto da suíte de regressão.

Ao encontrar uma entidade afetada, a técnica pára o caminhamento no fluxo que estava sendo analisado, e marca, como entidades afetadas, todos os vértices e transições que seguem, em profundidade, a partir daquele ponto. Dessa forma, é observado na matriz de rastreabilidade os casos de teste que exercitam estas entidades deste fluxo e são, portanto, adicionados ao subconjunto selecionado. A técnica seleciona também, os demais elementos do fluxo, a partir daquela entidade afetada. A técnica assume que os demais passos do fluxo são afetados pela modificação, e devem ser, portanto, testados. A busca por modificações continua, então, em outro fluxo do diagrama.

A primeira etapa de seleção pára uma vez que todos os fluxos do diagrama foram analisados. Os casos de teste afetados, obtidos na matriz de rastreabilidade, são os *Targeted Tests*. De acordo com os autores da técnica, estes casos de teste são os mais críticos para a realização do teste de regressão, pois as entidades modificadas são as que possuem mais chance de apresentar as faltas de regressão. Uma vez concluída a seleção dos *Targeted Tests*,

a técnica passa para a segunda etapa da seleção.

Observando o exemplo da Figura 3.6, verificamos que a única modificação realizada foi a remoção da atividade "Requisitar novamente o saldo". Portanto, nesta primeira etapa, os casos de teste que exercitam esta atividade são selecionados, com o objetivo de capturar as faltas de regressão que possam ocorrer devido à esta remoção.

3.2.2 Seleção dos *Safety Tests*

Os autores argumentam que apenas os *Targeted Tests* não são suficientes para a realização de um teste de regressão, pois outros casos de teste podem ser também críticos para a execução. Portanto, é necessário observar também, os casos de teste que exercitem funcionalidades críticas para a aplicação, ou seja, funcionalidades cujo custo de correção é muito alto quando faltas são encontradas. Estas funcionalidades podem não estar relacionadas com as modificações, mas podem ser afetadas por efeitos colaterais das modificações, ou integração de algum componente modificado [Binder 1999].

Diante disto, a segunda etapa da seleção é identificar os casos de teste que não são afetados pelas modificações, mas que apresentam um alto risco de apresentar alguma falta cuja correção é custosa. Sob esta perspectiva, a técnica realiza uma análise de riscos a partir de informações fornecidas pelo testador. O modelo de análise de risco é baseado em valores definidos em um modelo de riscos proposto por Amland [Amland 2000]. Estes valores, para cada caso de teste, são:

- **Custo:** O testador deve atribuir um valor de custo para cada caso de teste na suíte de regressão, baseado nos requisitos que o caso de teste cobre. O valor deve ser ponderado pela importância dos requisitos cobertos, e pelo custo de correção de uma falta encontrada pelo caso de teste.

- **Quantidade de defeitos:** É definido o número de defeitos encontrado no caso de teste, através de informações no histórico de execuções do caso de teste de regressão. Se o histórico não está disponível, o testador pode estimar este número baseando-se em sua experiência ou *know-how*.

- **Gravidade de defeitos:** Este valor é definido, pelo testador, para cada defeito encontrado no caso de teste baseando-se no custo de correção do respectivo defeito.

- **Probabilidade de gravidade:** A probabilidade de gravidade representa a probabilidade de um defeito grave ser encontrado pelo caso de teste. Este valor é definido a partir do produto entre a quantidade de defeitos e a gravidade do defeito.

- **Exposição de risco:** A exposição de risco indica o risco de que um defeito grave seja encontrado em um caso de teste caro. Este valor é obtido a partir do produto entre o custo do caso de teste e a probabilidade de gravidade.

Os valores deste modelo de riscos são organizados em uma matriz, e os casos de teste que apresentam os maiores valores de exposição de risco são selecionados como *Safety Tests*. A quantidade de *Safety Tests* selecionados é determinada pelo testador a partir, por exemplo, dos recursos disponíveis para o teste de regressão, ou de um limiar de riscos definido por gerentes ou clientes.

É possível obeservar que esta etapa da seleção é dependente de diversos aspectos definidos pelo testador. Porém, o objetivo é complementar a cobertura das modificações obtida através da seleção dos *Targeted Tests*, a partir da seleção de casos de teste que possam não estar relacionadas com as modificações mas que possuem um alto risco de encontrar faltas graves no software. No final desta etapa, a técnica encerra a execução, e o subconjunto da suíte de regressão é formado pelos *Targeted Tests* e pelos *Safety Tests* selecionados durante as duas etapas.

3.3 Re-teste baseado em Perfis

Esta técnica, proposta por Binder [Binder 1999], realiza a seleção do subconjunto da suíte de testes de acordo com um perfil operacional. Quando não há recursos disponíveis para realizar a execução de toda a suíte de regressão, a técnica é utilizada para selecionar os casos de teste relacionados aos casos de uso com maior frequência de uso.

Binder argumenta que esta técnica apresenta um baixo custo para ser aplicada quando o perfil já é definido [Binder 1999]. Apesar de selecionar os casos de teste relacionados aos casos de uso executados mais frequentemente, a técnica não está relacionada diretamente com as modificações, podendo não capturar, portanto, as faltas de regressão.

3.4 *Weighted Similarity Approach*

Weighted Similarity Approach, ou WSA, é uma técnica proposta por Bertolino et al. utilizada para a seleção automática de casos de testes [Bertolino et al. 2008]. Por ser uma técnica de seleção, a técnica executa com o objetivo de selecionar casos de teste de acordo com a quantidade de recursos disponíveis para o processo de teste.

Esta técnica possui, como objetivo, selecionar casos de teste que cobrem diferentes transições do modelo, identificando similaridades entre os casos de testes. Casos de teste que cobrem as mesmas transições do modelo (ou passos da aplicação) são considerados redundantes. Outro aspecto considerado na seleção é a utilização de uma abordagem baseada em valor.

As autoras da técnica consideram como "valores" os parâmetros e escolhas úteis para adaptar a estratégia de teste para as necessidades do usuário. Ou seja, as diferentes funcionalidades do sistema não apresentam a mesma "importância" para o usuário, e isto deve ser considerado na estratégia de teste da aplicação. Diante disto, é necessário definir os critérios que determinam uma funcionalidade como importante para o propósito do teste. Exemplos destes critérios são a complexidade dos componentes, ou a frequência de uso da funcionalidade [Bertolino et al. 2008].

Sob esta perspectiva, a técnica considera como funcionalidades importantes aquelas funcionalidades que são frequentemente exercitadas por um usuário da aplicação ou sistema. As informações de frequência de utilização das funcionalidades são organizadas em um perfil de uso da aplicação ou sistema. No trabalho proposto por Bertolino et al. [Bertolino et al. 2008], o perfil de uso é expresso através de valores de probabilidades nas transições do modelo. A partir destas informações, é calculado um valor de probabilidade para cada caso de teste representando a frequência com que o cenário exercitado pelo respectivo caso de teste será exercitado também pelo usuário, na versão final do software.

Dessa forma, as entradas da técnica são: a suíte de testes, o perfil de uso e a cobertura desejada. Inicialmente, a técnica constrói uma matriz, onde cada linha e cada coluna representa um caso de teste da suíte. Cada valor da matriz é obtido através da similaridade entre um par de casos de teste (um correspondente a uma linha, e o outro a uma coluna da matriz) dividido pela probabilidade do caso de teste da respectiva linha. A simi-

laridade entre dois casos de teste é obtida através da fórmula proposta por Cartaxo et. al [Cartaxo et al. 2007, Cartaxo et al. 2009] (apresentada na equação 4.1 do Capítulo 4).

Com esta estratégia, a matriz apresenta as informações de similaridade entre os casos de teste da mesma suíte, e o valor de probabilidade de cada caso de teste obtido a partir do perfil de uso. Estes valores são utilizados para decidir quais casos de teste não devem ser selecionados. Diante disto, a técnica itera, então, nas linhas e colunas da matriz, procurando identificar os maiores valores. Ao encontrar empates, ou seja, dois ou mais casos de teste com o maior valor da matriz, a técnica realiza uma escolha aleatória para decidir qual caso de teste não será selecionado.

A técnica repete o processo de remoção dos casos de teste da suíte até que esta apresente a quantidade de casos de teste especificada pela cobertura definida pelo testador. Os valores de probabilidade são inseridos no modelo pelo usuário da técnica (testador), e o processo de geração automática realiza o cálculo dos valores de probabilidades de cada caso de teste.

É possível observar que o conceito de similaridade entre os casos de teste utilizado pela téncica pode ser aplicado ao contexto de teste de regressão. Sob esta perspectiva, casos de teste muito similares são casos de teste que não sofreram muitas modificações, e observando esta característica entre os casos de teste de regressão das diversas versões do software, é possível identificar trechos da aplicação que foram modificados. Além disto, a perspectiva de seleção baseada no perfil de uso é semelhante também à abordagem de re-teste baseado em perfis proposta por Binder [Binder 1999] (Seção 3.3).

Diante disto, a técnica de WSA foi adaptada para o contexto de teste de regressão com o objetivo de utilizar o conceito da seleção baseada na similaridade e perfil de uso em uma técnica de re-teste seletivo baseado em especificação. Esta adaptação (*WSA for Regression Testing*, ou WSA-RT) é uma das contribuições deste trabalho, e os detalhes desta técnica são descritos no Capítulo 4.

3.5 Técnica baseada em *Clusters*

Uma das técnicas escolhidas para a realização do estudo experimental é a seleção baseada em *clusters*. Esta técnica foi apresentada por Laski e Szermer [Laski and Szermer 1992], e possui o objetivo de identificar *clusters* em um grafo de fluxo de controle. Estes *clusters* são

subgrafos com uma entrada e uma saída que representem porções modificadas do modelo (seja ele o código ou a especificação). No contexto de código, o GFC representa os fluxos de execução do código, enquanto que no contexto de especificação, o GFC representa os fluxos de execução do software, por exemplo, os cenários dos casos de uso.

Em um estudo empírico para técnicas baseadas em código, conduzido por Rothermel e Harrold [Rothermel and Harrold 1996], esta técnica apresentou um bom desempenho de inclusão e precisão, capturando muitas faltas de regressão durante sua execução. Diante disto, decidimos investigar se este mesmo desempenho é observado quando aplicamos a técnica no contexto de especificação, uma vez que o modelo da técnica pode ser utilizado no re-teste seletivo baseado em especificação [Chen et al. 2002].

Para executar, a técnica necessita dos dois modelos: um modelo correspondente à versão base, e um modelo correspondente à versão delta da aplicação. Ilustramos dois GFC nas figuras 3.7(a) e 3.7(b), representando os modelos da versão base e da versão delta de uma aplicação, respectivamente.

Em uma primeira etapa, a técnica realiza um caminhamento em ambos modelos, procurando identificar os elementos (transições e estados) que foram modificados. Durante este caminhamento, a técnica investiga quais transições foram removidas, adicionadas ou tiveram seu rótulo modificado.

É possível observar na Figura 3.7 as seguintes modificações que foram realizadas no modelo: a transição entre os estados 8 e 10 foi removida; o estado 11 foi removido; e a transição entre os estados 16 e 12 foi adicionada. O primeiro passo da técnica é identificar os *clusters* dos GFC.

Os *clusters* são identificados através de um caminhamento simultâneo entre os dois modelos, iniciando na raiz (estado de rótulo 0) dos GFC. Estes *clusters* são subgrafos do GFC que contemplam dois ou mais fluxos do modelo. Ao encontrar um estado de decisão (i.e. estado que possui mais de uma aresta direcionadas para outros estados), a técnica inicia a verificação de um possível *cluster*. A técnica caminha pelos dois ou mais fluxos encontrados, até que estes fluxos se encontrem novamente, em um fluxo comum da aplicação.

Observando o GFC da Figura 3.7 (a), a técnica inicia a busca por *clusters* já na raiz, onde o fluxo é dividido para os estados 1, e 6. O primeiro cluster é encontrado caminhando a partir do estado 3, até o estado 20, contemplando os estados $3, 4$ e 5. No segundo fluxo, são

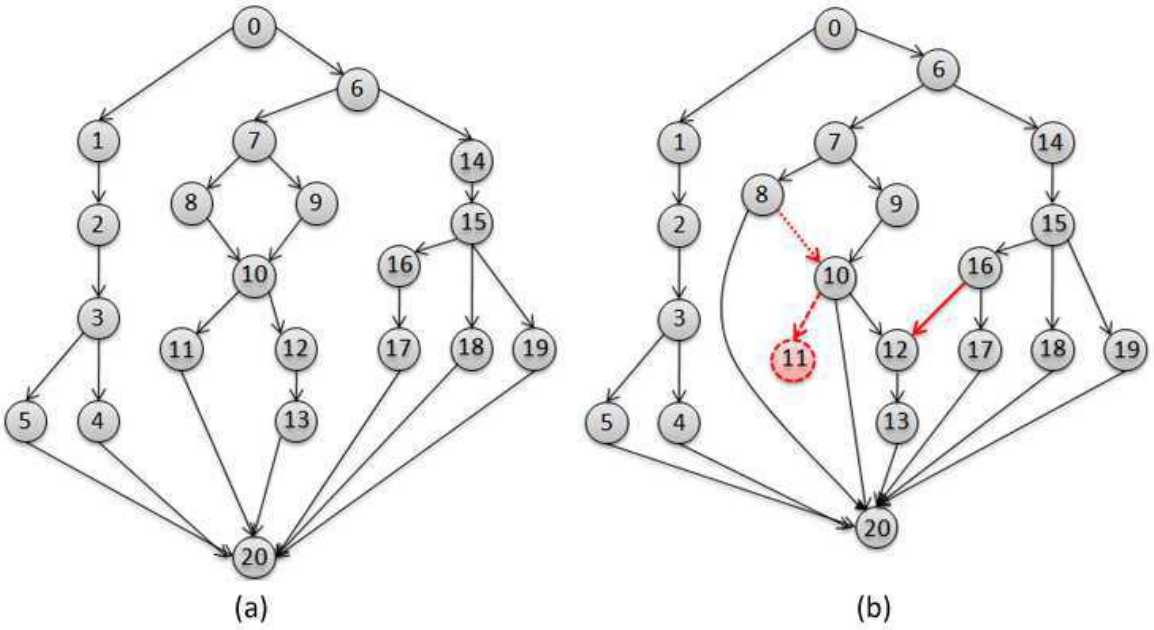

Figura 3.7: Exemplos de Grafos de Fluxo de Controle de uma versão base (a) e uma versão modificada (b) de uma aplicação.

encontrados dois *clusters*, que iniciam nos estados 7 e 10, e contemplam, respectivamente, os estados $7, 8$ e 9, e os estados $10, 11, 12$ e 13. No terceiro fluxo, iniciado no estado 14, é identificado um *cluster* no estado 15, contemplando os estados $15, 16, 17, 18$ e 19.

O mesmo processo é realizado no GFC da Figura 3.7 (b). Pelas modificações realizadas os *clusters* encontrados são diferentes dos encontrados no GFC da versão base. Na versão delta são encontrados apenas os *clusters* do estado 3, que contempla os estados $3, 4$ e 5, e um *cluster* no estado 6, que contempla os estados $6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18$ e 19. Os *clusters* identificados nos GFC podem ser observados nas figuras 3.8(a) e 3.8(b).

Após esta primeira etapa, é verificado se os dois grafos são isomorfos. Caso não seja observado o isomorfismo, a técnica reinicia o processo de identificação de *clusters*. Neste segunda etapa, ela identifica para o GFC da Figura 3.8 (a), um *cluster* no estado 6, que contempla os estados $6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18$ e 19. Após estabelecer este *cluster*, é verificado que os dois GFC são isomorfos, como é possível observar na Figura 3.9.

Neste momento, a técnica observa cada subgrafo (*cluster*) identificando as diferenças entre eles. Através desta comparação entre os clusters, é possível identificar as modificações e os elementos do GFC que são afetados por elas, ou seja, é possível identificar cada modifica-

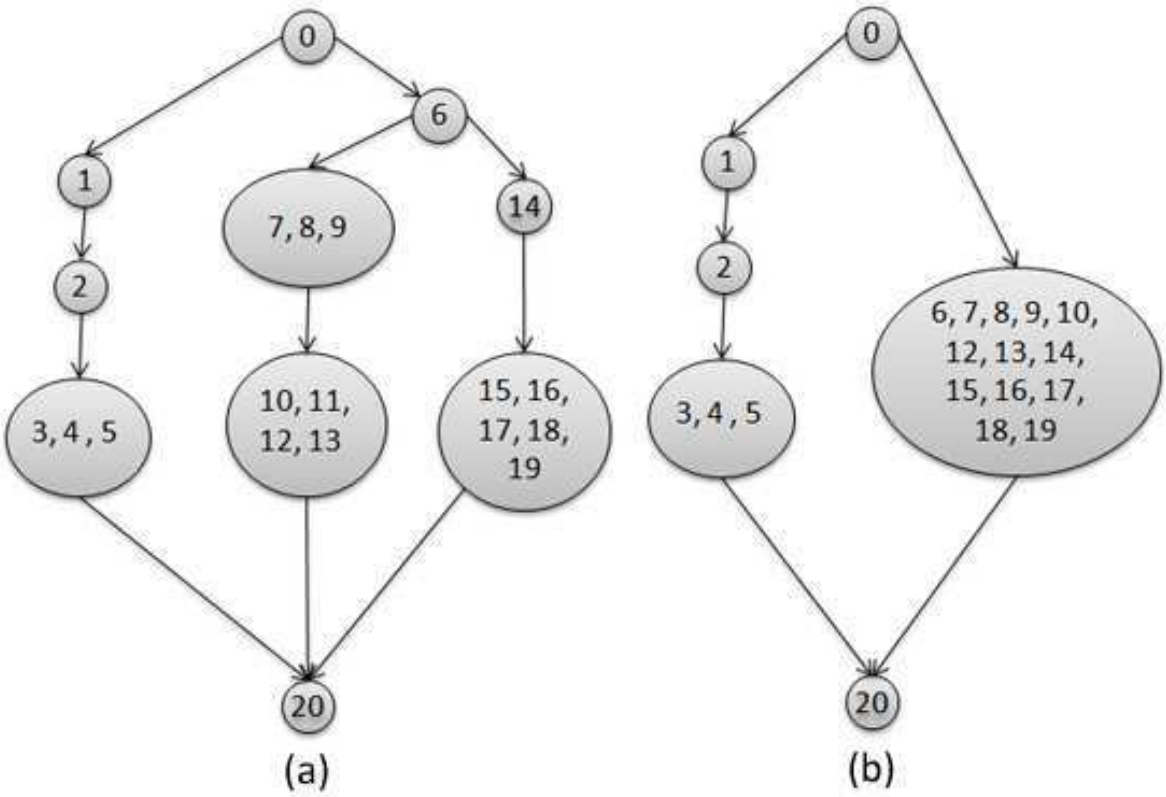

Figura 3.8: *Clusters* encontrados nos modelos da versões base (a) e modificada (b).

ção e até onde (trechos do fluxo) esta se propaga. No caso do exemplo ilustrado, é verificado ao analisar o *cluster* dos estados $3, 4$ e 5 que não houve modificação. No entanto, ao observar o *cluster* dos estados $6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18$ e 19, a técnica identifica todas as modificações realizadas no modelo.

Uma vez identificadas as modificações, são selecionados os casos de teste que exercitam os estados do *cluster* que possui as modificações. No nosso exemplo, são selecionados apenas os casos de teste que exercitam os estados $6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18$ e 19. Como podemos observar, e como é apontado pelos autores da técnica, a obtenção de *clusters* com muitos estados pode diminuir o desempenho da técnica, pois quanto maior a quantidade de estados em um *cluster*, maior a quantidade de casos de teste que exercitam este *cluster*, e maior a quantidade de casos de teste que devem ser selecionados.

Além disto, nem todos os estados de um determinado *cluster* podem revelar faltas de regressão. No nosso exemplo, é possível observar que os estados $6, 14, 15, 18$ e 19 não estão diretamente relacionados com nenhuma das modificações, o que diminui as chances de que

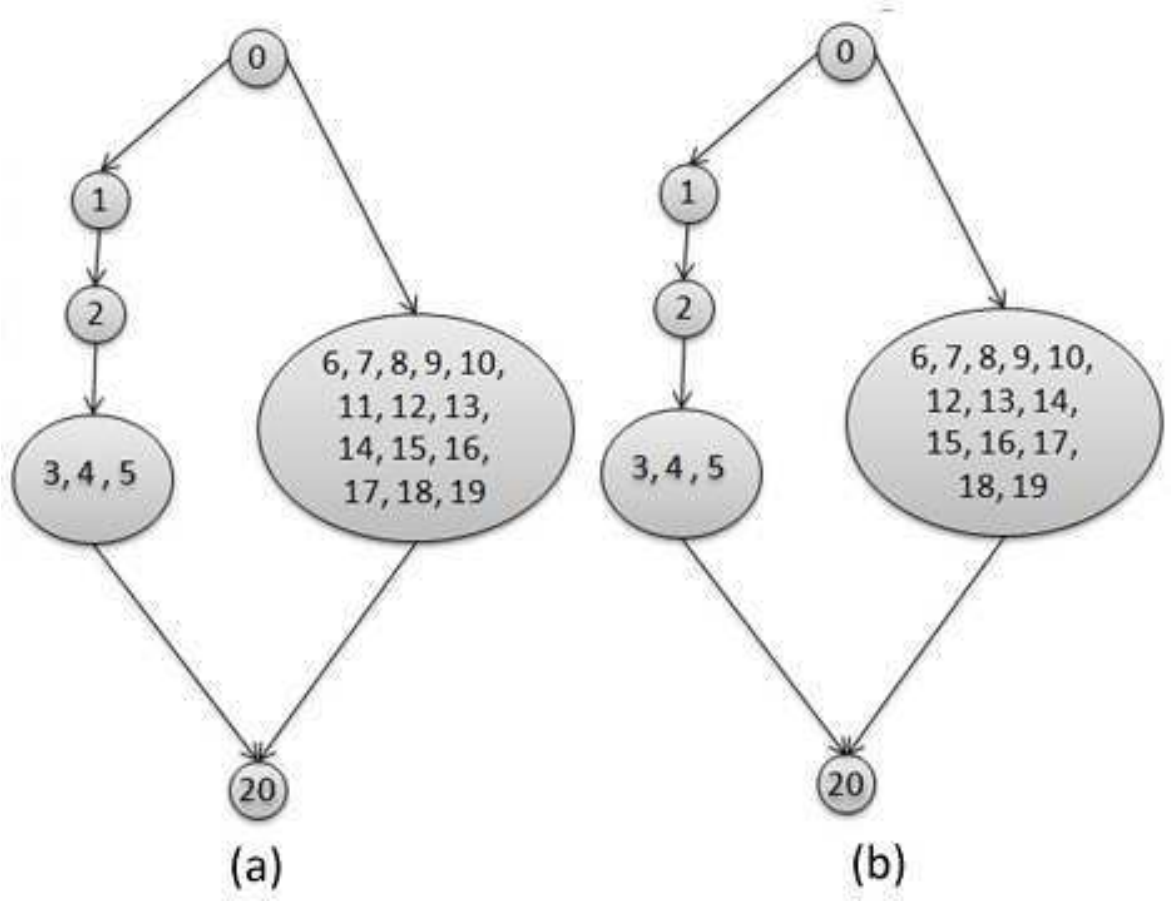

Figura 3.9: *Clusters* encontrados nos modelos para a obtenção dos GFC isomorfos.

faltas de regressão sejam encontradas nestes estados. No entanto, eles devem ser selecionados pela técnica, pois fazem parte do fluxo de execução de alguns trechos modificados.

Sob o contexto de código, os *clusters* são identificados apenas no contexto intraprocedural, dessa forma, a técnica não leva em consideração os relacionamentos interprocedurais, ou seja, as dependências entre procedimentos. Dessa forma, são obtidos, geralmente, diversos *clusters* pequenos, o que melhora a identificação das faltas. Apesar de possuir esta relação com o tamanho dos *clusters*, a técnica é considerada segura [Rothermel and Harrold 1996], no contexto de código, pois seleciona todos os casos de teste que exercitem algum trecho de código modificado, detectando, portanto, as faltas de regressão.

3.6 Técnica de Seleção Aleatória de Casos de Teste

A técnica de seleção aleatória de casos de teste é utilizada em diversos contextos tanto industriais quanto acadêmicos [Cartaxo et al. 2009, Graves et al. 1998]. Devido à sua ampla utilização, esta técnica foi escolhida para a realização do estudo experimental. Além disto, esta técnica apresenta características que fornecem uma boa perspectiva comparativa entre uma técnica que utiliza um critério de seleção, e uma que não utiliza.

Esta técnica é utilizada na indústria quando os prazos são muito restritos, e não há ferramentas para realizar a seleção dos casos de teste de regressão [Graves et al. 1998]. Geralmente, pela falta de suporte ferramental, os testadores ou os desenvolvedores selecionam os casos de teste procurando ligar, sem nenhum mecanismo automático, as funcionalidades aos casos de teste.

Na comunidade acadêmica, a técnica de seleção aleatória é utilizada em estudos de casos e estudos experimentais para comparar o desempenho das técnicas [Cartaxo et al. 2009, Cartaxo et al. 2007, Graves et al. 1998]. Esta comparação é utilizada para, por exemplo, mostrar as vantagens de aplicar a técnica em um cenário onde uma técnica de seleção não é utilizada (que geralmente indica a utilização de uma seleção *ad hoc* dos casos de teste). Diante disto, é esperado que as técnicas apresentem um desempenho melhor que a técnica de seleção aleatória, para as propriedades de inclusão, precisão, densidade de faltas, e potencial de redução.

Para utilizar a seleção aleatória, é necessário definir a quantidade de casos de teste que devem ser selecionados. Esta quantidade atua como a condição de parada da técnica, e representa a quantidade de casos de teste que podem ser executados de acordo com os recursos disponibilizados para a etapa de testes. Uma vez definida a cobertura desejada, a técnica seleciona, aleatoriamente, casos de teste até atingir a cobertura especificada.

Sob esta perspectiva, o potencial de redução da técnica é definido pelo testador. Apesar disto, é importante comparar o potencial de redução desta técnica com as demais para complementar as informações obtidas a respeito da cobertura de faltas de regressão. Dessa forma, é possível observar se o potencial de redução é adequado para a cobertura de faltas de regressão atingida.

3.7 Considerações Finais do Capítulo

Neste capítulo foram apresentadas técnicas de re-teste seletivo e de seleção de casos de teste em TBM. As técnicas executam em elementos do modelo da especificação, como requisitos, casos de uso, ou os próprios casos de teste funcionais do sistema sob teste. Através da seleção de um subconjunto da suíte de teste, a quantidade de casos de teste utilizada no processo de teste é menor, e portanto, o custo desta etapa é reduzida.

Dentre as técnicas apresentadas neste capítulo, algumas foram investigadas no estudo experimental (as técnicas das seções 3.1, 3.2, 3.5 e 3.6), enquanto outras (as técnicas das seções 3.4 e 3.3) foras utilizadas para estruturar a técnica *Weighted Similarity Approach for Regression Testing* (WSA-RT) proposta neste trabalho. WSA-RT é apresentada no próximo capítulo e investigada no estudo experimental.

Capítulo 4

Weighted Similarity Approach para Teste de Regressão

Uma das contribuições deste trabalho é a adaptação da técnica *Weighted Similarity Approach* (WSA) proposta por Bertolino et al. [Bertolino et al. 2008]. Observamos que o conceito de similaridade apresentado no trabalho, assim como as considerações a respeito do perfil de uso utilizado para a seleção dos casos de teste, seriam bem utilizados no contexto de teste de regressão.

O conceito de similaridade entre casos de testes apresentado no trabalho original, pode ser utilizado para verificar a similaridade entre casos de testes regressão. A expectativa é de que os casos de teste de versões diferentes da aplicação, menos similares entre si, cubram mais modificações realizadas no modelo. Esta perspectiva indica que o conceito de similaridade utilizado por Bertolino et al., possa ser utilizado para identificar e selecionar os casos de testes modificados. Diante disto, a técnica é capaz de indentifcar modificações observando os casos de teste, enquanto que a maioria das técnicas propostas, necessitam analisar transições ou estados do modelos da aplicação.

Por sua vez, o perfil de uso, considerado pela técnica, é utilizado para identificar os casos de teste com maior probabilidade de serem executados por um usuário. Ou seja, é útil para identificar e selecionar os cenários mais executados pelo usuário. A presença de faltas nestes cenários é muito crítica para o produto, pois o custo de uma manutenção em uma funcionalidade importante para o usuário, ou cliente, é muito alto [Chen et al. 2002]. Sob esta perspectiva, o perfil de uso considerado por WSA, pode ser utilizado para realizar

o re-teste baseado em perfis de uso [Binder 1999].

O re-teste baseado em perfis é uma técnica que seleciona os casos de teste de regressão que seriam exercitados pelo usuário. Esta técnica necessita de um perfil de uso do produto, geralmente fornecido pelo testador, contendo as informações referentes aos fluxos executados frequentemente pelo usuário.

A adaptação realizada na técnica, para o contexto de teste de regressão, foi submetida ao estudo experimental com o objetivo de investigar e avaliar seu desempenho junto com as demais técnicas analisadas. Portanto, a técnica WSA para teste de regressão (*Weighted Similarity Approach for Regression Testing*, ou WSA-RT) é descrita nas próximas seções.

4.1 WSA para Teste de Regressão (WSA-RT)

O primeiro aspecto que deve ser incorporado na adaptação de WSA é a capacidade de identificar modificações. Um dos principais objetivos de uma técnica de re-teste seletivo é capturar faltas de regressão, e estas estão relacionadas com as modificações realizadas no software. Para WSA-RT, a matriz de similaridade é o recurso utilizado para identificar as modificações. Portanto, é necessário observar a similaridade entre duas versões distintas do software.

Utilizando técnicas de geração automática de casos de teste em TBM, são geradas duas suítes de testes. Uma a partir do modelo da versão base, e outra a partir do modelo da versão delta. Estas duas suítes são referenciadas, ao longo do texto, como T e T', respectivamente. T' já apresenta as modificações realizadas no software (e.g. adição, remoção e mudanças nos passos da aplicação). O modelo da versão delta, assim como os casos de teste gerados a partir dele, possuem as informações do perfil de uso fornecido pelo testador. O perfil de uso é definido a partir das diretrizes estabelecidas por Bertolino et al., em que:

- A probabilidade associada a uma transição representa a probabilidade de que a respectiva transição será executada por um usuário final da aplicação.
- A soma dos valores de probabilidade associados às transições cujo predecessor direto comum apresenta um grau de saída *maior que* 1, deve somar 1.
- As transições de um vértice cujo predecessor direto apresenta um grau de saída *igual a* 1, devem possui valor de probabilidade 1.

A técnica WSA-RT utiliza, como entrada, o perfil de uso do modelo modificado e as duas suítes de teste (T e T'). A partir destes elementos, a técnica é executada em quatro etapas ilustradas na Figura 4.1. É importante observar que as etapas, devem ser executadas na sequência especificada pela técnica, e que, a última etapa (Etapa 4), é opcional. Apesar de opcional, recomendamos a realização da última etapa da técnica, com o objetivo de prover um melhor resultado da técnica. Cada etapa de WSA-RT será detalha nas seções seguintes.

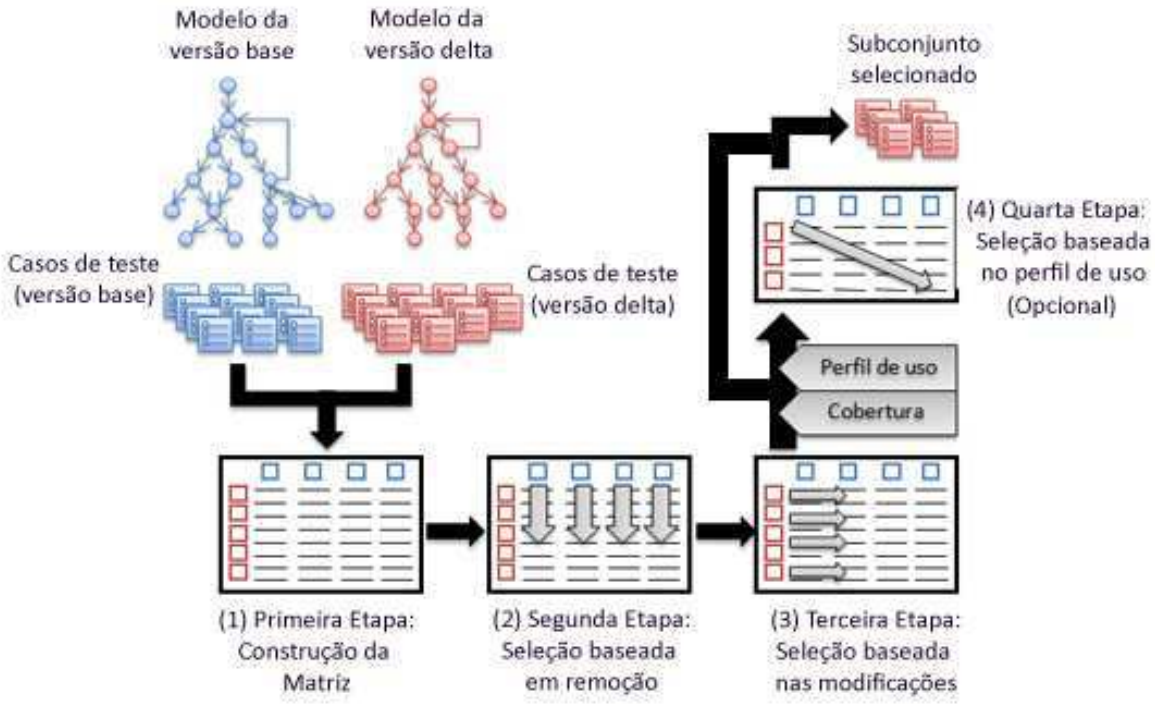

Figura 4.1: Resumo da execução de WSA-RT.

A suíte resultante é uma união de todos os casos de teste selecionados nas etapas 2, 3 e 4 da técnica. Outro aspecto da técnica é que a suíte selecionada é subconjunto de T'. Dessa forma, o subconjunto selecionado não apresenta casos de teste obsoletos, e cobre as funcionalidades modificadas [Subramaniam et al. 2009, Mahdian et al. 2009].

Para ilustrar as etapas e os passos da técnica, será utilizado um exemplo. Para manter a simplicidade do modelo e da explicação da técnica, foram elaborados os sistemas de transições rotuladas (STR) da Figura 4.2 para ilustrar a técnica. Nesta figura são especificados o modelo da versão base (a) e o modelo da versão delta (b).

As informações de probabilidade do perfil de uso para o modelo modificado (Figura 4.2

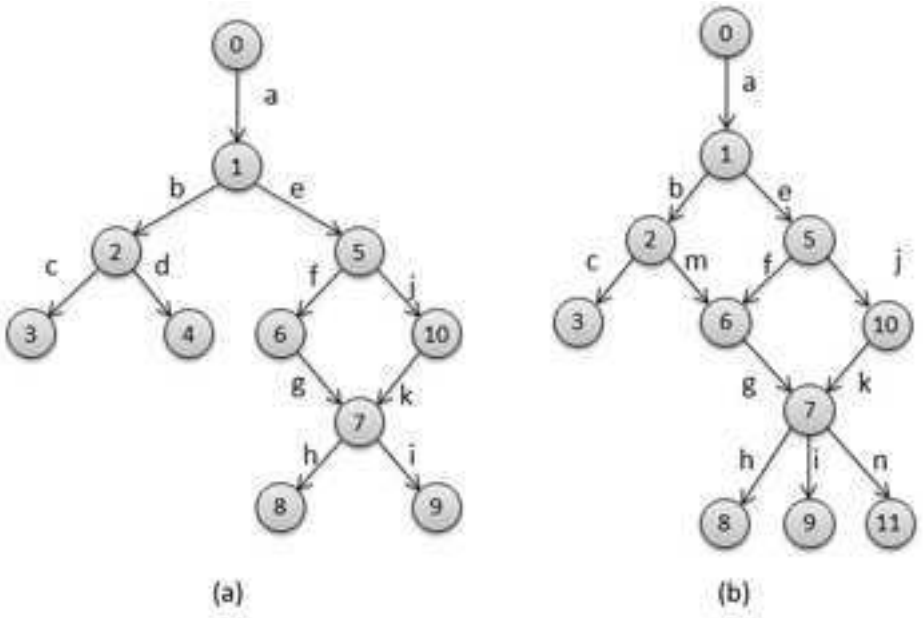

Figura 4.2: Modelos da versão base (a) e delta (b) utilizados para ilustrar a técnica WSA-RT.

(b)) são apresentadas na Tabela 4.1.

Como podemos ver, foram realizadas apenas 3 modificações no modelo: a remoção da transição de rótulo "d", a adição da transição "m" entre os estados 2 e 6; e a adição da transição de rótulo "n" no estado 7. A partir dos modelos do exemplo, foram gerados os casos de teste apresentados na Tabela 4.2. Esta tabela apresenta os casos de teste de T e de T', assim como o valor de probabilidade de cada caso de teste de T' (apresentado na última coluna). Este valor foi obtido a partir do perfil de uso especificado no modelo, multiplicando o valor de probabilidade em cada transição do caso de teste [Bertolino et al. 2008].

As próximas seções descrevem cada etapa de execução da técnica. O exemplo, ilustrado através da Figura 4.2 e das Tabelas 4.1 e 4.2, é utilizado para demonstrar o processo de seleção realizado pela técnica.

4.2 Construção da Matriz de Similaridade

A primeira etapa da técnica é a construção da matriz de similaridade. Em WSA, a matriz apresenta a similaridade entre pares de casos de teste provenientes de uma mesma suíte de testes. Por outro lado, em WSA-RT, a matriz apresenta a similaridade entre pares de casos

Tabela 4.1: Perfil de uso do modelo da versão delta.

Estado	Transição	Probabilidade de execução
0	a	1
1	b	0,5
	e	0,5
2	c	0,5
	m	0,5
5	f	0,6
	j	0,4
6	g	1
10	k	1
7	h	0,8
	i	0,1
	n	0,1

Tabela 4.2: Casos de testes gerados automaticamente a partir dos modelos.

Suítes geradas a partir dos modelos										
Versão base						**Versão delta**						**Probabilidade de execução**
TC_1	a	b	c			TC'_1	a	b	c			0,25
TC_2	a	b	d			TC'_2	a	b	m	g	h	0,2
TC_3	a	e	f	g	h	TC'_3	a	b	m	g	i	0,025
TC_4	a	e	f	g	i	TC'_4	a	b	m	g	n	0,025
TC_5	a	e	j	k	h	TC'_5	a	e	f	g	h	0,24
TC_6	a	e	j	k	i	TC'_6	a	e	f	g	i	0,03
						TC'_7	a	e	f	g	n	0,03
						TC'_8	a	e	j	k	h	0,16
						TC'_9	a	e	j	k	i	0,02
						TC'_{10}	a	e	j	k	n	0,02

de teste provenientes de duas suítes de testes diferentes. Apesar de diferentes, é necessário que estas suítes tenham sido geradas a partir de duas versões distintas do mesmo software.

Em WSA-RT, cada linha da matriz representa um caso de teste de T', enquanto que cada

coluna da matriz representa um caso de teste de T. Portanto, a informação de similaridade entre cada par de caso de teste (TC'_i, TC_j), de forma que $TC'_i \in T'$, e $TC_j \in T$, representa a similaridade entre os casos de teste das versões base e delta (colunas e linhas, respectivamente). Neste trabalho, é utilizado o conceito de similaridade apresentado por Cartaxo et al. [Cartaxo et al. 2009]. Dessa forma, é obtida uma matriz $A_{|T'|\times|T|}$ onde, cada valor $a[i,j]$ da matriz é calculado da seguinte forma:

$$a[i,j] = \frac{QtdPassosIguais(TC'_i, TC_j)}{TamanhoMédio(TC'_i, TC_j)}; \tag{4.1}$$

$$TamanhoMédio(TC'_i, TC_j) = \frac{|TC'_i| + |TC_j|}{2}. \tag{4.2}$$

A função $QtdPassosIguais(TC'_i, TC_j)$ no numerador da Equação 4.1 calcula a quantidade de passos iguais entre os casos de testes especificados como parâmetros. Neste trabalho é utilizado o conceito de passos iguais especificado por Cartaxo et al. [Cartaxo et al. 2007], onde dois passos (ou transições) são iguais, em um STR, quando possuem o mesmo vértice de origem, o mesmo rótulo e o mesmo vértice de destino.

O denominador da Equação 4.1 equilibra a quantidade de passos iguais pela média aritmética entre o tamanho de cada caso de teste do par analisado (Equação 4.2). Esta divisão é realizada com o objetivo de normalizar os valores de similaridade com relação aos diversos tamanhos dos casos de testes. Neste trabalho, o tamanho de um caso de teste ($|TC_x|$) é a quantidade de passos (no nosso caso, transições no modelo) que este possui.

Para ilustrar a utilização da fórmula, calculamos a similaridade entre os casos de teste TC'_3 e TC_5. Utilizando os dados dos casos de teste na Equação 4.1, temos:

$$
\begin{aligned}
TamanhoMédio(TC'_3, TC_5) &= \frac{|TC'_3| + |TC_5|}{2} \\
&= \frac{5+5}{2} \\
&= 5; \\
a[3,5] &= \frac{QtdPassosIguais(TC'_3, TC_5)}{TamanhoMédio(TC'_3, TC_5)} \\
&= \frac{|[a,b,m,g,i] \cap [a,e,j,k,h]|}{5} \\
&= \frac{|[a]|}{5} \\
&= \frac{1}{5} \\
&= 0,20;
\end{aligned}
$$

Portanto, a similaridade encontrada entre TC'_3 e TC_5 é $0,2$. A similaridade entre dois casos de testes que exercitam as mesmas transições do modelo é 1. Ou seja, os valores de similaride 1 são utilizados para identificar o mesmo caso de teste nas duas versões: $Similaridade(TC'_i, TC_j) = 1 \Rightarrow TC'_i = TC_j$. Podemos observar esta característica da técnica, utilizando a Equação 4.1 para TC'_1 e TC_1 ($TC_1 = TC'_1 = [a,b,c]$).

Na Tabela 4.3, apresentamos a matriz de similaridade, obtida a partir das suítes de teste da Tabela 4.2.

Tabela 4.3: Matriz de similaridade obtida a partir das suítes de testes utilizadas no exemplo.

Casos de teste	TC_1	TC_2	TC_3	TC_4	TC_5	TC_6
TC'_1	1,00	0,66	0,25	0,25	0,25	0,25
TC'_2	0,50	0,50	0,60	0,40	0,40	0,20
TC'_3	0,50	0,50	0,40	0,60	0,20	0,40
TC'_4	0,50	0,50	0,40	0,40	0,20	0,20
TC'_5	0,25	0,25	1,00	0,80	0,60	0,40
TC'_6	0,25	0,25	0,80	1,00	0,40	0,60
TC'_7	0,25	0,25	0,80	0,80	0,40	0,40
TC'_8	0,25	0,25	0,60	0,40	1,00	0,80
TC'_9	0,25	0,25	0,40	0,60	0,80	1,00
TC'_{10}	0,25	0,25	0,40	0,40	0,80	0,80

Os valores de similaridade na matriz serão utilizados para identificar os casos de teste da

versão delta que são mais similares aos da versão base. Os casos de teste menos similares entre si exercitam mais modificações, uma vez que as duas suítes são geradas a partir de versões diferentes de uma mesma aplicação. Após construir a matriz de similaridade, o processo de seleção é iniciado. As informações de probabilidade são incorporadas na matriz apenas durante a última etapa da técnica.

4.3 Seleção Baseada nos Casos de Teste Obsoletos

Dentre as diversas modificações em um software, podemos encontrar a remoção de cenários ou passos da aplicação que não são mais desejados pelo cliente. A remoção destes passos, do software, faz com que alguns casos de teste da versão base não possam mais ser executados. Estes casos de teste são denominados *obsoletos*, e são encontrados na suíte de teste da versão base.

É necessário testar as transições não removidas de um caso de teste obsoleto, pois estas podem revelar as faltas de regressão inseridas devido à remoção realizada no modelo. Este tipo de falta de regressão ocorre quando a aplicação atinge um estado, que não deveria ser atingido na versão delta, devido à remoção de uma transição. Diante disto, para capturar estas faltas é necessário exercitar a maior quantidade possível, de estados e transições, exercitadas pelos casos de teste obsoletos [Korel et al. 2002, Chen et al. 2007].

Os casos de teste obsoletos são encontrados na versão base, em WSA-RT, através das colunas da matriz de similaridade. As informações dos casos de teste obsoletos, presentes na coluna da matriz, podem ser utilizadas para identificar as faltas de regressão devido à remoção de transições. Portanto, utilizamos as colunas da matriz para encontrar, dentre os casos de teste da versão delta (as linhas da matriz), os mais similares aos casos de testes obsoletos.

O primeiro passo é iterar sobre as colunas da matriz, ou seja, iterar sobre os casos de teste da versão base. Em cada coluna, iteramos sobre as linhas verificando os valores de similaridade. Ao encontrar uma linha com valor de similaridade 1, identificamos algum caso de teste da versão delta que exercita os mesmos passos que o respectivo caso de teste na versão base. Dessa forma, o caso de teste não exercita nenhuma transição removida do modelo, e portanto, não é obsoleto. Portanto, nenhum caso de teste da versão delta,

semelhante à ele, precisa ser selecionado.

Se não for encontrado um valor de similaridade igual a 1 dentre as linhas da coluna analisada, identificamos que o caso de teste é obsoleto, e portanto, cobre pelo menos uma transição removida do modelo. Diante disto, selecionamos o caso de teste da versão delta mais similar ao caso de teste obsoleto. Ou seja, selecionamos o caso de teste correspondente à linha que apresenta o maior valor de similaridade da coluna analisada.

Através desta estratégia de seleção, o caso de teste selecionado exercita a maior quantidade de transições exercidas pelo caso de teste obsoleto, aumentando a probabilidade de encontrar alguma falta de regressão devido à remoção [Korel et al. 2002, Bo 2005, Chen et al. 2007]. Se, ao analisar uma coluna, for encontrado um empate entre duas ou mais linhas, a linha selecionada é escolhida aleatoriamente.

Observando a matriz da Tabela 4.3, verificamos que a coluna referente ao caso de teste TC_2 é a única que não apresenta similaridade 1 nas suas linhas. Portanto, TC_2 é um caso de teste obsoleto. Podemos observar que a identificação, realizada por WSA-RT, é correta, pois TC_2 exercita a única transição removida do modelo – transição "d" da Figura 4.2 (a).

Iterando sobre cada linha, da coluna de TC_2, observamos que o maior valor de similaridade encontrado, é $0,66$. A linha em que o maior valor é encontrado corresponde ao caso de teste mais similar ao TC_2. Diante disto, nesta primeira etapa da técnica, e para o exemplo apresentado, é selecionado o caso de teste TC'_1 da suíte da versão delta.

Ao observar o valor de similaridade 1 nas colunas de TC_1, TC_3, TC_4, TC_5 e TC_6, identificamos que esses casos de teste da versão base não sofreram modificações, e portanto, não exercitam transições modificadas no modelo. De acordo com esta etapa da técnica, não é necessário selecionar casos de testes similares a eles, pois a probabilidade de que estes casos de teste encontrem faltas de regressão por remoção é menor que a do caso de teste selecionado (TC_2). Os demais tipos de modificações são identificadas, durante as próximas etapas da ténica.

4.4 Seleção Baseada na Similaridade entre Casos de Testes Adicionados ou Modificados

Após identificar as transições removidas, WSA-RT observa as demais modificações que possam ser realizadas: mudança nos rótulos, e a adição de transições. A primeira não causa modificações na estrutura do modelo, porém, pode inserir faltas de regressão, por exemplo, devido à falta de atualização em demais componentes da aplicação (e.g. Interface Gráfica) dependentes deste rótulo. Por sua vez, a adição de transições muda a estrutura do modelo, aumentando a quantidade de casos de teste de regressão, pois novos casos de teste devem exercitar as novas trasições. Estas duas modificações são identificadas na suíte da versão delta.

WSA-RT identifica estas modificações observando as linhas da matriz. Para cada linha é verificado se o caso de teste correspondente não cobre modificações no modelo (i.e. apresenta o valor de similaridade igual a 1), e portanto, apresenta baixa probabilidade de capturar faltas de regressão [Binder 1999]. Se nenhum valor de similaridade 1 é observado na linha analisada, o caso de teste correspondente à linha analisada é selecionado.

Observando o exemplo da Tabela 4.3, é possível identificar que as linhas TC'_1, TC'_5, TC'_6, TC'_8 e TC'_9 apresentam o valor de similaridade 1, quando comparados, respectivamente, aos casos de teste TC_1, TC_3, TC_4, TC_5 e TC_6 da versão base. É importante observar que TC'_1 já foi selecionado durante a segunda etapa da técnica, pois é o caso de teste mais semelhante a TC_2 da versão base. Uma vez que este caso de teste já foi selecionado durante a segunda etapa, apenas os casos de teste TC'_5, TC'_6, TC'_8 e TC'_9 não são selecionados para a suíte final.

Após iterar sobre todas as linhas da matriz, a técnica remove da matriz, as linhas dos casos de teste não selecionados, e conclui a terceira etapa. A próxima etapa é a seleção baseada no perfil de uso. No exemplo ilustrado, até essa etapa da técnica, os caso de teste selecionados foram: TC'_1, TC'_2, TC'_3, TC'_4, TC'_7 e TC'_{10}.

4.5 Seleção baseada no Perfil de Uso

Como podemos observar, até o fim da terceira etapa utilizando o exemplo especificado, a redução observada na suíte da versão delta foi de 4 casos de teste. Considerando a situação

em que o testador deseja reduzir mais a suíte obtida no fim da terceira etapa, a próxima seleção realizada pela técnica considera o perfil de uso da aplicação. Diante disto, esta etapa da técnica, apesar de recomendada, é considerada opcional.

O testador especifica, no modelo da versão delta, valores de probabilidade para os vértices de decisão do STR. Estes valores de probabilidade representam a probabilidade de que o usuário execute a aplicação seguindo um determinado fluxo [Barbosa et al. 2007]. Dessa forma, é importante selecionar os casos de teste com maior probabilidade de serem executados pelo usuário, pois o custo da correção de faltas encontradas por estes casos de teste, e.g. em uma versão funcional do software, é maior [Chen et al. 2002].

Diante disto, a técnica, utiliza a informação do perfil de uso para selecionar os casos de teste com maior probabilidade de serem executados por um usuário, dentre os casos de teste que exercitam modificações no modelo (a suíte resultante da terceira etapa). Executando estes casos de teste, a probabilidade de que um usuário encontre uma falta de regressão é menor, pois estas faltas seriam capturadas durante o teste de regressão.

Esta etapa também utiliza a matriz de similaridade, no entanto, os valores de similaridade de cada linha são divididos pela probabilidade do respectivo caso de teste da versão delta. Nesta técnica, assim como na técnica proposta por Bertolino et al., o valor de probabilidade de um caso de teste é obtido multiplicando os valores de probabilidade especificados nas transições do caso de teste. Para o nosso exemplo, estes valores de probabilidade podem ser observados na Tabela 4.2.

É importante observar que, ao dividir o valor de similaridade pelo valor de probabilidade, os valores da matriz mudam de acordo com a probabilidade obtida para o caso de teste da respectiva linha. Dessa forma, um pequeno valor de similaridade pode aumentar significativamente, quando o caso de teste analisado apresenta uma baixa probabilidade de ser executado. A Tabela 4.4 apresenta a matriz do exemplo, após as multiplicações pelos valores de probabilidade dos casos de teste.

Após multiplicar os valores da matriz de similaridades, o testador especifica uma porcentagem de cobertura, referente à quantidade de casos de teste desejada. Este valor representa a quantidade de casos de teste que deve ser selecionada, a partir do perfil de uso especificado. Nesta etapa, os casos de teste da suíte resultante da terceira etapa são removidos até que a quantidade desejada seja alcançada.

Tabela 4.4: Matriz de similaridade após a multiplicação dos valores de probabilidade de cada linha.

Casos de teste	TC_1	TC_2	TC_3	TC_4	TC_5	TC_6
TC'_1	4	2, 6	1	1	1	1
TC'_2	2, 5	2, 5	3	2	2	1
TC'_3	20	20	16	24	8	16
TC'_4	20	20	16	16	8	8
TC'_7	8, 3	8, 3	26, 6	26, 6	13, 3	13, 3
TC'_{10}	12, 5	12, 5	20	20	40	40

Diante disto, a técnica realiza uma busca dentre os valores da matriz para remover o caso de teste correspondente à linha em que o maior valor da matriz é encontrado. Este valor representa o caso de teste mais similar aos da versão anterior e/ou menos provável de ser executado pelo usuário. Antes de realizar a próxima busca na matriz, a linha selecionada é removida da matriz, para que a busca prossiga apenas nas linhas dos casos de teste restantes. Em caso de empates nos valores da matriz, a linha a ser removida é escolhida aleatoriamente, dentre as que estão empatadas.

A partir da matriz da Tabela 4.4 e considerando uma cobertura de 50% dos caso de teste da suíte obtida na terceira etapa, iniciamos a execução da quarta etapa da técnica. A partir da cobertura especificada é desejada uma suíte com metade dos casos de teste. Portanto, os casos de teste serão removidos da suíte, e da matriz, até restarem apenas 3 casos de teste.

Durante a primeira busca na matriz, o maior valor encontrado é 40, na última linha da matriz. Dessa forma, o TC'_{10} é removido da suíte e da matriz. Uma vez que a cobertura não foi atingida, é realizada outra busca na matriz. Desta vez, é encontrado o valor 26 na linha correspondente a TC'_7 que é, então, removido. É possível observar que estes dois casos de teste apresentam a menor cobertura de modificações e os menores valores de probabilidade encontrados na suíte analisada. Portanto, WSA-RT os remove primeiro.

Durante a próxima busca, 24 é encontrado como maior valor da matriz, resultando na remoção do caso de teste TC'_3. Após esta remoção, a cobertura especificada é alcançada e a técnica encerra a execução. A suíte resultante apresenta os casos de teste TC'_1, TC'_2 e TC'_4. Como é possível observar na Tabela 4.2, os casos de teste da suíte resultante, cobrem as modificações realizadas e apresentam os maiores valores de probabilidade, de acordo com

o perfil de uso especificado.

É importante observar que foram selecionados casos de teste que cobrem as modificações, tanto as adições como a remoção de transições realizadas no modelo. Dessa forma, é possível reduzir a quantidade de casos de teste executados e cobrir os casos de teste que podem revelar faltas de regressão críticas para o produto.

4.6 Considerações Finais do Capítulo

Neste capítulo foi apresentada a técnica WSA-RT, proposta neste trabalho. Este trabalho foi inspirado na técnica desenvolvida por Bertolino et al. [Bertolino et al. 2008], porém, aplicada para o contexto de teste de regressão. Dessa forma, alguns elementos da técnica foram utilizados (função de similaridade e o formato do perfil de uso), enquanto que outros foram adaptados para o contexto de regressão (estrutura da matriz e o processo de seleção).

Uma descrição resumida de cada etapa de WSA-RT é apresentada abaixo. Os detalhes destas etapas estão descritos nas seções deste capítulo, ilustrados através de um exemplo.

1. **Construção da matriz de similaridade:** Calculamos a similaridade entre os pares de casos de teste, de versões diferentes.

2. **Seleção baseada nos casos de teste obsoletos:** Iteramos sobre as colunas da matriz para identificar, e selecionar, os casos de teste, da versão delta, mais similares aos casos de teste obsoletos.

3. **Seleção baseada em similaridade dos casos de testes adicionados ou modificados:** Iteramos sobre as linhas da matriz para identificar, e selecionar, os casos de teste da versão base que cobrem as transições adicionadas e cujo rótulo foi alterado.

4. **Seleção baseada no perfil de uso:** Se o testador deseja reduzir ainda mais a quantidade de casos de teste, o perfil de uso é incorporado na matriz de similaridade para representar a similaridade e a probabilidade de execução dos casos de teste. A seleção então ocorre de forma similar a WSA [Bertolino et al. 2008].

Após desenvolver esta técnica, foi iniciado o processo do estudo experimental, seguindo o processo descrito no Capítulo 2, seção 2.5. Nos próximos capítulos serão apresentadas

as etapas do processo experimental realizadas com o objetivo de investigar e avaliar WSA-RT e as demais técnicas de re-teste seletivo, baseado em especificação, consideradas neste trabalho.

Capítulo 5

Definição e Planejamento do Estudo Experimental

Neste capítulo serão descritos a definição e o planejamento do estudo experimental. Os elementos a serem apresentados seguem a perspectiva de definição e planejamento do experimento, caracterizando a metodologia do trabalho.

Os aspectos do processo seguem as diretrizes propostas por Wohlin et. al [Wohlin et al. 2000], como: a hipótese geral, os objetos de estudo, a seleção do contexto, a definição das variáveis dependentes e independentes, a definição do projeto experimental, as hipóteses investigadas, dentre outros.

Os demais capítulos deste trabalho apresentam informações que complementas as que são apresentadas neste capítulo, como por exemplo, a fundamentação a respeito das propriedades da técnica (Capítulo 2), as técnicas analisadas (Capítulos 3 e 4) e os modelos e implementação utilizados (Capítulo 6).

As seções a seguir contemplam a caracterização do experimento que será realizado para avaliar as propriedades das técnicas de re-teste seletivo em questão. Estas seções estão organizadas como segue: inicialmente os elementos do experimento são definidos na Seção 5.1; na Seção 5.2 é apresentado o contexto do experimento; a Seção 5.3, por sua vez, descreve as variáveis do experimento; a Seção 5.4 apresenta o fator, e os respectivos níveis utilizados; na Seção 5.5 são enunciadas as hipóteses que se deseja investigar. A Seção 5.6 apresenta a caracterização do papel dos sujeitos no experimento, seguida pela Seção 5.7 onde é discutido o objeto utilizado no experimento, i.e. a especificação submetida como entrada para as técni-

cas. Os elementos da instrumentação são discutidos na Seção 5.8, enquanto que a Seção 5.9 apresenta alguns aspectos de implementação do ambiente e elementos da execução do estudo experimental. A Seção 5.10 caracteriza o projeto experimental seguido para a avaliação das hipóteses, e por fim, é apresentada, na Seção 5.11 a discussão envolvendo as ameaças à validade.

5.1 Definição do experimento

A partir do problema que motivou este trabalho, o experimento tem o objetivo de expor as limitações e benefícios das técnicas de re-teste seletivo baseado em especificação, sob um conjunto comum de propriedades. Estas propriedades são definidas na literatura e contemplam a inclusão, precisão, eficiência, potencial de redução e a densidade de faltas das técnicas de re-teste seletivo. Sob esta perspectiva, é possível definir uma hipótese geral que se deseja investigar:

"*As técnicas de re-teste seletivo são diferentes quanto à inclusão, precisão, eficiência, potencial de redução e densidade de faltas*."

A partir desta hipótese, os elementos do estudo experimental foram definidos seguindo o *template* de definição proposto por Wohlin et al. [Wohlin et al. 2000] (apresentado no Capítulo 2, deste trabalho). Os elementos definidos através do template são descritos abaixo.

Analisar ***Técnicas de re-teste seletivo baseado em especificação***
com o propósito de ***Investigação***
com respeito às ***Propriedades das técnicas***
do ponto de vista do ***Testador***
no contexto de ***Teste de regressão***

Diante dos elementos do *template*, foram definidos os seguintes elementos:

- **Objetos de estudo:** As técnicas de re-teste seletivo que serão executadas no estudo expeirmental.

- **Propósito:** O propósito é investigar estas técnicas, ou seja, obter informações a respeito de cada técnica, comparando e observando desempenho de cada técnica com relação ao foco de qualidade.

- **Foco da qualidade:** Esta característica reflete os efeitos sob estudo, ou seja, as características observadas pelo estudo experimental. Durante o estudo experimental, os elementos do foco de qualidade são referenciados como as variáveis dependentes do experimento.

- **Perspectiva:** A perspectiva utilizada é a do testador, ou seja, a pessoa, do processo de desenvolvimento, responsável pelo processo de teste, e portanto, pela utilização das técnicas.

- **Contexto:** O contexto utilizado é o mesmo em que as técnicas são aplicadas; o teste de regressão. Dessa forma, são considerados elementos como modificações, faltas de regressão, análise de dependências, e demais aspectos do teste de regressão.

Os outros elementos definidos foram os sujeitos e objetos utilizados no experimento. Os sujeitos são testadores responsáveis por configurar algumas das técnicas analisadas. Para este estudo foi definido apenas um objeto, que é uma especificação utilizada como entrada para as técnicas de re-teste seletivo. Mais detalhes sobre os sujeitos e o objeto especificados para este estudo são apresentados, respectivamente, nas Seções 5.6 e 5.7 deste capítulo.

A partir dos elementos definidos nesta etapa do experimento, é iniciada a etapa de planejamento. As seções a seguir contemplam os elementos de planejamento do estudo experimental realizado neste trabalho.

5.2 Seleção do Contexto

A primeira etapa do planejamento é a definição do contexto de realização do experimento. Dentre os contextos apresentados por Wohlin et. al [Wohlin et al. 2000], este experimento utiliza o contexto de "*Student vs Professional*". Ou seja, o experimento é caracterizado pela utilização de estudantes com experiência na área em que o experimento é realizado, para assumir o papel de profissionais, com o objetivo de atuar como sujeitos do experimento.

No caso deste experimento, foram escolhidos alunos de um curso de graduação e pós-graduação em ciência da computação com experiência na área de teste de software. O objetivo desta escolha foi diminuir o custo do experimento, uma vez que não havia disponibilidade de recursos financeiros para a contratação de testadores reais para a realização do experimento.

5.3 Variáveis

Um dos principais elementos de um estudo experimental são as variáveis dependentes e independentes. A partir destas variáveis, são estruturados o projeto experimental, os testes de hipóteses, e o processo de execução do experimento. A partir das definições de variáveis (Seção 2.5.2 do Capítulo 2), foram definidas, para este trabalho:

- **Variáveis Dependentes:** As propriedades (inclusão, precisão, eficiência, potencial de redução e densidade de faltas) obtidas durante a execução das técnicas.
- **Variáveis Independentes:** As técnicas e os seus respectivos parâmetros de configuração fornecidos pelo sujeito, quando necessário.

A descrição da cada variável dependente está no Capítulo 2, Seção 2.4.1, onde é explicado o que caracteriza cada uma destas propriedades e como elas podem ser obtidas observando a suíte de testes reduzida. Considerando que os casos de teste obtidos neste experimento são abstratos, i.e. não são executáveis automaticamente, é inviável seguir um modelo de custos utilizado na análise de técnicas baseadas em código. Este modelo, proposto por Leung e White [Leung and White 1991], mede a eficiência da técnica através dos custos de tempo de obtenção, execução e análise da suíte de testes reduzida. Portanto, para este estudo experimental, o único aspecto observado para a variável dependente de eficiência foi o tempo de execução da técnica.

Uma outra propriedade investigada em estudos experimentais de técnicas de re-teste seletivo é a generalidade da técnica [Rothermel and Harrold 1996, Mahdian et al. 2009]. Devido à natureza qualitativa e subjetiva desta propriedade, e às restrições de cronograma e escopo deste estudo experimental, não foi possível analisar estatisticamente (a partir de intervalos de confiança ou testes de hipóteses) a generalidade das técnicas. Portanto, foi realizada uma

breve análise descritiva a respeito da generalidade observada em cada técnica. Esta análise investiga aspectos como a facilidade de aplicação e a versatilidade de uso da técnica de re-teste seletivo através de elementos como: as dependências ferramentais e operacionais da técnica, a necessidade de recursos humanos (i.e. testadores, desenvolvedores e analistas de sistema) ou financeiros para executar a técnica, a natureza da análise realizada, dentre outros elementos.

5.4 Fator e Níveis

Este experimento apresenta um único fator, que se trata da técnica de re-teste seletivo a ser utilizada. Diante disto, cada nível deste fator é uma técnica de re-teste seletivo utilizada neste experimento. A descrição de cada técnica é apresentada nos Capítulos 3 e 4. Esses níveis são categóricos e referenciados a partir dos identificadores, apresentados a seguir, que serão utilizados ao longo do texto. Os níveis e os respectivos identificadores são:

- T_1: Seleção baseada em Análise de Dependência em Máquinas de Estados Finitas Estendidas [Chen et al. 2007];
- T_2: Seleção baseada em Análise de Riscos e Diagramas de Atividade em UML [Chen et al. 2002];
- T_3: *Weighted Similarity Approach for Regression Testing* (WSA-RT);
- T_4: Técnica baseada em *clustering* de Laski e Szermer [Laski and Szermer 1992];
- T_5: Seleção aleatória de casos de teste.

As técnicas T_2 e T_3 necessitam de informações fornecidas por um testador para que possam ser executadas. Os sujeitos do experimento atuaram como testadores e devem, portanto, fornecer estas informações como parte da configuração da técnica. Além da configuração do sujeito, T_2 e T_3 necessitam de um valor de cobertura, que especifica a quantidade de casos de teste que devem ser selecionados durante as etapas de, respecitivamente, análise de risco, e seleção baseada em perfis de uso.

Estas duas etapas, para ambas as técnicas, são realizadas para complementar a seleção realizada através da análise das modificações entre as versões. Neste estudo expeirmental,

estas etapas foram realizadas considerando uma cobertura de 25%, para as duas técnicas. Este valor foi escolhido pois, não é grande o suficiente para comprometer o potencial de redução das técnicas, nem pequeno o suficiente para descaracterizar o desempenho fornecido pelas respectivas etapas de seleção de T_2 e T_3.

Por sua vez, T_5 é dependente de um fator de cobertura que indique o limiar de seleção da suíte de testes de regressão (i.e. a quantidade de casos de teste que devem ser selecionados). Diante disto, foram escolhidos 3 valores de cobertura para a técnica T_5: 25%, 50% e 75%. Estes valores, utilizados em um estudo experimental realizado por Graves et. al [Graves et al. 1998], são escolhidos pois fornecem uma visão da suíte reduzida ao serem realizadas, respectivamente, uma grande, uma média e uma pequena redução no tamanho da suíte de regressão.

O impacto destas diferentes configurações no desempenho das técnicas, com relação às variáveis dependentes analisadas, deve ser levado em consideração durante a realização do projeto experimental. Diante disto, foi realizada uma investigação fundamentada em recursos estatísticos para avaliar se as diferentes configurações afetavam significativamente os resultados da técnica (detalhes desta investigação estão no Apêndice B). Ao detectar uma diferença estatisticamente significativa no resultado da técnica para configurações distintas, a técnica e a respectiva configuração eram considerados um nível do fator. Caso contrário, os dados resultantes das diferentes configurações eram consolidades a partir de uma média aritmética.

5.5 Hipóteses

As hipóteses caracterizam um dos elementos mais críticos do experimento. A partir das hipóteses formuladas em um experimento, é possível aceitar, ou rejeitar, os resultados esperados da execução do experimento. Diante disto, é importante que as hipóteses sejam bem estruturadas no contexto do experimento, e não possuam ambigüidades.

Estas hipóteses são obtidas a partir da hipótese geral estabelecida durante a definição do experimento, e devem contemplar as variáveis dependentes e independentes [Wohlin et al. 2000]. Portanto, a partir da hipótese geral que motivou a realização do experimento, foram derivadas as hipóteses nulas, e alternativas. Considerando T_x, $x = 1, 2, \ldots, 5$,

uma das técnicas utilizadas no experimento, as propriedades $I(T_x)$ (inclusão), $P(T_x)$ (precisão), $E(T_x)$ (eficiência), $R(T_x)$ (potencial de redução) e $D(T_x)$ (densidade de faltas) foram utilizadas para definir as seguintes hipóteses nulas e as respectivas hipóteses alternativas[1]:

- **Hipóteses Nula 1 ($H0_1$):** Todas as técnicas de re-teste seletivo selecionadas para o experimento são semelhantes com relação à propriedade de inclusão.

- **Hipótese Alternativa 1 ($H1_1$):** Todas as técnicas de re-teste seletivo selecionadas para o experimento possuem um comportamento diferenciado com relação à propriedade de inclusão.

$$H0_1\text{: } I(T_1) = I(T_2) = I(T_3) = I(T_4) = I(T_5)$$
$$H1_1\text{: } I(T_1) \neq I(T_2) \neq I(T_3) \neq I(T_4) \neq I(T_5)$$

- **Hipóteses Nula 2 ($H0_2$):** Todas as técnicas de re-teste seletivo selecionadas para o experimento são semelhantes com relação à propriedade de precisão.

- **Hipótese Alternativa 2 ($H1_2$):** Todas as técnicas de re-teste seletivo selecionadas para o experimento possuem um comportamento diferenciado com relação à propriedade de precisão.

$$H0_2\text{: } P(T_1) = P(T_2) = P(T_3) = P(T_4) = P(T_5)$$
$$H1_2\text{: } P(T_1) \neq P(T_2) \neq P(T_3) \neq P(T_4) \neq P(T_5)$$

- **Hipóteses Nula 3 ($H0_3$):** Todas as técnicas de re-teste seletivo selecionadas para o experimento são semelhantes com relação à propriedade de eficiência.

- **Hipótese Alternativa 3 ($H1_3$):** Todas as técnicas de re-teste seletivo selecionadas para o experimento possuem um comportamento diferenciado com relação à propriedade de eficiência.

$$H0_3\text{: } E(T_1) = E(T_2) = E(T_3) = E(T_4) = E(T_5)$$
$$H1_3\text{: } E(T_1) \neq E(T_2) \neq E(T_3) \neq E(T_4) \neq E(T_5)$$

[1]É importante observar que as inequações expressas como: $a \neq b \neq c$, representam $a \neq b \wedge b \neq c \wedge a \neq c$. Este formato é adotado durante todo este documento.

- **Hipóteses Nula 4 ($H0_4$):** Todas as técnicas de re-teste seletivo selecionadas para o experimento são semelhantes com relação à propriedade de potencial de redução.

- **Hipótese Alternativa 3 ($H1_4$):** Todas as técnicas de re-teste seletivo selecionadas para o experimento possuem um comportamento diferenciado com relação à propriedade de potencial de redução.

$$H0_4\text{: } R(T_1) = R(T_2) = R(T_3) = R(T_4) = R(T_5)$$
$$H1_4\text{: } R(T_1) \neq R(T_2) \neq R(T_3) \neq R(T_4) \neq R(T_5)$$

- **Hipóteses Nula 5 ($H0_5$):** Todas as técnicas de re-teste seletivo selecionadas para o experimento são semelhantes com relação à propriedade de densidade de faltas.

- **Hipótese Alternativa 5 ($H1_5$):** Todas as técnicas de re-teste seletivo selecionadas para o experimento possuem um comportamento diferenciado com relação à propriedade de densidade de faltas.

$$H0_5\text{: } D(T_1) = D(T_2) = D(T_3) = D(T_4) = D(T_5)$$
$$H1_5\text{: } D(T_1) \neq D(T_2) \neq D(T_3) \neq D(T_4) \neq D(T_5)$$

Estas hipóteses, após a execução do experimento, são submetidas a testes estatísticos, que fornecem informações definitivas para rejeitar (ou não) as hipóteses nulas em favor das hipóteses alternativas. Estes testes são realizados durante a análise dos resultados do experimento, e constituem a principal fonte de informação para as conclusões obtidas no estudo experimental.

5.6 Sujeitos

Os sujeitos de um experimento, são as pessoas envolvidas com o experimento. O papel e a efetividade dessas pessoas no experimento variam de acordo com os objetivos deste. Para o experimento especificado neste trabalho, foram organizados conjuntos de testadores para participarem do experimento, em especial para o planejamento da etapa de Operação. Nesta etapa, os sujeitos devem especificar os perfis de uso da aplicação utilizada no experimento para, possibilitar a execução da técnica WSA (T_3). Além disto, os sujeitos também atuarão

na definição dos custos e riscos de cada caso de teste da suíte de testes de regressão, para possibilitar a execução da técnica de seleção baseada em análise de riscos (T_2).

Para este experimento, foram selecionados 4 sujeitos, organizados em grupos compostos por 2 testadores, ou seja, foram utilizados 2 grupos, com 2 testadores cada. Estes grupos diferenciam-se pelo tempo de experiência de cada sujeito na área de testes, ou seja, um grupo é composto por testadores inexperientes, e o outro grupo é composto por testadores experiente. Todos os testadores fornecem uma configuração para a técnica T_2 e uma outra configuração para a técnica T_3.

As técnicas T_2 e T_3 são executadas com cada configuração fornecida, e estas configurações são utilizadas em todas as replicações do experimento. Diante disto, os dados referentes a um nível de experiência de um testador são consolidados através da média aritmética entre os dados dos testadores do respectivo grupo. Ou seja, os resultados de um testador inexperiente são calculados através da média aritmética dos resultados obtidos, utilizando as configurações dos testadores inexperientes. Os dados de um testador experiente são obtidos da mesma forma, utilizando, porém, as configurações dos testadores experientes.

De acordo com o contexto selecionado para o experimento (apresentado na Seção 5.2), todos os sujeitos são estudantes dos programas de Pós-graduação em Informática e da Graduação em Ciência da Computação, ambos da Universidade Federal de Campina Grande. Para a seleção destes sujeitos, foram observados aspectos curriculares da formação em Teste de Software. Estes aspectos são:

- Nível de formação (graduação ou pós-graduação).
- Quantidade de meses nos quais desenvolveu algum trabalho ou pesquisa na área de teste de software;
- Quantidade de trabalhos produzidos para empresas ou publicados em conferências da área de testes.

Por questões éticas em um estudo experimental, os nomes dos sujeitos que participam do estudo experimental não é divulgado neste documento. Estes são referenciados apenas por: "*testador inexperiente 1*", "*testador inexperiente 2*", "*testador experiente 1*" e "*testador experiente 2*". Sob a mesma justificativa, as informações curriculares destes sujeitos também não são apresentadas.

5.7 Objeto do Experimento

Nesta subseção é descrito o objeto utilizado no estudo experimental. Considerando que as técnicas necessitam de um modelo para executar, pois o re-teste seletivo realizado é baseado em especificação, é necessário definir a especificação de um software, utilizada como entrada para estas técnicas. Com o objetivo de manter um nível de controle sobre este elemento do estudo experimental, i.e. evitar ameaças à validade por problemas no modelo, decidimos escolher uma especificação disponível e familiar para os autores deste estudo experimental.

Dessa forma, escolhemos a especificação da ferramenta LTS-BT (*Labeled Transition System Based Testing*). LTS-BT é uma ferramenta utilizada para a geração e seleção automática de casos de teste para Testes Baseados em Modelos [Cartaxo et al. 2008]. Esta ferramenta é familiar para os autores deste estudo experimental, o que permitiu uma maior precisão na verificação da execução das técnicas de re-teste seletivo.

Para a execução das técnicas, é usada uma especificação dos casos de uso da ferramenta. Estes casos de uso seguem o formato proposto por Cabral e Sampaio [Cabral and Sampaio 2008]. A partir desta especificação, são gerados os modelos (o diagrama de atividades, o diagrama de estados, o grafo de fluxo de controle e o *Labeled Transition System*) sobre os quais as técnicas executam. Esta especificação possui um tamanho pequeno quando comparada com a especificação de sistemas complexos que possuem diversos componentes. No entanto, para a realização de análises de técnicas e estudos experimentais, este objeto é adequado pois é possível manter um controle e uma rastreabilidade de todos os seus elementos, durante a etapa de operação, assim como na etapa de análise dos resultados obtidos da execução do experimento.

A especificação utilizada possui 19 casos de uso, dos quais foram gerados 58 casos de teste. Esta especificação foi modificada para poder ser aplicada a um Teste de Regressão. Após a modificação, a especificação passou a gerar 65 casos de teste. O algoritmo de geração de casos de testes utilizado foi o caminhamento em profundidade nas transições do modelo, de forma que, cada cenário do modelo caracterizava um caso de teste. Como modificações na especificação, foram realizadas algumas remoções e adições de casos de uso, assim como modificações nas transições dos modelos (i.e. modificações em rótulos, ou nos vértices de origem e destino das arestas). Os detalhes da especificação dos modelos de entrada e de

faltas, modificações analisadas, da implementação e demais aspectos da isntrumentação são apresentados no Capítulo 6.

A utilização de apenas um objeto neste experimento, diminui a capacidade de generalização dos resultados deste estudo experimental. Além disto, os resultados apresentados pelo experimento estão relacionados com as características do objeto utilizado, já que as técnicas podem apresentar resultados diferentes em modelos com estruturas diferentes (e.g. modelos com laços, muitas divisões ou junções de fluxos, dentre outros). Porém, utilizando apenas um modelo para todas as técnicas obtemos uma análise mais precisa e controlada, com relação aos resultados das técnicas.

5.8 Instrumentação

Em diversas situações, a execução do experimento, ou pelo menos alguns elementos da execução, necessitam de um suporte instrumental. Este suporte instrumental caracteriza a instrumentação do experimento. Para este experimento, a instrumentação é caracterizada por três elementos: objetos, diretrizes e ferramentas de suporte [Wohlin et al. 2000]. Neste estudo experimental, estes elementos são:

- **Objeto:** A especificação da ferramenta LTS-BT (*Labelled Tansitions System Based Testing*).
- **Diretrizes:** Um modelo que descreve as modificações realizadas no sistema, documentos com instruções para os sujeitos do experimento e um modelo de faltas [Binder 1999].
- **Ferramentas de suporte:** A ferramenta LTS-BT será utilizada para dar suporte à implementação do ambiente de experimentação que será, posteriormente, incorporado à própria ferramenta.

LTS-BT foi utilizada na instrumentação do experimento, pois, é uma ferramenta que realiza geração e seleção de casos de teste implementando várias técnicas da literatura. Além de fornecer os mecanismos de geração, de acordo com as dependências das técnicas, esta ferramenta é também o ambiente de desenvolvimento e execução do experimento [Oliveira Neto and Machado 2008].

A partir da especificação de LTS-BT são propostas algumas modificações. Para manter a controlabilidade dos dados coletados e analisados, a respeito das faltas de regressão, é utilizado um modelo de faltas [Binder 1999] e um modelo para representar as modificações no objeto (ver Seção 5.7). Dessa forma, as faltas e modificações são inseridas em alguns pontos do modelo da especificação, e os casos de teste que exercitam estes pontos capturam as faltas de regressão.

5.9 Implementação

Para a realização do estudo experimental, é necessária a implementação de diversos elementos do experimento. Com o objetivo de automatizar as execuções do experimento, são implementadas: todas as técnicas analisadas no estudo experimental (Seção 5.4); os respectivos leitores de formatos de arquivos, e.g. Trivial Graph Format (TGF[2]), XML Metadata Interchange (XMI[3]); um injetor de faltas baseado no modelos de faltas especificado; a coleta de dados durante a execução do experimento; dentre outros.

Para a implementação é utilizada a Linguagem de Programação Java[4], e a IDE (*Integrated Development Environment*) Eclipse[5]. Para este experimento, desenvolvemos 68 classes, organizadas em pacotes dividindo estas classes de acordo com o contexto do experimento em que são utilizadas (e.g. execução, representação de digramas, técnicas de re-teste seletivo).

Para o fator, é utilizada uma interface, implementada por todos os níveis (i.e. a implementação de cada técnica de re-teste seletivo). Também são utilizados *design patterns* (e.g. *FactoryMethod*, *Adapter* e *Facade*) com o objetivo de fornecer flexibilidade ao ambiente de execução do experimento [Gamma et al. 1994], permitindo que, posteriormente, outras técnicas possam ser facilmente adicionadas à execução do experimento (e.g. incluir mais níveis no fator). Os detalhes da implementação são descritos na Seção 6.3 do Capítulo 6.

[2]http://www.yworks.com/products/yfiles/doc/developers-guide/tgf.html

[3]http://www.omg.org/technology/documents/formal/xmi.htm

[4]http://www.sun.com/java/

[5]http://www.eclipse.org/

5.10 Projeto Experimental

O processo de geração de conclusões sobre as hipóteses levantadas precisa levar em conta a concepção e as práticas sugeridas por um projeto experimental. Neste trabalho, são estruturados cinco projetos experimentais. Em cada projeto é avaliada uma propriedade das técnicas de re-teste seletivo ou, alternativamente, um par de hipóteses nula e alternativa. A Figura 5.1 sintetiza o que é coletado a cada experimento realizado.

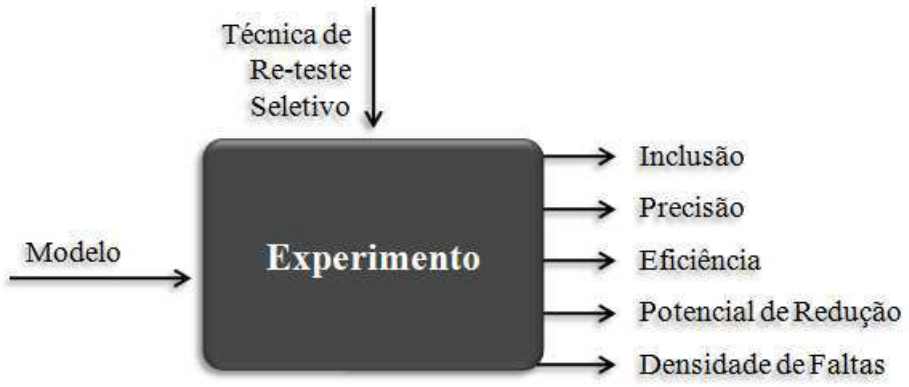

Figura 5.1: Visão geral dos experimentos que serão realizados.

O nível de confiança escolhido para este estudo experimental é de 95% ou, equivalentemente, o nível de significância é de $\alpha = 0,05$. Esta escolha leva em consideração sugestões da literatura de estatística para realização de investigações experimentais [Jain 1991]. Em cada projeto experimental, é calculado o tamanho mínimo da amostra para obter conclusões com significância estatística. Os valores de precisão e a quantidade de dados utilizados neste cálculo são especificados na subseção de cada projeto experimental.

É possível observar que todos os projetos experimentais apresentam um único fator, e mais de dois níveis categóricos. Diante disto, o teste estatístico recomendado é *One-way* ANOVA. Durante a análise, as premissas de ANOVA são investigadas para verificar se o teste pode ser aplicado nos dados coletados. Se for observado que os dados não respeitam estas premissas, é aplicado, então, o teste não-paramétrico de Kruskal-Wallis.

A caracterização de cada projeto experimental realizado é apresentado nas seções a seguir. Antes de tal caracterização, são apresentadas algumas considerações sobre os diferentes parâmetros de configurações das técnicas.

5.10.1 Considerações sobre as configurações das técnicas

No projeto experimental é considerado que cada técnica é auto-contida, ou seja, é assumido que as mesmas possuem todos os parâmetros para o seu funcionamento adequado. Porém, é possível configurar alguns destes parâmetros de forma diferente e nem todos os parâmetros constituem uma mesma entrada para todas as técnicas, afinal, cada técnica executa um algoritmo diferente.

Em algumas situações, vale ressaltar, é interessante analisar o comportamento de uma mesma técnica sob diferentes condições de inicialização, verificando se esta mudança caracteriza, por exemplo, em um desempenho melhor sob um determinado aspecto. Esta diferenciação é realizada por meio de testes estatísticos e está reportada em detalhes no Apêndice B.

Um resumo dos resultados desta análise é apresentada na Tabela 5.1. Nesta tabela, as colunas apresentam os resultados das comparações, enquanto que as linhas representam cada variável dependente analisada. É importante observar que as inequações expressas como: $T_{x1} \neq T_{x2} \neq T_{x3}$, representam $T_{x1} \neq T_{x2} \wedge T_{x2} \neq T_{x3} \wedge T_{x1} \neq T_{x3}$.

Tabela 5.1: Resultados da análise da diferença entre as configurações.

	T_2	T_3	T_5
Inclusão	$T_{2i} = T_{2e}$	$T_{3i} \neq T_{3e}$	$T_{5-25\%} \neq T_{5-50\%} \neq T_{5-75\%}$
Precisão	$T_{2i} = T_{2e}$	$T_{3i} \neq T_{3e}$	$T_{5-25\%} \neq T_{5-50\%} \neq T_{5-75\%}$
Eficiência	$T_{2i} \neq T_{2e}$	$T_{3i} \neq T_{3e}$	$T_{5-25\%} \neq T_{5-50\%} \neq T_{5-75\%}$
Potencial de Redução	$T_{2i} = T_{2e}$	$T_{3i} \neq T_{3e}$	$T_{5-25\%} \neq T_{5-50\%} \neq T_{5-75\%}$
Densidade de Faltas	$T_{2i} = T_{2e}$	$T_{3i} = T_{3e}$	$T_{5-25\%} = T_{5-50\%} = T_{5-75\%}$

As técnicas: T_{2i} e T_{2e} representam, respectivamtente, a técnica T_2 (seleção baseada em riscos) configurada por testadores inexperientes e experientes; T_{3i} e T_{3e}, representam a técnica T_3 (WSA-RT) configurada por testadores inexperientes e experientes, respectivamente; $T_{5-25\%}, T_{5-50\%}$ e $T_{5-75\%}$ representam a técnica T_5 (seleção aleatória) com 25%, 50% e 75% de cobertura.

Nos casos em que não é encontrada diferença estatisticamente significativa entre os níveis, é utilizada uma média aritmética entre os valores de cada configuração. Se há difer-

ença nos resultados com diferentes configurações, cada configuração é considerada como um nível do fator.

As seções a seguir utilizam os resultados obtidos nessa investigação para caracterizar os diferentes níveis do fator técnica de re-teste seletivo. As justificativas para as diferenças observadas, são apresentadas na análise de cada técnica (Seções 7.7.2, 7.7.3 e 7.7.5 do Capítulo 7).

5.10.2 Projeto Experimental 1 – Hipóteses $H0_1$ e $H1_1$

Este projeto experimental visa investigar as hipóteses $H0_1$ e $H1_1$, que representam a igualdade ou diferença das cinco técnicas de re-teste seletivo no tocante à propriedade de *inclusão*. Para determinar a quantidade de experimentos, necessários para garantir significância estatística, são realizadas 40 replicações de experimentos com cada uma das técnicas, considerando uma precisão de $\pm 5\%$. Esta quantidade de experimentos é sumarizada na Tabela 5.2.

Tabela 5.2: Dados estatísticos para o tamanho de amostra mínimo de cada uma das técnicas de re-teste seletivo no tocante à inclusão.

Técnica	T_1	T_2	T_{3i}	T_{3e}	T_4	$T_{5-25\%}$	$T_{5-50\%}$	$T_{5-75\%}$
Média	50	92,85	33,92	35,93	78,57	24,82	49,46	71,78
Desvio Padrão	0,00	0,00	1,61	1,81	0,00	6,15	7,03	7,448
Replicações Necessárias	1	1	4	4	6	95	31	17

O número de replicações calculado em função dos dados amostrais é utilizado ao longo do trabalho para assegurar significância estatística. Ou seja, este número representa a quantidade de execuções da técnica, e portanto, a quantidade de dados de inclusão medidos na suíte reduzida, necessários para que o resultado de inclusão obtido possa ser fundamentado por análises estatísticas, fornecendo maior credibilidade às conclusões do experimento. A partir dos aspectos estruturados no projeto experimental, é possível observar que a análise de inclusão envolve um único fator com 8 níveis categóricos (T_1, T_2, T_{3i}, T_{3e}, T_4, $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$).

5.10.3 Projeto Experimental 2 – Hipóteses $H0_2$ e $H1_2$

O objetivo deste projeto experimental é investigar as hipóteses $H0_2$ e $H1_2$. Estas hipóteses investigam a igualdade ou diferença das cinco técnicas de re-teste seletivo no tocante à propriedade de *precisão*. Assim como o projeto experimental de inclusão, o projeto experimental de precisão é caracterizado por um único fator com oito níveis categóricos (T_1, T_2, T_{3i}, T_{3e}, T_4, $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$).

A quantidade de experimentos necessários para garantir significância estatística neste projeto experimental é calculada considerando uma precisão de $\pm 5\%$, e 40 replicações de experimentos com cada um dos níveis. Os resultando obtidos estão sumarizados na Tabela 5.3.

Tabela 5.3: Dados estatísticos para o tamanho de amostras mínimo de cada uma das técnicas de re-teste seletivo no tocante à precisão.

Técnica	T_1	T_2	T_{3i}	T_{3e}	T_4	$T_{5-25\%}$	$T_{5-50\%}$	$T_{5-75\%}$
Média	28,16	85,13	29,91	25,89	35,13	24,38	49,02	75,36
Desvio Padrão	3,34	0,00	0,95	1,01	0,00	4,58	5,35	5,44
Replicações Necessárias	23	1	2	3	1	55	19	9

5.10.4 Projeto Experimental 3 – Hipóteses $H0_3$ e $H1_3$

O Projeto Experimental 3 é construído com o objetivo de investigar as hipóteses $H0_3$ e $H1_3$, referentes à propriedade de *eficiência* das técnicas de re-teste seletivo. Este projeto experimental apresenta um fator, e 9 níveis (T_1, T_{2i}, T_{2e}, T_{3i}, T_{3e}, T_4, $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$).

Recursos estatísticos são utilizados para determinar a quantidade de experimentos necessários para garantir significância estatística. Para determinar esta quantidade, são realizadas 40 execuções do experimento considerando precisão de $\pm 12,5\%$. O valor de precisão de $\pm 12,5\%$, é maior que nos demais projetos experimentais, pois, para a eficiência, sob uma precisão de $\pm 5\%$, os valores de amostras obtidos são altos (em torno de 1000 amostras). Diante disto, aumentamos a precisão para obter um tamanho de amostras menor.

É importante ressaltar que esta diferença não afeta significativamente os resultados pois a eficiência é medida em nanossegundos, uma dimensão pequena para tempo de execução

na maioria das técnicas (que utilizam, geralmente, milissegundos para executar). Com estes dados, os tamanhos de amostras para cada nível foram calculados, resultando nos dados sumarizados na Tabela 5.4.

Tabela 5.4: Dados estatísticos para o tamanho de amostras mínimo de cada uma das técnicas de re-teste seletivo no tocante à eficiência.

Técnica	T_1	T_{2i}	T_{2e}	T_{3i}	T_{3e}
Média	$1,52 \cdot 10^{11}$	$36632908,56$	$27676304,38$	$14707116,6$	14785851,98
Desvio Padrão	3269801419	$8384937,16$	$13221207,2$	$640195,29$	$1339077,72$
Replicações Necessárias	1	13	57	1	3

Técnica	T_4	$T_{5-25\%}$	$T_{5-50\%}$	$T_{5-75\%}$
Média	$17507,07$	$17507,07$	$12481,63$	$9127,41$
Desvio Padrão	$9968,61$	$9968,61$	$6383,83$	$9347,98$
Replicações Necessárias	80	80	65	258

5.10.5 Projeto Experimental 4 – Hipóteses $H0_4$ e $H1_4$

O *potencial de redução* das técnicas de re-teste seletivo analisadas no experimento é contemplado pelas hipóteses $H0_4$ e $H1_4$ do Projeto Experimental 4. Verificando a execução das técnicas com diferentes configurações, é observado que este projeto experimental apresenta um único fator, com 8 níveis categóricos (T_1, T_2, T_{3i}, T_{3e}, T_4, $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$).

Para determinar a quantidade de execuções necessárias para obter significância estatística, são utilizados os resultados de 40 execuções de cada técnica sob uma precisão de $\pm 5\%$. Os tamanhos de amostras obtidos para cada técnica pode ser observado na Tabela 5.5.

Tabela 5.5: Dados estatísticos para o tamanho de amostras mínimo de cada uma das técnicas de re-teste seletivo no tocante ao potencial de redução.

Técnica	T_1	T_2	T_{3i}	T_{3e}	T_4	$T_{5-25\%}$	$T_{5-50\%}$	$T_{5-75\%}$
Média	60	12,06	70,05	68,46	46,15	75,38	50,76	26,15
Desvio Padrão	3,11	0,00	0,91	0,92	0,00	0,00	0,00	0,00
Replicações Necessárias	5	1	1	1	1	1	1	1

5.10.6 Projeto Experimental 5 – Hipóteses $H0_5$ e $H1_5$

As hipóteses $H0_5$ e $H1_5$, referentes à *densidade de faltas* das técnicas, são investigadas no Projeto Experimental 5. Ao investigar o efeito das diferentes configurações das técnicas, não é observada diferença estatisticamente significativa na densidade de faltas destas técnicas. Diante disto, são consideradas as médias aritméticas das densidades de faltas de cada configuração para as respectivas técnicas analisada. Portanto, este projeto experimental apresenta um único fator, com 5 níveis categóricos (T_1, T_2, T_3, T_4 e T_5).

Após observar que as configurações não afetam, significativamente, os resultados da suíte, é investigada a quantidade de execuções necessárias para obter uma análise com significância estatística. Assim como nos demais projetos experimentais, são utilizados dados de 40 execuções de cada técnica, sob uma precisão de $\pm 5\%$. Os resultados são apresentados na Tabela 5.6.

Tabela 5.6: Dados estatísticos para o tamanho de amostras mínimo de cada uma das técnicas de re-teste seletivo no tocante à densidade de faltas.

Técnica	T_1	T_2	T_3	T_4	T_5
Média	41,39	38,23	49,16	62,85	41,90
Desvio Padrão	5,57	0,00	1,89	0,00	4,37
Replicações Necessárias	28	1	3	1	17

5.11 Avaliação de Validade

Nesta seção será discutida a avaliação de validade definida para este estudo experimental. É importante considerar os aspectos de validade do experimento ainda no planejamento, para que uma avaliação adequada dos resultados do experimento seja planejada. Os aspectos que caracterizam uma validade adequada, assim como, os aspectos de avaliação de validade são apresentadas no Capítulo 2, Seção 2.5.3.

A validade de estudos experimentais está sujeita a ameaças. Para identificar as ameaças à validade deste estudo experimental, é utilizado um *checklist* definido por Cook [Cook and Campbell 1979] em que as ameaças são identificadas de acordo com o tipo da validade (interna, externa, de conclusão ou construção).

Neste estudo experimental, a validade de conclusão é mantida através dos elementos do projeto e do design experimental, como o tamanho das amostras e os testes estatísticos utilizados. Para manter a significância estatística dos dados, a quantidade de execuções realizadas é maior que a quantidade definida no tamanho das amostras de cada projeto experimental (Seção 5.10).

Diante disto, são planejadas 100 execuções de cada técnica, para as variáveis dependentes investigadas. Em uma primeira execução, são capturados os dados de inclusão e eficiência; enquanto que, em uma segunda execução, são coletados os dados de precisão, potencial de redução e densidade de faltas. Com esta divisão na coleta de dados, foi possível iniciar a análise das variáveis de inclusão e eficiência, de forma que, ao término da segunda execução, parte da análise já estaria concluída, evitando atrasos no cronograma por eventuais problemas de execução (e.g. quedas de energia, erros na manipulação dos arquivos, dentre outros).

O único tamanho de amostra planejado que não pode ser atingido, ao realizar as 100 execuções é o da técnica $T_{5-75\%}$ (onde são necessárias 258 execuções). Porém, como esclarecido na Seção 5.10.4, devido à dimensão dos dados (nanossegundos) e o nível de precisão utilizados, esta redução na quantidade de execuções para esta técnica não caracteriza um impacto significativo nos resultados de eficiência.

Todas as análises realizadas consideram um nível de confiança de 95%, i.e. um nível de significância $alpha = 0,05$. Escolhemos este nível de confiança através de sugestões na literatura de estatística [Jain 1991, Siegel and Junior 1988]. Além disto, este nível de confiança fornece uma perspectiva satisfatória a respeito das zonas de rejeição das hipóteses nulas de cada projeto experimental.

A principal ameaça à validade de conclusão está no poder e uso apropriado dos recursos estatístico, na violação de premissas de testes estatísticos e na confiabilidade da implementação dos tratamentos. Para lidar com as ameaças na utilização dos recursos estatísticos, o investigador realizou um amplo estudo dos aspectos estatísticos com relação à experimentação em engenharia de software.

Durante esse estudo, o investigador cursou as disciplinas: T.E.C.C. de Fundamentos de Pesquisa em Ciência da Computação, e Engenharia de Software Experimental do Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande, cujos respectivos programas contemplam muitos conceitos de investigação e análise estatís-

tica. Os conceitos vistos e aplicados durante estas disciplinas contribuíram significativamente para a fundamentação estatística utilizada neste experimento.

Uma das principais ameaças à validade interna é o controle do experimento. Considerando que a execução das técnicas é automática, o controle é inserido na implementação e execução dos algoritmos, de forma que o ambiente de execução do experimento não seja influenciado por outro processo, ou programa, da máquina. Ou seja, não são utilizados algoritmos distribuídos ou concorrentes na implementação, para evitar problemas de controle, como condições de corridas e *deadlocks*.

Dessa forma, cada algoritmo executa por vez, e os dados referentes à técnica em execução são capturados e armazenados. Além disto, a máquina em que o experimento executa, não é utilizada durante a etapa de execução, evitando que processos de usuários prejudiquem a execução das técnicas.

Os outros aspectos controlados para não prejudicar a validade interna são os sujeitos e o objeto do experimento. A atuação dos sujeitos é controlada através de documentos que descrevem os valores de configuração que devem ser especificados. Neste documento é descrito o que cada valor representa, assim como, as faixas de valores que devem ser utilizadas.

O objeto, por sua vez, é controlado através de suas características estruturais (vértices e transições). Para manter um controle na rastreabilidade da cobertura de faltas e modificações, estruturas de laços, a integração de muitos fluxos alternativos são evitados durante a modelagem do objeto. Além de manter um controle na rastreabilidade do modelo, a ausência destas estruturas complexas previne uma sobrecarga (*overhead*) no processo de geração de casos de teste, e análise de dependência.

As ameaças à validade de construção encontradas para este experimento estão relacionadas com o aspecto social do experimento, ou seja, os sujeitos. Um tipo comum de ameaça é observada quando o sujeito procurar guiar suas ações de acordo com palpites que possam fornecer um melhor resultado para, por exemplo, rejeitar (ou não) a hipótese nula. Para evitar esta ameaça, as hipóteses dos projetos experimentais não foram reveladas para os sujeitos.

Wohlin et. al [Wohlin et al. 2000] apresenta uma breve descrição de como o próprio investigador pode caracterizar uma ameaça à validade de construção. Esta ameaça é baseada na própria expectativa do investigador acerca do experimento. Para evitar esta ameaça, Wohlin

et. al recomenda que outras pessoas estejam envolvidas na elaboração e no processo do estudo experimental. Neste experimento, a orientadora do investigador desenvolveu este papel.

Os papéis e perfis dos sujeitos, o objeto utilizado, assim como a instrumentação e os projetos experimentais apresentados neste estudo caracterizam a validade externa dos resultados obtidos.

Diante disto, os resultados podem ser generalizados para configurações semelhantes a estas, em especial quando consideramos o objeto. Por sua vez, as condições que limitam a generalização dos resultados do experimento caracterizam as ameaças à validade externa. Para este experimento, a principal ameaça à validade externa é o objeto.

A utilização de apenas um objeto neste experimento prejudica a generalidade, pois é possível que modelos com estruturas muito diferentes apresentem um melhor ou pior resultado para a técnica. Por outro lado, o conhecimento acerca do único objeto utilizado no experimento aumenta significativamente a precisão da análise realizada, uma vez que os aspectos estruturais e comportamentais do modelo são facilmente rastreáveis pelo investigador. Diante disto, é possível observar os aspectos estruturais do modelo que possam caracterizar o desempenho das técnicas.

É esperado que modelos com as características semelhantes (e.g. tamanho da suíte de regressão gerada ou a quantidade de vértices ou transições) ao objeto utilizado neste experimento, apresentem os mesmos resultados. Esta expectativa é baseada na quantidade de execuções realizadas e nos recursos estatísticos (testes de hipóteses, dados obtidos em grandes amostras, nível de confiança, dentre outros) utilizados durante o planejamento e a análise.

Além destas ameaças à validade outras podem ser encontradas durante as demais etapas do processo experimental. Documentar as ameaças é importante para viabilizar a reprodução do experimento por demais pesquisadores. Além disto, estas ameaças podem caracterizar a motivação de um melhoramento no experimento, aumentando a dinamização da investigação destas técnicas.

5.12 Considerações Finais do capítulo

Este capítulo apresentou o planejamento do estudo experimental proposto neste trabalho. Este planejamento utiliza as etapas sugeridas por Wohlin et al. [Wohlin et al. 2000] onde são

apresentados os sujeitos, o contexto, o objeto, as ameaças à validade do experimento, dentre outros elementos. A fundamentação a respeito deste processo é apresentada no capítulo de Fundamentação Teórica deste trabalho (Capítulo 2, Seção 2.5).

Outras informações, como a investigação necessária para estabelecer quantos e quais são os níveis dos projetos experimentais na Seção 5.10, são encontradas nos Apêndices deste documento. Esta análise foi apresentada nos Apêndices pois constitui uma investigação estatística necessária para estruturar o projeto experimental. O objetivo disto é manter o foco da leitura deste capítulo na metodologia realizada.

Com relação às ameaças apresentadas na Seção 5.11, é importante lembrar que apesar de identificar as ameaças na etapa de planejamento, novas ameaças podem ser encontradas durante a etapa operacional ou de análise do experimento. Além disto, é possível que alguma medida estabelecida para lidar com estas ameaças não se mostre suficiente. Diante disto, no Capítulo 7 é apresentada uma Seção que discute as ameaças encontradas durante a execução do experimento, e como estas afetam o estudo experimental realizado.

Capítulo 6

Instrumentação

Neste capítulo será apresentada a instrumentação realizada para a execução do estudo experimental. Durante este capítulo, as técnicas são referenciadas a partir dos identificadores apresentados durante o planejamento. São eles:

- T_1**:** Seleção baseada em Análise de Dependência em Máquinas de Estados Finitas Estendidas [Chen et al. 2007];
- T_2**:** Seleção baseada em Análise de Riscos e Diagramas de Atividade em UML [Chen et al. 2002];
  - T_{2i}: Técnica T_2 configurada por um testador inexperiente;
  - T_{2e}: Técnica T_2 configurada por um testador experiente;
- T_3**:** *Weighted Similarity Approach for Regression Testing* (WSA-RT);
  - T_{3i}: WSA-RT configurada por um testador inexperiente;
  - T_{3e}: WSA-RT configurada por um testador experiente;
- T_4**:** Técnica baseada em *clustering* de Laski e Szermer [Laski and Szermer 1992];
- T_5**:** Seleção aleatória de casos de teste.
  - $T_{5-25\%}$: Seleção aleatória de 25% dos casos de teste de regressão.
  - $T_{5-50\%}$: Seleção aleatória de 50% dos casos de teste de regressão.

– $T_{5-75\%}$: Seleção aleatória de 75% dos casos de teste de regressão.

A instrumentação contempla as ferramentas utilizadas, a criação dos modelos, aspectos da implementação, as modificações e o modelo de faltas utilizados no estudo experimental. Cada um destes elementos serão discutidos nas seções a seguir.

6.1 Ferramentas

Para realizar um estudo experimental diversas ferramentas são utilizadas. Neste trabalho, estas ferramentas são necessárias para a criação dos modelos de especificação, realização dos cálculos estatístico, construção de gráficos apresentados na etapa de análise, e suporte à implementação do ambiente de execução do estudo experimental. Neste experimento são utilizadas as ferramentas: *Labeled Transitions System - Based Testing*, *Magic Draw*, e o Minitab. Cada uma destas ferramentas são descritas nas subseções a seguir.

6.1.1 *Labeled Transitions System - Based Testing* – LTS-BT

Labeled Transitions System - Based Testing (LTS-BT), é uma ferramenta que realiza geração e seleção automática de casos de teste [Cartaxo et al. 2008]. Para o estudo experimental a especificação de LTS-BT caracteriza o objeto submetido às técnicas analisadas. A partir de um documento que possui informações dos requisitos, casos de uso e cenários de execução da ferramenta, são gerados: o diagrama de atividades, a máquina de estados, o grafo de fluxo de controle (GFC) e o sistema de transições rotulada (STR). Este documento de especificação de LTS-BT segue o formato proposto por Cabral e Sampaio [Cabral and Sampaio 2008].

Além de ser utilizada como objeto do estudo experimental, LTS-BT é o ambiente no qual as técnicas e o estudo experimental (coleta de dados, execução das técnicas, etc.) são implementados. Por ser uma ferramenta que implementa algumas técnicas de geração e seleção de casos de teste em TBM, LTS-BT apresenta uma arquitetura na qual a implementação deste trabalho é facilmente adicionada. Além da implementação das técnicas, também foi implementado, em LTS-BT, o ambiente que executa o experimento, ou seja, o ambiente que configura as técnicas, realiza as replicações do experimento e armazena os dados em arquivos. Os detalhes desta implementação são apresentados na Seção 6.3 deste capítulo.

6.1.2 *Magic Draw*

O *Magic Draw*[1] é uma ferramenta utilizada para modelagem de diagramas em *Unified Modeling Language*[2] (UML). Uma vez que as técnicas analisadas neste experimento são baseadas em especificação, algumas delas necessitam de uma especificação em UML, como um diagrama de atividades ou um diagrama de máquinas de estados. Diante disto, a especificação utilizada como objeto do estudo experimental é modelada utilizando o *Magic Draw* na versão 15.1.

A ferramenta foi escolhida pois a criação e a visualização dos modelos é facilitada pela interface gráfica. Além disto, ela possui uma baixa curva de aprendizado e exporta os modelos no formato *XML Metadata Interchange*[3] (XMI), que é utilizado por LTS-BT para realizar a leitura dos diagramas. Apesar da facilidade na sua utilização, o *Magic Draw* necessita de muitos recursos de *hardware* para executar, como memória física e aleatória, e processamento. Portanto, é uma ferramenta de execução um pouco mais lenta que outras que necessitam de menos recursos computacionais.

6.1.3 Minitab

Para a realização da análise estatística é utilizada a ferramenta Minitab[4]. Através do Minitab é possível realizar testes estatísticos como, testes de normalidade (e.g Anderson-Darling, Kolmogorov-Smirnoff), testes não-paramétricos (por exemplo os testes de Mann-Whitney, Kruskal-Wallis), e testes paramétricos (como ANOVA, teste t de Student). Os dados obtidos durante a execução do experimento são submetidos ao Minitab, para a realização do teste estatístico adequado.

O Minitab foi escolhido pois possui uma interface gráfica que facilita a realização da análise estatística e organização dos dados. Cada teste estatístico possui uma formatação de saída específica. A documentação de ajuda da ferramenta possui as instruções adequadas para a correta compreensão da saída de cada teste, bem como um exemplo que mostra todo o passo a passo, desde a formatação das entradas.

[1] http://www.magicdraw.com/

[2] http://www.uml.org/

[3] http://www.omg.org/spec/XMI/

[4] http://www.minitab.com/

Todos os resultados dos testes estatísticos são exibidos em função do p. Diferentemente de outras ferramentas, nos casos em que o p é menor que 1, o Minitab apresenta quantas casas decimais nulas há antes do primeiro dígito significativo. Por exemplo, quando a saída de um teste é $p = 0,0000$ deve ser interpretada como $p = 0,00001$. Nos resultados exibidos ao longo deste trabalho, o valor de p reportado já foi alterado para a notação convencional.

6.2 Modelos de Entrada

As técnicas de re-teste seletivo baseado em especificação executam a partir de informações obtidas na especificação do software. Diante disto, é necessário fornecer, para as técnicas analisadas, o modelo de especificação utilizado no estudo experimental, ou seja, o modelo de especificação de LTS-BT. Uma vez que cada técnica usa um formato específico de modelo para executar, a especificação de LTS-BT é modelada nos diversos formatos necessários para a execução das técnicas analisadas.

A partir de um documento de requisitos, a especificação de LTS-BT é modelada em um diagrama de atividades, um diagrama de máquina de estados, um gráfico de fluxo de controle e um sistema de transições rotuladas. Não é utilizado nenhum mecanismo formal para garantir a semântica entre estes diagramas, no entanto, para cada diagrama modelado, é verificado se todas as funcionalidades e cenários da especificação são contemplados em todos os formatos modelados.

As técnicas T_2 e T_4 comparam as especificações da versão base e versão delta, procurando identificar as modificações realizadas (transições removidas, estados adicionados, dentre outros). Diante disto, é necessário construir os modelos das versões base e delta para o diagrama de atividades e para o Grafo de Fluxo de Controle (GFC). O GFC é construído manualmente a partir do diagrama de atividades seguindo um processo proposto por Chen et al. [Chen et al. 2002], onde as atividades e transições destes diagramas são mapeados em estados e arestas do GFC, respectivamente.

O Sistema de Transições Rotuladas (STR) da especificação é utilizado apenas para a geração automática da suíte de testes de regressão a ser reduzida, pois as técnicas analisadas que utilizam este modelo (T_3 e T_5) realizam operações apenas na suíte de testes de regressão, e não no modelo em si. Portanto, são modelados os STR das versões base e delta com

o objetivo de gerar automaticamente a suíte de testes. Para T_3, o STR da versão delta é utilizado também para a configuração da técnica pelo testador, ou seja, a especificação do perfil de uso. A partir do perfil de uso no STR, os casos de teste são gerados contendo os valores de probabilidade.

O diagrama de máquina de estados, utilizado pela técnica T_1 é modelado a partir dos requisitos e cenários de LTS-BT. É modelada, apenas, a máquina de estados da versão base do objeto, pois a própria técnica T_1 realiza as modificações no diagrama a partir de um arquivo que possui as modificações [Bo 2005]. A partir das máquinas de estados das versões base e delta, a técnica constrói um gráfico estático de dependência, relacionando as dependências entre as transições e os estados do modelo.

A máquina de estados da versão delta, gerada pela técnica T_1, é utilizada para gerar os casos de teste automaticamente. No entanto, como esta suíte gerada apresenta muita redundância, ou seja, muitos casos de teste cobrem as mesmas transições, esta suíte de teste de regressão de T_1 apresenta 8079 casos de teste. Uma vez que não é viável comparar a seleção nesta suíte de regressão com as das demais técnicas, onde cada uma apresenta 65 casos de teste, é necessário aplicar uma seleção na suíte de regressão de T_1 para reduzir esta redundância.

Através da técnica de seleção de casos de teste baseada em similaridade de caminhos [Cartaxo et al. 2009], foi possível reduzir a suíte de de regressão de T_1 para 65 casos de teste, alcançando o mesmo tamanho das demais suítes. Mesmo com esta redução, foi possível manter a cobertura de todos os estados e transições da máquina de estados. Esta manutenção na suíte de T_1 não prejudica o desempenho da técnica, pois os próprios autores argumentam que o desempenho de T_1 é independente da suíte de regressão utilizada como entrada. Os autores argumentam que o desempenho da técnica depende apenas dos modelos e modificações.

Todos os modelos construídos contemplam todos os requisitos e cenários especificados para o objeto do experimento, a ferramenta LTS-BT. Os modelos do diagrama de atividades e do diagrama de máquina de estados foram exportados no formato XMI, para serem processados pelas técnicas. O GFC foi convertido a partir do diagrama de atividades, enquanto que o STR foi construído no formato *Trivial Graph Format* (TGF). Todos os conversores de formatos de arquivos XMI e TGF foram implementados para possibilitar a leitura dos diagramas durante a geração automática das suítes de teste de regressão.

6.3 Implementação

Para executar este estudo experimental, é necessário implementar diversos elementos como: as técnicas, a leitura dos modelos, o injetor de faltas configurado a partir de um modelo de faltas, a coleta de dados, dentre outros. Um dos primeiros aspectos investigados durante a definição deste estudo experimental foi a utilização de implementações disponíveis das técnicas analisadas. Porém, como não foi encontrada uma única ferramenta que apresentasse a implementação de todas as técnicas analisadas decidimos implementá-las.

Utilizar implementações, de cada técnica, realizadas por fontes diferentes pode caracterizar um viés no experimento, pois implementações diferentes podem apresentar diferentes resultados de eficiência, ou erros de medições. Portanto, definimos que a implementação de todas as técnicas seria realizada pelo investigador, para manter a precisão no controle e na rastreabilidade da execução de cada técnica. Diante disto, a implementação do estudo experimental é realizada na ferramenta LTS-BT.

Nesta seção é apresentada, inicialmente, uma visão arquitetural de LTS-BT, e como os aspectos do estudo experimental são incorporados nesta arquitetura. Em seguida, são discutidos os elementos da implementação, (e.g. ambiente de desenvolvimento, classes, *design patterns* utilizados, e pacotes), e por fim, é apresentada como foi realizada a Verificação e Validação da implementação.

6.3.1 Arquitetura de LTS-BT

Uma vez que o estudo experimental é implementado em LTS-BT, os elementos arquiteturais da ferramenta devem ser observados durante a etapa de instrumentação. Um dos objetivos deste trabalho é adicionar à ferramenta os elementos do estudo experimental (e.g. técnicas, conversores de arquivos, coleta de dados, dentre outros), permitindo que outros estudos experimentais sejam realizados em LTS-BT. Portanto, os novos elementos de implementação devem ser incorporados em uma nova arquitetura da ferramenta. Um esquema arquitetural de LTS-BT, antes da adição das novas funcionalidades, pode ser visto na Figura 6.1.

Nesta figura, os retângulos representam módulos da ferramenta. Um módulo, em LTS-BT representa um conjunto de classes e/ou pacotes responsáveis por uma determinada função da ferramenta, e.g. geração de casos de teste, leitura de modelos, escrita em arquivos, dentre

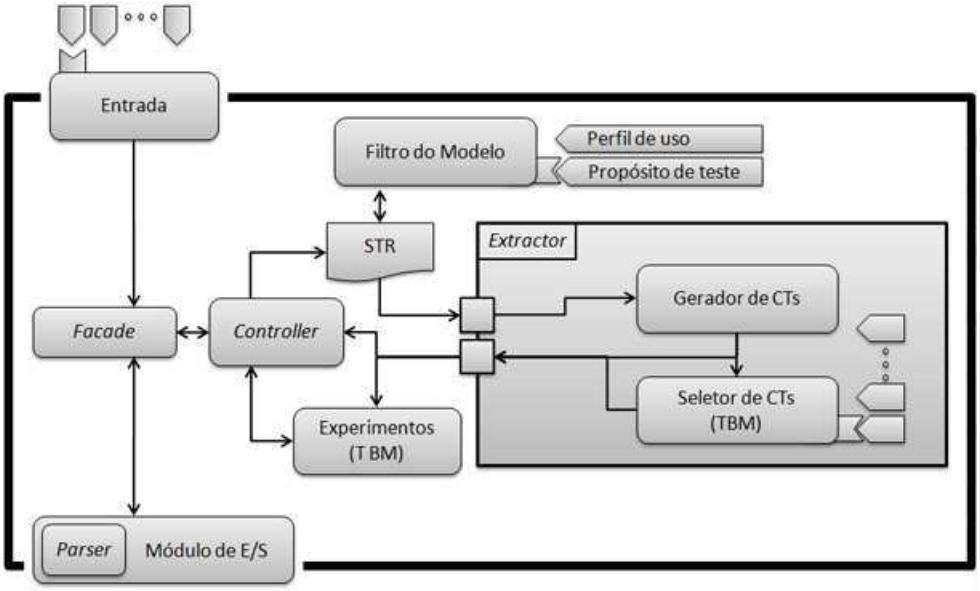

Figura 6.1: Arquitetura de LTS-BT antes da implementação do estudo experimental.

outros. Os módulos se relacionam entre si, através das setas no esquema. Esta relação representa uma dependência entre os módulos da ferramenta.

A partir da arquitetura ilustrada na Figura 6.1 é possível observar os módulos de LTS-BT responsáveis pelas funcionalidades das ferramentas (e.g. o módulo para seleção de casos de teste a partir de uma técnica especificada). O módulo "Experimentos (TBM)" da Figura 6.1 é responsável por executar experimentos em técnicas de Seleção de Casos de Teste [Oliveira Neto and Machado 2008] para Teste Baseado em Modelos. Apesar de realizar experimentos, este módulo não é utilizado para o estudo experimental deste trabalho, pois lida com aspectos diferentes da teoria de teste de software.

Para a instrumentação, elaboramos um esquema arquitetural adicionado à LTS-BT. Este esquema arquitetural, ilustrado na Figura 6.2, é integrado com a arquitetura da Figura 6.1. Na nova arquitetura, os módulos já existentes em LTS-BT que sofreram alguma modificação estão em vermelho. Os elementos do estudo experimental são executados a partir de um controlador (Controlador do Experimento) que se comunica com o controlador da ferramenta (*Controller*).

Os elementos necessário para a configuração do experimento são ilustrados na Figura 6.2. São eles: o objeto do experimento (O), o modelo de faltas (M_f), e a quantidade de replicações desejada (r). Quaisquer aspectos diferentes destes são incorporados na imple-

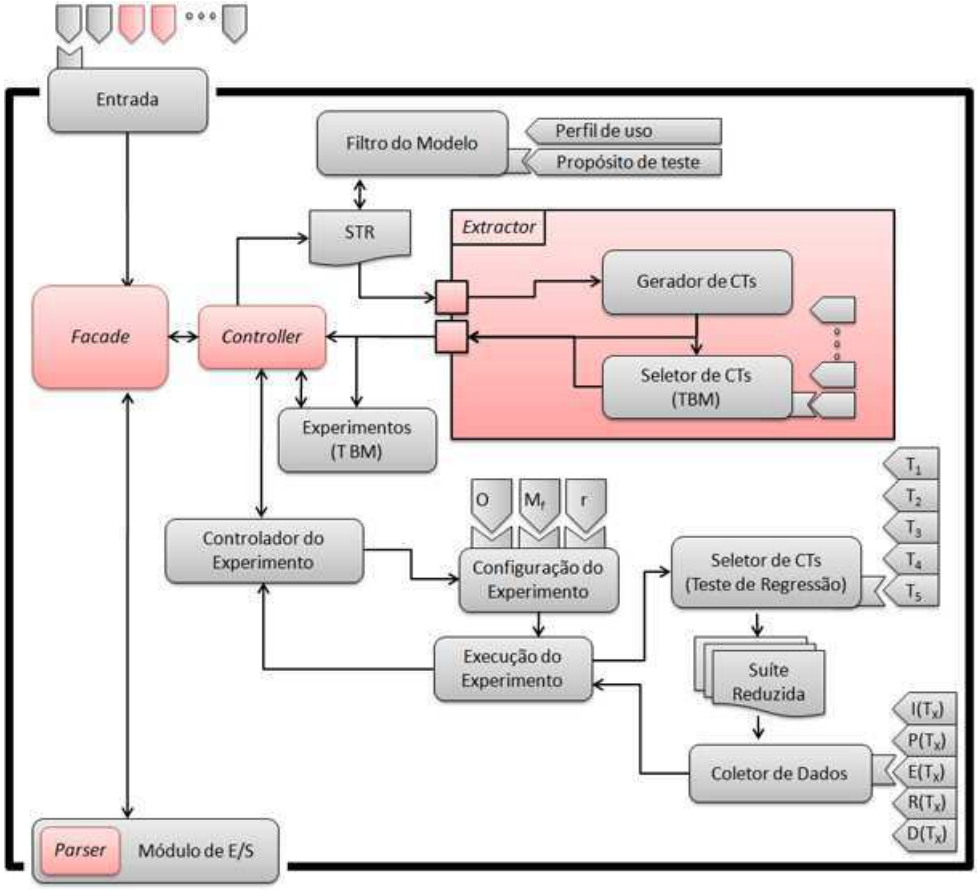

Figura 6.2: Arquitetura de LTS-BT após a implementação do estudo experimental.

mentação específica da técnica, os denominados parâmetros de configuração das técnicas (e.g. valores de custos de casos de teste, requisitos de teste, dentre outros). Ainda sob esta perspectiva, o investigador deve definir quais técnicas são analisadas. Após definidas, estas técnicas são executadas no estudo experimental. Outro aspecto considerado nesta arquitetura é a coleta de dados.

Para este estudo experimental, os dados coletados são referentes às variáveis dependentes analisadas (inclusão, precisão, eficiência, potencial de redução e densidade de faltas). Uma vez que a densidade de faltas, a inclusão e a precisão estão relacionados com as faltas de regressão na suíte reduzida, é necessário fornecer ao experimento um modelo de faltas. Este modelo pode ser especificado pelo investigador, a partir de seu conhecimento acerca do(s) objeto(s), ou a partir de um histórico real de execução e relatório de faltas no teste do software [Binder 1999]. À medida que novos dados a serem observados são encontrados na literatura ou em outros estudos experimentais, a arquitetura permite que estes novos aspectos sejam

incorporados na coleta de dados do estudo experimental.

Este esquema arquitetural é adicionado à arquitetura de LTS-BT, com o objetivo de fornecer à ferramenta, um ambiente para realização de estudos experimentais com técnicas de re-teste seletivo baseado em especificação. Além disto, as técnicas podem ser incorporadas ao processo de LTS-BT de seleção automática de casos de teste. Portanto, este trabalho também contribui para o suporte ferramental às técnicas de re-teste seletivo baseado em especificação.

6.3.2 Estrutura do Código

Ao implementar o estudo experimental, foram observados aspectos da estrutura do código (pacotes e classes) que facilitassem o entendimento e manutenção do código para outros investigadores que desejassem utilizar a implementação para reproduzir ou melhorar o experimento. Dessa forma, o código-fonte está escrito em inglês. O primeiro passo da implementação é a definição de pacotes para estabelecer um escopo com relação ao objetivo de cada classe. Dessa forma as entidades do código estão organizadas nos seguintes pacotes:

- **`parsers:`** Este pacote contém as classes responsáveis pela leitura e conversão entre os formatos de arquivos que possuem os modelos de entrada (e.g. XMI, TGF, AUT) e as modificações realizadas na ferramenta.

- **`techniques:`** Este pacote contém classes e pacotes que, por sua vez, possuem a implementação das técnicas de re-teste seletivo analisadas neste estudo experimental:

  - **`dependence_analysis:`** As entidades responsáveis pela execução da técnica de análise de dependência em máquinas de estados estão contidas neste pacote. Dentre estas entidades, podemos destacar as classes que realizam a análise de dependência e a identificação dos padrões de interação.

  - **`risk_analysis:`** A técnica de seleção baseada em análise de riscos é implementada pelas classes deste pacote. Estas classes são responsáveis por identificar as modificações, e realizar a análise de risco baseado no valor de exposição de risco calculado.

  - **`wsart:`** A implementação da técnica WSA-RT está presente neste pacote. As classes deste pacote realizam a construção da matriz de similaridade, assim como, o processo de seleção de casos de teste baseada na análise do perfil de uso fornecido pelo testador.
  - **`cluster_tech:`** As classes da técnica de seleção baseada em *clusters* estão neste pacote. Dentre estas classes, podemos destacar as responsáveis pela construção dos *clusters* e identificação das modificações estão presentes neste pacote.
  - **`random_tech:`** Este pacote contém apenas uma classe, que implementa a técnica de seleção aleatória de casos de teste. Esta classe é responsável por selecionar, aleatoriamente, uma quantidade de casos de teste especificada pelo testador.

- **`models:`** Este pacote contém as classes responsáveis pela representação dos modelos de especificação utilizados no experimento.
  - **`state_machine:`** Este pacote contém as classes responsáveis pela representação da máquina de estados, como estados e transições.
  - **`sdg:`** A técnica de seleção baseada em análise de riscos gera um diagrama de dependência a partir da análise de dependências realizada na máquina de estados. Os elementos deste diagrama estão contidos nas classes deste pacote.
  - **`activity_diagram:`** Os elementos do diagrama de atividades (e.g. atividades e transições) estão representados nas classes deste pacote.
  - **`cfg:`** Os elementos do grafo de fluxo de controle, como os vértices, transições e *clusters* são representados pelas classes deste pacote.
  - **`lts:`** Este pacote contém as classes que representam os elementos de um STR, como os estados e transições (com ou sem os valores de probabilidade, especificados pelo testador).

- **`experiment:`** Este pacote possui apenas uma classe responsável por executar e configurar o estudo experimental. A implementação de estudos experimentais desenvolvidos com outras configurações (e.g. mais objetos, outros modelos de faltas, dentre outros) devem ser incluídas neste pacote.

- **`util`:** Este pacote contém classes com diversas funções que facilitam o uso e processamento das demais classes. Alguns exemplos são: as classes para o processamento de cadeias de caracteres (*strings*), escrita de arquivos de saída, e formatação de dados (decimais, porcentagem, dentre outros).

A disposição das principais entidades utilizadas na execução do experimento são apresentados na Figura 6.3. A classe `ExperimentalStudy` é responsável por configurar e executar o experimento. Além disto, esta classe armazena os dados em arquivos para a análise dos resultados. Os parâmetros do estudo experimental são fornecidos pelo investigador para a classe. Exemplos destes parâmetros são: a quantidade de execuções, as técnicas a serem executadas, os dados a serem capturados, e quais os modelos de entrada utilizados.

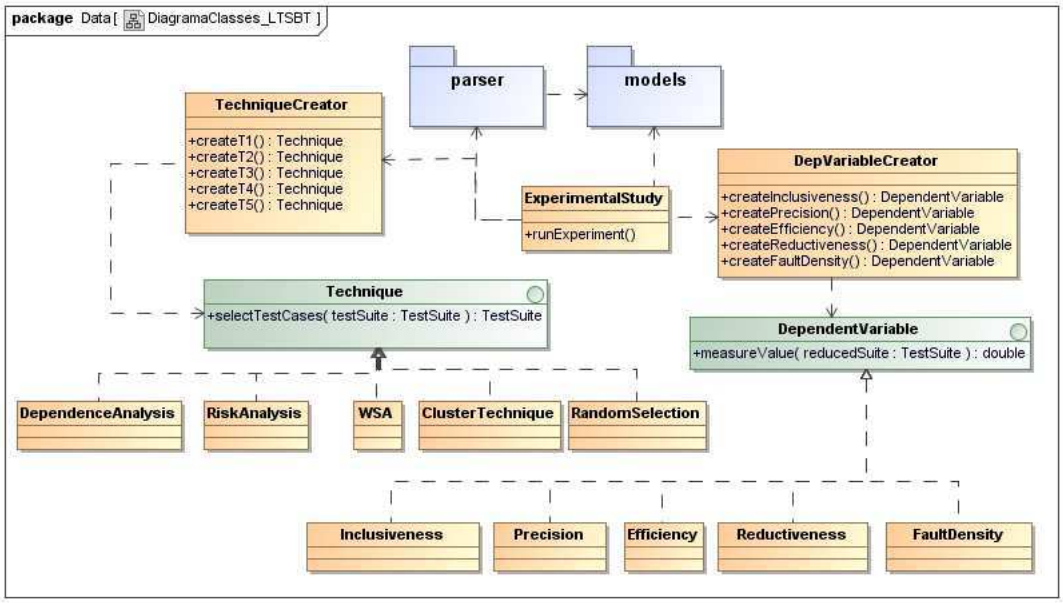

Figura 6.3: Diagrama com as principais classes da implementação do estudo experimental em LTS-BT.

Como podemos observar na Figura 6.3, é utilizado o *design pattern Factory Method* nas classes `TechniqueCreator` e `DepVariableCreator`. No início da execução do experimento, o investigador especifica como parâmetro da classe `ExperimentalStudy`, através de palavras reservadas, as técnicas e variáveis dependentes a serem analisadas. A classe passa então estas palavras reservadas para os respectivos *creators*, para obter as técnicas a serem executadas e as variáveis dependentes a serem observadas.

Essa estrutura no código do estudo experimental é adotada com o objetivo de facilitar a adição de novas técnicas e variáveis dependentes. Além disto, foram definidas interfaces para caracterizar o comportamento destas entidades (interfaces `Technique` e `DependentVariable` da Figura 6.3). Dessa forma, durante a etapa de execução, o estudo experimental executa as técnicas e coleta os dados de acordo com as implementações disponíveis.

É importante lembrar que o objetivo destas técnicas é fornecer uma subconjunto da suíte de testes a partir da suíte de regressão especificada. Portanto, novas técnicas podem ser adicionadas no estudo experimental, desde que a classe desta nova técnica implemente a interface `Technique` do pacote `techniques`. Além disto, é necessário que o `TechniqueCreator` seja atualizado para possibilitar a obtenção da nova técnica implementada. Estas considerações também se aplicam à adição de novas variáveis dependentes, através das entidades `DependentVariable` e `DepVariableCreator`.

A classe `ExperimentalStudy` utiliza os conversores de formatos de arquivos (pacote `parser`) para obter os modelos de entrada específicos para cada técnica. Os conversores de arquivos fornecem à classe *ExperimentalStudy* os respectivos formatos de modelos requisitados. Estes modelos são utilizados pelas demais entidades da implementação (e.g. injetor de faltas, construtores das técnicas, dentre outros) para estruturar a execução de cada técnica. Após finalizada a configuração dos modelos e técnicas, é iniciado um processo iterativo, no qual, a cada iteração, cada uma das técnicas analisadas são executadas, e os dados referentes às variáveis dependentes são armazenados.

A implementação deste trabalho é realizada na linguagem de programação Java[5], utilizando a IDE (*Integrated Development Environment*) Eclipse[6]. Além disto, é importante lembrar que alguns módulos da ferramenta LTS-BT já existentes antes da realização deste trabalho são utilizados. Dentre eles o módulo de geração automática de casos de teste em STR e as classes de representação da suíte de teste e casos de testes (módulo *Extractor* da Figura 6.2). Estas entidades de LTS-BT são aproveitadas, através da herança entre classes, para a implementação das classes que representam a suíte de testes de regressão e os casos de teste de regressão, respectivamente.

[5]http://www.java.com/en/

[6]http://www.eclipse.org/

6.3.3 Verificação e Validação

Durante e após o desenvolvimento de uma implementação, é necessário checar se esta contempla os aspectos especificados e provê as funcionalidades esperadas pelos clientes. O processo que checa estes aspectos é denominado Verificação e Validação. O objetivo da Verificação é checar se o produto está sendo construído corretamente, enquanto que a Validação procura checar se o produto certo está sendo construído [Sommerville 2010]. Para a implementação deste estudo experimental, a Verificação é realizada através de testes de unidade no código desenvolvido, enquanto que a Validação é realizada através de estudos de casos para as técnicas de re-teste seletivo analisadas.

Para realizar a verificação é utilizado o *framework* JUnit[7] para a execução automática e o desenvolvimento dos testes de unidade. Dessa forma, são desenvolvidas classes de teste para cada classe criada na implementação. Foi coberto, pelos teste, cerca de 90% do código desenvolvido. O percentual de código não coberto pelos testes caracterizam operações de entrada e saída e manipulação de caracteres, como formatação de cadeias de caracteres (palavras reservadas, números reais e porcentagens).

A validação das técnicas é realizada através da execução de estudos de casos. Dessa forma, utilizamos os estudos de caso propostos pelos autores das técnicas. Além disto, outros exemplos "*toy*" foram desenvolvidos pelo investigador do experimento, com o objetivo de validar aspectos do algoritmo (e.g. modificações em laços e fluxos específicos do modelo) que não são abordados no estudo de caso dos autores da respectiva técnica.

Os recursos de verificação e validação, fazem parte da validade do experimento, pois é necessário que a implementação e execução das técnicas estejam certas. Os elementos de ameaça à validade relacionados com a verificação e validação da implementação são discutidos na Seção 5.11 do Capítulo 5.

6.4 Modificações

Ao realizar um Teste de Regressão, um dos principais aspectos a serem observados são as modificações realizadas no sistema. No contexto de especificação, estas modificações podem ser diversas, como a adição de funcionalidades, refinamentos arquiteturais, mudança

[7] http://www.junit.org/

nos cenários da aplicação, dentre outras. Para investigar as técnicas, é necessário estabelecer modificações no objeto do experimento para possibilitar a caracterização das técnicas de acordo com a natureza das modificações realizadas no sistema. Portanto, foram contemplados os 3 tipos de modificações usualmente considerados no teste de regressão baseado em especificação [Korel et al. 2002, Chen et al. 2007]: a adição, remoção e modificação de funcionalidades.

As modificações estabelecidas para este estudo experimental não estão implementadas no objeto do estudo experimental, pois é necessário muito tempo para realizar estas modificações, além de esforço de implementação e testes no objeto modificado (versão delta da ferramenta). Estas modificações foram estabelecidas a partir de modificações realizadas em outras versões da ferramenta, e a partir de cenários que seriam importantes observar nas técnicas, para identificar que tipos de modificações as técnicas seriam capazes de analisar. Para este estudo experimental, planejamos as seguintes modificações em LTS-BT.

- Adicionar a técnica WSA de Bertolino et al. no processo de geração com propósito de teste.
- Adicionar a técnica WSA-RT (Capítulo 4) no processo de seleção automática de casos de teste de LTS-BT.
- Remover a geração de casos de teste baseada no propósito de teste de rejeição.
- Remover a geração de casos de teste no formato CSP.
- Modificar um rótulo nos menus da interface gráfica da ferramenta.
- Realizar o tratamento de arquivos inválidos em um único módulo do sistema.

Portanto, foram realizadas: 2 adições de funcionalidades, 2 remoções de funcionalidades, 1 modificação simples que não envolve aspectos estruturais no modelo (i.e. posição de transições e vértices) e 1 modificação complexa que envolve a re-estruturação de elementos do modelo. No total, foram realizadas 6 modificações. Estas modificações são, então, utilizadas para caracterizar o comportamento e desempenho da técnica, durante a execução do experimento.

6.5 Modelo de Faltas

Um dos principais aspectos analisados no teste de regressão é a quantidade de casos de teste que capturam as faltas de regressão. Este estudo experimental analisa esta característica através das variáveis dependentes de inclusão, precisão e densidade de faltas. Para medir estas faltas, é necessário obter as faltas de regressão do software. Geralmente, estas faltas são obtidas através de histórico de execuções passadas do software relacionando-as com as modificações realizadas.

Em algumas situações este histórico não está disponível, ou não há registro das faltas encontradas durante o desenvolvimento do software. Para analisar as faltas de um software quando estas não são conhecidas, é possível utilizar um modelo de faltas [Binder 1999]. As variáveis dependentes de inclusão, precisão e densidade de faltas são analisadas baseadas nas faltas de regressão capturadas pelos casos de teste.

Estabelecemos estas faltas de regressão, neste trabalho, a partir de algumas faltas encontradas no histórico de modificações de LTS-BT [Oliveira Neto and Machado 2008]. No entanto, uma vez que estas faltas não cobrem os cenários desejados para prover uma boa avaliação do comportamento e desempenho das técnicas, são utilizadas também faltas estabelecidas a partir de um modelo de faltas criado pelo investigador do experimento.

O modelo de faltas é utilizado para estabelecer passos no modelo que poderiam apresentar faltas devido às modificações. Selecionamos estes passos com faltas a partir de trechos no modelo que deveriam ser selecionados pelas técnicas, pois são trechos que integram muitos fluxos da aplicação, ou caracterizam funcionalidades críticas para o funcionamento adequado da ferramenta. Portanto, o modelo de faltas utilizado é composto pelas seguintes faltas de regressão:

1. A especificação do perfil de uso após a definição do propósito de teste não insere os valores de probabilidade no modelo.

2. Alguns valores do perfil de uso são removidos ao aplicar o propósito de teste.

3. Falta por efeito colateral no módulo de experimentos com seleção de casos de teste em TBM, devido à adição da técnica WSA.

4. A tela que exibe as técnicas de seleção em TBM está exibindo também a técnica WSA-RT, que é para teste de regressão.

5. A seleção de propósito a partir dos rótulos das transições não funciona após a remoção do propósito de rejeição.

6. Casos de teste gerados por arquivos TGF são apresentados no formato CSP.

7. O rótulo da interface gráfica não foi devidamente alterado, interrompendo a comunicação do botão com a lógica de negócios.

8. A ferramenta identifica como inválido um diagrama de atividade sem nós de decisão.

9. A ferramenta identifica como válido um diagrama de máquina de estados sem estados iniciais.

10. O tratamento de arquivos XMI inválidos apresenta uma falta após integrado com os dos demais foramtos (TGF, AUT, etc.).

As faltas 1, 2 e 3 deste modelo são referentes ao histórico de modificações da ferramenta, portanto, representam faltas reais. As demais faltas ($4, 5, 6, 7, 8, 9$ e 10) foram elaboradas pelo investigador, com o objetivo de aumentar a quantidade e a variedade de faltas analisadas no experimento. Cada uma destas faltas estão relacionadas com as modificações especificadas na Seção 6.4, e representam um passo da aplicação no modelo (i.e. uma transição ou estado, de acordo com o modelo analisado). Os casos de teste que exercitam uma ou mais destas transições são os casos de teste que capturam as faltas de regressão. Neste estudo experimental, 28 dos 65 casos de teste da versão delta capturam estas 10 faltas de regressão.

6.6 Considerações Finais do Capítulo

Este capítulo abordou os aspectos de instrumentação do estudo experimental. Dentre os elementos de instrumentação, foram discutidos os modelos, as ferramentas utilizadas no experimento, assim como a implementação realizada para a execução do estudo experimental, as modificações analisadas e o modelo de faltas definido para identificar as faltas de regressão.

A implementação pode ser utilizada para outros estudos experimentais, desde que os pesquisadores adicionem as implementações de suas respectivas técnicas. Portanto, os elementos de instrumentação caracterizam uma das contribuições deste trabalho. A utilização de *design patterns* no código proporcionou maior flexibilidade e modularização do código [Gamma et al. 1994], facilitando a adição de novas técnicas de re-teste seletivo baseado em especificação para comparar com as já implementadas neste estudo.

A etapa de instrumentação do estudo experimental necessitou de muito esforço para ser concluída devido à complexidade de implementação de alguns trechos dos algoritmos das técnicas. Esta complexidade está relacionada com o caminhamento e operações realizadas nos modelos para identificar e analisar as modificações e dependências.

Após finalizar a etapa de instrumentação, o estudo experimental foi executado. A execução foi realizada em uma máquina com processador Intel®Core^{TM} 2 Duo de $2,13$ GHz, 1 GB de memória de acesso aleatório, e 140 GB de memória secundária (disco rígido). Durante a etapa de execução, os dados das suítes de testes obtidas foram coletados e analisados através de recursos estatísticos. Os detalhes destes resultados e da análise realizada são apresentados no capítulo a seguir.

Capítulo 7

Resultados e Análise do Experimento

Neste capítulo são apresentados e discutidos os resultados obtidos para os projetos experimentais especificados na Seção 5.10. Inicialmente, são apresentados os passos seguidos para a realização do teste de hipótese em todos os projetos experimentais especificados. Após descrever estes passos, são apresentados os resultados obtidos a partir da execução das técnicas, para cada projeto experimental, assim como as respectivas conclusões a respeito das variáveis dependentes e das técnicas analisadas.

7.1 Investigação das Hipóteses

No planejamento realizado neste estudo experimental, foram estabelecidos 5 projetos experimentais, a partir dos quais, foram estruturados 5 testes de hipóteses, que verificam se as técnicas possuem um comportamento semelhante para cada variável dependente analisada. Diante disto, para cada projeto experimental, seguimos os seguintes passos, para realizar os testes de hipóteses:

1. **Investigação das premissas da Análise de Variância (ANOVA):** A partir dos dados obtidos, são realizados testes visuais com os resíduos dos dados observados, com o objetivo de verificar se o teste de ANOVA pode ser aplicado.

2. **Aplicação do teste de Kruskal-Wallis:** Após verificar que as premissas de ANOVA não são respeitadas pelos dados, é decidido aplicar o teste estatístico de Kruskal-Wallis, que é o teste não-paramétrico indicado quando o teste de ANOVA não pode

ser realizado. Para esta etapa, foi utilizado o software Minitab.

3. **Investigar o p:** Como resultado do teste de Kruskal-Wallis, é obtido um p, que deve então ser comparado com o nível de significância estabelecido para o teste de hipótese, permitindo rejeitar, ou não, a hipótese nula em favor da hipótese alternativa.

4. **Analisar resultados:** A partir do resultado do teste de hipóteses, os aspectos que caracterizem as diferenças ou semelhanças no comportamento das técnicas são, então, verificados e as conclusões são apresentadas.

A partir dos passos descritos acima, são realizadas as 5 análises para cada projeto experimental estruturado. Para todas as análises realizadas, as premissas de ANOVA não são respeitadas pelos dados. Portanto, é necessário realizar o teste de Kruskal-Wallis em todos os testes de hipóteses. Para a realização deste teste, é utilizada a ferramenta Minitab. Os resultados destes testes, para cada projeto experimental, é apresentado na Tabela 7.1.

7.2 Projeto Experimental 1 - Inclusão

A partir dos dados do Projeto Experimental 1 (inclusão das técnicas), são executados os passos de análise descritos na Seção 7.1. Os resultados de cada passo são descritos a seguir:

1. O teste das premissas de ANOVA foi realizado no conjunto de dados obtidos e foi verificado (Apêndice C.1) que não é adequado utilizar ANOVA neste projeto experimental. Portanto, foi utilizado o teste de Kruskal-Wallis.

2. Os resultados do teste de Kruskal-Wallis são apresentados na Tabela 7.1 (a).

3. O p obtido como resultado do teste foi $0,0001$. Uma vez que o p é menor que o nível de significância considerado no teste ($\alpha = 0,05$), podemos rejeitar a hipótese nula em favor da hipótese alternativa.

4. A partir do teste de hipótese, podemos afirmar, com um nível de confiança de 95%, que as técnicas se comportam de forma diferente com relação à propriedade de inclusão.

Tabela 7.1: Resultados dos testes de Kruskal-Wallis para cada variável dependente.

Resultados Kruskal-Wallis

Inclusão (a)

Técnicas	Mediana	Posto Médio	Z
T_1	50,00	404,5	0,19
T_2	92,86	750,5	16,19
T_{3e}	35,71	224,8	-8,13
T_{3i}	33,93	174,7	-10,45
T_4	78,57	629,0	10,57
$T_{5-25\%}$	25,00	57,0	-15,89
$T_{5-50\%}$	50,00	391,9	-0,40
$T_{5-75\%}$	75,00	571,6	7,92
H = 756,41	gl = 7	$p = 0,0001$	
H = 764,42	gl = 7	$p = 0,0001$	(empates)

Precisão (b)

Técnicas	Mediana	Posto Médio	Z
T_1	27,78	228,4	-7,96
T_2	85,14	749,5	16,15
T_{3e}	25,68	127,5	-12,63
T_{3i}	29,73	302,1	-4,55
T_4	35,14	444,5	2,04
$T_{5-25\%}$	24,32	150,1	-11,59
$T_{5-50\%}$	48,65	550,5	6,94
$T_{5-75\%}$	72,97	651,5	11,61
H = 722,47	gl = 7	$p = 0,0001$	
H = 726,29	gl = 7	$p = 0,0001$	(empates)

Eficiência (ns) (c)

Técnicas	Mediana	Posto Médio	Z
T_1	$1,51 \cdot 10^{11}$	850,5	16,32
T_{2e}	24641188	652,5	8,24
T_{2i}	34933107	747,0	12,10
T_{3e}	14430249	491,3	1,67
T_{3i}	14510029	511,1	2,47
T_4	10541	272,6	-7,26
$T_{5-25\%}$	10541	272,6	-7,26
$T_{5-50\%}$	7667	162,9	-11,73
$T_{5-75\%}$	4792	93,8	-14,55
H = 839,46	gl = 8	$p = 0,0001$	
H = 839,75	gl = 8	$p = 0,0001$	(empates)

Redução (d)

Técnicas	Mediana	Posto Médio	Z
T_1	60,00	450,8	2,32
T_2	12,07	50,5	-16,19
T_{3e}	68,46	562,0	7,47
T_{3i}	70,00	638,7	11,02
T_4	46,15	250,5	-6,94
$T_{5-25\%}$	5,38	750,5	16,19
$T_{5-50\%}$	50,77	350,5	-2,31
$T_{5-75\%}$	26,15	150,5	-11,57
H = 782,52	gl = 7	$p = 0,0001$	
H = 790,69	gl = 7	$p = 0,0001$	(empates)

Densidade de Faltas (e)

Técnicas	Mediana	Posto Médio	Z
T_1	41,14	176,6	-5,72
T_2	38,24	97,5	-11,84
T_3	48,96	333,5	6,42
T_4	62,86	450,5	15,48
T_5	42,19	194,4	-4,34
H = 377,97	gl = 4	$p = 0,0001$	
H = 384,13	gl = 4	$p = 0,0001$	(empates)

Legenda: *gl*: Graus de Liberdade
H: estatística do teste
p: $p - valor$

Após observar que as técnicas são diferentes, é realizada uma análise para identificar o desempenho de inclusão de cada técnica. Detalhes desta análise podem ser encontrados no Apêndice D.1. A partir dos intervalos de confiança, e de testes de Mann-Whitney, é possível observar que as técnicas possuem o seguinte desempenho de inclusão: $T_2 > T_4 > T_{5-75\%} > T_1 = T_{5-50\%} > T_{3e} > T_{3i} > T_{5-25\%}$, onde T_2 apresentou o melhor resultado de inclusão.

7.2.1 Conclusões sobre os resultados de Inclusão

Observando os resultados de inclusão de cada técnica, assim como, os resultados dos testes estatísticos realizados nos dados, detectamos que as técnicas com melhores resultados de inclusão são T_2 (técnica beaseada em análise de riscos) e T_4 (técnica baseada em *clusters*). Diferentemente das demais técnicas, essas duas técnicas priorizam a seleção dos casos de teste que exercitam os trechos modificados do modelo.

O diferencial entre T_2 e T_4 que pode ter caracterizado o melhor desempenho de T_2, é a análise de riscos que T_2 realiza. Além de selecionar os casos de teste que exercitam as modificações, T_2 utiliza as informações dos custos e riscos dos casos de teste fornecidos pelo testador do sistema para selecionar casos de teste que não exercitam modificações e que possam revelar faltas de regressão por efeito colateral (i.e. faltas em componentes não modificados do sistema que foram afetados pelas modificações). A técnica T_1 ao contrário de T_2 e T_4 foca o processo de seleção nos casos de teste que exercitam as dependências das modificações, e portanto, pode selecionar menos casos de teste que capturam faltas, por focar nas dependências e não nas modificações em si.

Para as técnicas de seleção aleatória ($T_{5-25\%}$, $T_{5-50\%}$, $T_{5-75\%}$) é observado um desempenho esperado, onde as técnicas apresentam (em média, e após muitas replicações) uma inclusão semelhante à sua respectiva taxa de cobertura. A técnica WSA-RT configurada por testadores experientes e inexperientes (T_{3e} e T_{3i}, respectivamente) apresenta, nos dois casos, uma inclusão baixa quando comparada com as demais. Foi observado que esta baixa inclusão está relacionada com o alto potencial de redução da técnica. Além disto, a maior parte desta remoção foi observada nas etapas 2 e 3 do algoritmo de WSA-RT (ver Capítulo 4). Diante disto, mesmo com o filtro de seleção (especificado para a etapa 4 do algoritmo) confiurado para 100%, ou seja não realizar nenhuma seleção baseada no perfil de uso, esta baixa inclusão

seria observada, devido ao alto potencial de redução das etapas 2 e 3 da técnica.

Portanto, ao excluir muitos casos de teste, a inclusão de WSA-RT é reduzida significativamente. Como podemos observar, a partir do melhor desempenho de T_{3e} quando comparado com T_{3i}, um testador experiente permite a configuração da técnica para selecionar uma suíte de teste com melhor inclusão (i.e. mais casos de teste que capturam faltas de regressão).

7.3 Projeto Experimental 2 - Precisão

Os dados de precisão obtidos durante as 100 execuções do experimento são analisados de acordo com os passos descritos na Seção 7.1. Os resultados de cada passo são apresentados a seguir:

1. Verificando as premissas de ANOVA no conjunto de dados de precisão das técnicas (Apêndice C.2), concluímos que não é adequado o uso de ANOVA na análise de precisão. Dessa forma, é utilizado o teste de Kruskal-Wallis.

2. Os resultados do teste de Kruskal-Wallis são apresentados na Tabela 7.1 (b).

3. Para os dados submetidos ao teste é obtido um p de $0,0001$. Como o p é menor que o nível de significância considerado no teste ($\alpha = 0,05$), rejeitamos a hipótese nula em favor da hipótese alternativa.

4. Diante do resultado do teste de hipótese, podemos afirmar, com um nível de confiança de 95%, que as técnicas se comportam de forma diferente com relação à propriedade de precisão.

A partir dos resultados dos testes de Kruskal-Wallis para a precisão, são realizados testes visuais e testes de Mann-Whitney entre os pares de técnicas, para determinar a técnica com o melhor e o pior desempenhos de precisão. Após esta análise (detalhes no Apêndice D.2), verificamos que as técnicas possuem a seguinte ordenação (decrescente) de desempenho de precisão: $T_2 > T_{5-75\%} > T_{5-50\%} > T_4 > T_{3i} > T_1 > T_{3e} = T_{5-25\%}$.

7.3.1 Conclusões sobre os resultados de Precisão

De acordo com os resultados, a técnica T_2 apresenta o melhor resultado de precisão. Isso pode ser atribuído ao fato de que T_2 apresenta o menor percentual de redução (da suíte de testes) dentre as técnicas. Devido à alta inclusão, a técnica é capaz de descartar casos de teste que não revelam faltas, pois assegura a seleção dos casos de teste que exercitam modificações. Ou seja, como poucos casos de teste são removidos e é observada uma alta inclusão, T_2 apresenta uma alta precisão.

Para as técnicas de seleção aleatória ($T_5 - 25\%$, $T_5 - 50\%$, $T_5 - 75\%$), é observado, novamente, um desempenho esperado, no qual a precisão foi próxima da inclusão da técnica (para a cobertura especificada e as várias replicações). A técnica T_4 apresentou uma precisão (média $35,13$, e desvio-padrão zero) muito inferior ao seu resultado de inclusão (média de $78,57$ e desvio-padrão zero). Algumas vezes, os *clusters* envolvem diversos fluxos de casos de uso, e alguns desses fluxos podem não apresentar faltas. Portanto, a técnica seleciona todos os casos de teste daquele fluxo, o que fornece uma alta inclusão (porcentagem de casos de teste que revelam faltas), porém diminui a precisão (porcentagem de casos de teste que não revelam faltas que foram descartados da suíte reduzida).

Semelhantemente à T_4, a técnica T_1 seleciona muitos casos de teste que exercitam vários caminhos do modelo (i.e. as dependências das modificações no modelo). Diante disto, T_1 não seleciona parte destes casos de teste, já que a maioria das faltas estavam nos trechos modificados do modelo (e não nas dependências da modificação).

O resultado obtido entre as técnicas T_{3i} e T_{3e}, parece intrigante a princípio. Era esperado que a técnica configurada pelo testador experiente (T_{3e}) apresentasse melhor precisão que a técnica configurada por um testador inexperiente (T_{3i}). No entanto, a configuração fornecida pelo testador experiente faz com que a técnica selecione mais casos de teste, permitindo uma maior cobertura dos casos de teste que o usuário poderia executar. Dessa forma, ao selecionar mais casos de teste de diversos fluxos (incluindo os não modificados), a técnica adiciona à suíte selecionada mais casos de teste que não revelavam faltas de regressão. Enquanto isto, a configuração fornecida pelo testador inexperiente, explora poucos fluxos, resultando em uma suíte de testes menor. Uma vez que menos casos de teste são selecionados por T_{3i}, a precisão da técnica aumenta.

7.4 Projeto Experimental 3 - Eficiência

Durante as 100 execuções, os dados da eficiência de cada técnica (o tempo de execução em nanossegundos - ns) são coletados. Estes dados são submetidos à análise descrita na Seção 7.1, e os resultados de cada passo são apresentados a seguir:

1. As premissas de ANOVA não satisfazem os dados de eficiência (Apêndice C.3), portanto é necessário utilizar o teste de Kruskal-Wallis.

2. Aplicamos o teste de Kruskal-Wallis e os resultados são apresentados na Tabela 7.1 (c).

3. O p obtido como resultado do teste foi $0,0001$. Uma vez que o p é menor que o nível de significância considerado no teste ($\alpha = 0,05$), podemos rejeitar a hipótese nula em favor da hipótese alternativa.

4. Diante do resultado do teste de hipótese, podemos afirmar, com um nível de confiança de 95%, que as técnicas se comportam de forma diferente com relação à propriedade de eficiência.

Para obter informações a respeito do desempenho comparativo entre as técnicas, realizamos testes visuais com os intervalos de confiança, assim como testes de Mann-Whitney. Os detalhes desta verificação estão no Apêndice D.3. A partir destes testes, verificamos que a eficiência das técnicas estão ordenadas (de forma decrescente) da seguinte forma: $T_{5-75\%} > T_{5-50\%} > T_{5-25\%} = T_4 > T_{3i} > T_{3e} > T_{2e} > T_{2i} > T_1$. Diante disto, afirmamos que a técnica com o pior desempenho de eficiência é T_1 enquanto que a técnica mais eficiente é $T_{5-75\%}$.

7.4.1 Conclusões sobre os resultados de Eficiência

Observando os dados de tempo de execução (em ns) obtidos durante as execuções, a técnica T_1 apresenta uma diferença muito grande quando comparada com as demais. Esta técnica necessita de muito tempo para executar, pois seu algoritmo precisa realizar uma análise de dependência entre os elementos do modelo, para cada modificação.

No caso deste experimento, onde são estabelecidas apenas 6 modificações, a técnica necessita de muito tempo para realizar a análise de dependência. Considerando a perspectiva de que este tempo ainda é pouco quando comparado ao tempo de executar e analisar os casos de teste não selecionados, a técnica possui uma eficiência inferior às demais técnicas que realizam a mesma atividade (seleção de casos de teste) e necessitam de muito menos tempo para executar.

As técnicas T_{2e} e T_{2i} apresentam uma baixa eficiência. Esta baixa eficiência é observada devido aos passos de realizar a análise de riscos realizados pela técnica. Nesta análise de riscos, uma matriz é construída e percorrida diversas vezes. Além da análise de riscos, a comparação entre os diagramas de atividades (i.e. identificar as modificações no modelo) é custosa, considerando que é necessário manter uma sincronia entre o caminhamento no modelo da versão anterior e da versão delta.

Assim como T_{2e} e T_{2i}, as técnicas T_{3e} e T_{3i} também utilizam uma matriz para a análise das modificações. No entanto, T_{3e} e T_{3i} não realizam um caminhamento no modelo, o que caracteriza a melhora de eficiência observada nestas técnicas quando são comparadas com T_{2i} e T_{2e}.

A técnica T_4 não realiza nenhuma análise com matrizes, apenas identifica porções modificadas no modelo da versão delta realizando comparações com as transições e vértices do modelo da versão base do software. Portanto, T_4 é mais eficiente que T_{2e}, T_{2i}, T_{3e} e T_{3i}, uma vez que estas técnicas utilizam matrizes cujo processamento demanda mais tempo.

As técnicas $T_{5-25\%}$, $T_{5-50\%}$, $T_{5-75\%}$ apresentaram o comportamento esperado. Estas técnicas apresentam os melhores resultados de eficiência, pois estas técnicas apenas percorrem a suíte selecionado, aleatoriamente, os casos de teste de regressão. Quanto maior a cobertura, menor a quantidade de seleções que o algoritmo deve realizar, e mais rápido o algoritmo pára. Esta relação justifica a eficiência obtida entre os níveis de cobertura para estas técnicas ($T_{5-25\%} < T_{5-50\%} < T_{5-75\%}$).

7.5 Projeto Experimental 4 - Potencial de Redução

O potencial de redução de cada técnica analisada foi observado durante as 100 execuções do experimento. Estes dados foram submetidos aos passos de análise descritos na Seção 7.1, e

os respectivos resultados são apresentados a seguir:

1. A partir do teste inicial das premissas de ANOVA, nos dados de redução das técnicas, foi verificado (Apêndice C.4) que não é adequado utilizar ANOVA neste projeto experimental. Portanto, decidimos utilizar o teste de Kruskal-Wallis.

2. Os resultados do teste de Kruskal-Wallis são apresentados na Tabela 7.1 (d).

3. O teste de Kruskal-Wallis apresentou, como resultado, um $p = 0,0001$. Observando que o p é menor que o nível de significância $\alpha = 0,05$ considerado no teste, podemos rejeitar a hipótese nula em favor da hipótese alternativa.

4. A partir do teste de hipótese, podemos afirmar, com um nível de confiança de 95%, que as técnicas se comportam de forma diferente com relação potencial de redução.

A partir deste resultado, é realizada uma investigação comparando os potenciais de redução de cada técnica, procurando identificar o melhor desempenho, i.e. qual técnica é capaz de prover uma maior redução da suíte de testes utilizada no experimento. Os detalhes desta análise podem ser observados no Apêndice D.4. Como resultado desta investigação é observado que as técnicas possuem o seguinte desempenho de potencial de redução: $T_{5-25\%} > T_{3i} > T_{3e} > T_1 > T_{5-50\%} > T_4 > T_{5-75\%} > T_2$, onde as técnicas com os maiores potenciais de redução são a técnica de seleção aleatória com 25% de cobertura ($T_{5-25\%}$), e a técnica WSA-RT (T_{3e} e T_{3i}).

7.5.1 Conclusões sobre o Potencial de Redução das técnicas

A partir do desempenho obtido considerando cada técnica de re-teste seletivo, podemos observar que as técnicas com o melhor potencial de redução são: a técnica de seleção aleatória com 25% de cobertura ($T_{5-25\%}$), e a técnica WSA-RT (T_{3i} e T_{3e}). As técnicas de seleção aleatória apresentam o resultado esperado, em que o nível de cobertura determinou o potencial de redução da técnica. No caso de $T_{5-25\%}$, a baixa cobertura garantiu um alto potencial de redução na suíte de testes de regressão.

A técnica WSA-RT, quando comparada com as demais técnicas, apresenta um melhor potencial de redução, possibilitando a remoção de cerca de 70% (para T_{3i}) e 68% (para T_{3e})

dos casos de teste da suíte de teste de regressão. A técnica WSA-RT foi capaz de identificar que os casos de teste da suíte modificada muito similares aos casos de teste da versão base. Dessa forma, muitos casos de teste, da suíte de regressão, não precisam ser selecionados.

O melhor resultado obtido pela configuração do testador inexperiente pode ser explicado a partir da mesma justificativa apresentada no resultado de precisão, onde a configuração fornecida pelo testador inexperiente permite que a técnica não selecione casos de teste que exercitam diferentes fluxos da aplicação. Ou seja, os casos de teste com maiores valores de probabilidade pertencem ao mesmo fluxo, e o valor de similaridade entre eles é alto, e não são, portanto, selecionados pela técnica. A configuração fornecida pelo testador experiente, por outro lado, permite a exploração de mais fluxos, e portanto, inclui mais casos de teste que capturam faltas de regressão, como foi verificado na investigação da inclusão das técnicas.

As técnicas de Análise de Dependência (T_1) e de Seleção baseada em *Clusters* (T_4) apresentam um potencial de redução médio de 60% e 46%, e desvio-padrão de 3 e 0, respectivamente. Para T_1 o foco em cobrir as dependências das modificações diminui o potencial de redução da técnica, pois as dependências afetadas pelas modificações se propagam por diversos fluxos da aplicação. Dessa forma, a técnica seleciona mais casos de teste.

Semelhantemente, para a técnica T_4, os *clusters* encontrados no modelo utilizado no estudo experimental engoblam diversos casos de uso, o que causou a criação de alguns *clusters* grandes, e como algumas modificações estão inseridas nestes *clusters*, os casos de teste que exercitam estes *clusters* maiores tiveram de ser selecionados. Porém, ao contrário de T_1, estes *clusters* possuem diversas faltas de regressão, o que aumenta a capacidade de inclusão de T_4.

A técnica de Análise de Riscos (T_2) apresenta um potencial de redução muito baixo. Pelo algoritmo da técnica, ao identificar uma transição com a modificação, são selecionados todos os casos de teste que exercitam as transições do fluxo em que esta modificação é realizada. Neste caso, são selecionados muitos casos de teste similares entre si, i.e. que possuem muitas transições em comum e portanto, exercitam os mesmos passos. Sob esta perspectiva, apenas alguns casos de teste que exercitam a aresta modificada poderiam ser selecionados. Apesar de não possuir um alto potencial de remoção a técnica apresenta uma alta inclusão e precisão, o que é consistente sob uma perspectiva geral, pois, como são removidos poucos casos de teste (cerca de 12% da suíte de regressão), a técnica apresenta uma alta cobertura de faltas

de regressão.

7.6 Análise do Projeto Experimental 5 - Densidade de faltas

Durante as 100 execuções das técnicas, são coletados os dados a respeito da densidade de faltas das técnicas. Os passos da análise descritos na Seção 7.1, são aplicados nos dados deste projeto experimental e os resultados são apresentados a seguir:

1. Após verificar que as premissas de ANOVA não são respeitadas pelos dados de densidade de faltas analisados (Apêndice C.5), decidimos utilizar o teste de Kruskal-Wallis.

2. Os resultados do teste de Kruskal-Wallis são apresentados na Tabela 7.1 (e).

3. O resultado do teste foi um $p = 0,0001$, que é menor que o nível de significância $\alpha = 0,05$ considerado no teste. Diante disto, podemos rejeitar a hipótese nula em favor da hipótese alternativa.

4. A partir do teste de hipótese, podemos afirmar, com um nível de confiança de 95%, que as técnicas se comportam de forma diferente com relação à densidade de faltas.

A partir do resultado do teste de Kruskal-Wallis, é realizada uma investigação entre cada técnica, com o objetivo de identificar a técnica com a melhor densidade de faltas. Os detalhes desta investigação podem ser encontrados no Apêndice D.5. O resultado obtido apresenta o seguinte desempenho de densidade de faltas entre as técnicas: $T_4 > T_3 > T_1 = T_5 > T_2$. Portanto, a que apresenta maior densidade de faltas, é T_4, enquanto que T_2 apresenta a menor densidade de faltas.

7.6.1 Conclusões sobre os resultados da densidade de faltas das técnicas

Diante do desempenho de densidade de faltas obtido ao comparar os resultados das execuções das técnicas, podemos concluir os seguintes aspectos com relação às técnicas. A técnica de seleção baseada em *Clusters* (T_4) apresenta o melhor resultado de densidade de

faltas, indicando que esta técnica é capaz de reduzir a suíte de teste e manter uma alta cobertura de faltas de regressão. Esta propriedade pode ser atribuída à divisão do modelo em *clusters* para identificar as modificações.

Além de possibilitar a localização das modificações no modelo, a criação dos *clusters* também possibilita a identificação dos elementos do modelo (e.g. transições, fluxos e estados) que são afetados por estas modificações e um perímetro que limita a propagação destas modificações. A partir desta 'segmentação' do modelo em *clusters*, é possível selecionar casos de teste que cobrem determinados conjuntos de fluxos da aplicação mais propícios a apresentarem faltas de regressão.

A técnica WSA-RT (T_3) também apresenta um bom resultado de densidade de faltas, pois é capaz de reduzir significativamente a suíte de teste (cerca de 68% a 70% dos casos de teste de regressão) e manter uma boa cobertura de faltas de regressão (cerca de 50% da suíte reduzida captura faltas de regressão). Este desempenho é atribuído pela análise de similaridade realizada entre a suíte de testes da versão base e a suíte de testes da versão delta.

WSA-RT é capaz de remover casos de teste muito similares entre as duas suíte, evitando que as mesmas funcionalidades modificadas sejam testadas repetidamente. A configuração fornecida pelo testador, ainda que não afete diretamente a densidade de faltas da técnica, permite que mais casos de teste da suíte de regressão sejam selecionados, diminuindo o potencial de redução, mas aumentando a capacidade de revelar faltas de regressão da suíte reduzida.

A técnica de Análise de Dependência em Máquinas de Estados (T_1) apresenta uma densidade de faltas semelhante à técnica de seleção aleatória (T_5). T_1 apresenta uma densidade de faltas média de $41,28\%$ com desvio-padrão de $5,54\%$, enquanto que T_5 apresenta uma densidade média de $42,37\%$ e desvio-padrão de $4,63\%$. Para cada execução de T_1 as dependências que são selecionadas podem causar uma redução na quantidade de casos de teste que não capturam faltas de regressão selecionados para a suíte reduzida.

Ao selecionar menos casos de teste que não capturam faltas de regressão, a densidade de faltas da suíte reduzida aumenta. No caso de T_1, isto explica alguns resultados bons para a técnica, uma vez que, em algumas execuções, é observada uma densidade de faltas de 50% na suíte selecionada. No entanto, o alto desvio-padrão de T_1 faz com que sua densidade de faltas se assemelhe à densidade de T_5. Diante disto, a escolha das dependências analisadas

e, portanto, dos casos de teste selecionados para a suíte selecionada causa um alto desvio-padrão em T_1 diminuindo a sua densidade de faltas.

A técnica de Análise de Riscos em Diagramas de Atividades (T_2) apresenta a pior densidade de faltas. Este resultado é contrastante com os resultados obtidos a respeito da inclusão e precisão da técnica, no entanto, coerente com o resultado obtido a respeito do potencial de redução do algoritmo. Uma vez que T_2 não reduz significativamente a suíte de teste de regressão, a técnica mantém uma alta cobertura de casos de teste que revelam faltas de regressão. Porém, a suíte permanece com muitos casos de teste.

Diante disto, a suíte de teste fica pouco densa. Se observarmos o quociente entre quantidade de casos de teste que capturam faltas de regressão (numerador), e a quantidade de casos de teste na suíte reduzida (denominador) (Equação 2.4), verificamos que o denominador é muito alto, quando comparado com as demais técnicas (T_2 reduz, em média, apenas 12% da suíte de teste, enquanto que todas as outras técnicas reduzem mais que 25% da suíte de teste de regressão). Portanto, para que a densidade de faltas da técnica aumente, é necessário que o potencial de redução do algoritmo de T_2 seja melhorado.

7.7 Análise das Técnicas

Nesta seção são apresentadas as características avaliadas em cada técnica. As características são apresentadas a partir da análise obtida em cada projeto experimental, com o objetivo de prover uma visão do resultado de cada técnica, relacionando as respectivas vantagens e desvantagens de utilizá-la.

7.7.1 Técnica de Análise de Dependência em Máquinas de Estados Finitas Estendidas – T_1

O algoritmo executado pela técnica T_1 realiza uma análise nas transições e estados do modelo identificando dependências de dados e controle entre as modificações realizadas na máquina de estados e os estados, transições e variáveis. A partir destas dependências, são identificados os elementos do modelo afetados pela modificação, e então, os casos de teste que exercitam estes elementos afetados são selecionados para a suíte de testes reduzida. Um resumo dos

resultados de T_1 é apresentado na Tabela 7.2.

Tabela 7.2: Resumo dos resultados (média aritmética) observados para T_1.

Inclusão	**Precisão**	**Eficiência (*ns*)**	**Potencial de Redução**	**Densidade de Faltas**
50%	$28,16$	$1,52 \cdot 10^{11}$	$60,15\%$	$34,78\%$

Considerando que a técnica realiza uma densa análise de dependência entre todos os elementos do modelo, a partir das modificações, é esperado um grande tempo de processamento da técnica. Este aspecto foi observado no experimento, onde a técnica apresentou o pior desempenho de eficiência. A técnica apresentou um tempo de processamento médio de $152,25$ segundos (com desvio-padrão de $2,3$ s), ou seja $2,53$ minutos para executar. Se for considerado também o tempo de configuração da técnica (redução da redundância na suíte de teste, leitura dos arquivos XMI, etc.), o tempo de processamento aumenta para cerca de 30 minutos.

A impressão obtida durante a execução da técnica no experimento é que o aumento na quantidade de modificações no modelo caracteriza um aumento significativo na quantidade de tempo na análise de dependências. Diante disto, é recomendado melhorar aspectos da técnica, como as condições de parada e o caminhamento no modelo, assim como melhorar aspectos de execução do algoritmo, como estruturas de repetições ou estruturas de dados.

Apesar de ser pouco eficiente, T_1 apresenta um comportamento regular para as demais variáveis dependentes analisadas. A partir dos resultados de inclusão da técnica, verificamos que T_1 selecionou 50% dos casos de teste que revelam faltas de regressão, sendo capaz de capturar em média $6,57$ (com desvio-padrão de $0,68$) faltas de um total de 10 faltas especificadas.

Podemos observar, então, que, apesar de selecionar apenas 50% dos casos de teste que capturam faltas de regressão T_1 foi capaz de cobrir mais que a metade das faltas presentes no modelo. Esta característica é observada na técnica, pois são selecionados casos de teste com diferentes padrões de interação (i.e. possuem um mesmo conjunto de dependências no modelo). Dessa forma, os casos de teste selecionados são diferentes, aumentando a quantidade de caminhos cobertos pela suíte selecionada.

Por sua vez, a precisão apresentada pela técnica é baixa, em especial quando comparada com as técnicas T_2 e T_4. Este aspecto é atribuído à análise de dependência. Durante esta

análise, a técnica identifica os elementos do modelo afetados pela modificação e seleciona os casos de teste que cobrem estas dependências afetadas.

Diante disto, observamos que durante o experimento, a técnica seleciona diversos casos de teste que não estão relacionados com as modificações em si, no entanto estão relacionados com algum elemento do modelo afetado pela modificação. Estes casos de teste possuem, geralmente, padrões de interações distintos dos padrões encontrados nos casos de teste que possuíam as faltas de regressão. Portanto, o algoritmo da técnica selecionou diversos casos de teste que não possuíam faltas de regressão, o que justifica a baixa precisão observada neste estudo experimental.

Apesar dos baixos resultados de eficiência, inclusão e precisão, a técnica apresentou resultados muito bons com relação ao seu potencial de redução. Assim como defendem os autores da técnica [Korel et al. 2002], T_1 apresentou um bom potencial de redução, sendo capaz de reduzir em média $60,15\%$ (com desvio padrão de $2,97$) da suíte de testes de regressão. Durante o experimento, foi observado que diversos dos casos de teste que exercitavam as modificações possuíam os mesmos padrões de interação, e portanto, estes casos de teste não foram selecionados para a suíte reduzida.

Sob a perspectiva da densidade de faltas da técnica, é observado que T_1 possui, em sua suíte reduzida, uma média de $41,29\%$ (com desvio-padrão de $5,54$) de casos de teste que capturam faltas de regressão. Assim como na análise de inclusão, T_1 apresenta um desempenho regular, onde não é observado nem o melhor nem o pior desempenho. A densidade de faltas regular de T_1 é justificada pelo seu desempenho de inclusão e precisão.

Apesar de selecionar 50% de todos os casos de teste que capturam falta de regressão, a maioria dos casos de teste da suíte reduzida não capturam faltas, portanto a densidade de faltas da técnica ficou abaixo de 50%. Esta densidade teria sido menor se não fosse observado um bom potencial de redução da técnica.

Em resumo, podemos destacar as seguintes características da técnica de análise de dependência em máquinas de estados:

- Vantagens:
  - A técnica apresenta um alto potencial de redução (cerca de 60% da suíte de regressão);

  - A análise de dependência aumenta a confiança no processo de seleção, por analisar as faltas por efeitos colaterais.
  - A técnica trata diversas modificações como remoção, adição e modificação de transições no modelo.

- Desvantagens:
  - É pouco eficiente, pois necessita de muito tempo para executar.
  - O aumento significativo na quantidade de modificações pode inviabilizar, computacionalmente, a execução do algoritmo de análise de dependência.
  - O custo de desenvolvimento da técnica é alto, pois requer diversas análises estruturais nos elementos do modelo.

Ao observar os resultados da técnica, verificamos que a análise de dependência contribui para o desempenho regular da técnica. Ainda que apresente pouco desempenho de precisão, a análise que a técnica realiza fornece confiança para a qualidade da suíte de testes resultantes, pois, pelo algoritmo, as dependências encontradas a partir das modificações guiam o processo de seleção. Para obter uma conclusão mais precisa a respeito do impacto desta análise de dependência no desempenho da técnica e na qualidade da suíte reduzida, é necessário realizar um outro estudo experimental, utilizando diversos modelos de faltas e máquinas de estados.

7.7.2 Técnica de Seleção baseada em Análise de Riscos – T_2

A técnica de seleção baseada em Análise de Riscos, referenciada como T_2 durante todo o estudo experimental, realiza uma análise que considera tanto aspectos do modelo da aplicação, como do custo relacionado com a descoberta de faltas de regressão. Diante disto, são consideradas 4 configurações da técnica: 2 configurações fornecidas por testadores experientes e 2 configurações fornecidas por testadores inexperientes. Os testadores fornecem, como configuração, os valores de custos e severidade de faltas necessários para a execução da técnica. A técnica T_2 foi executada, e os respectivos resultados são resumidos na Tabela 7.3.

Tabela 7.3: Resumo dos resultados (média aritmética) observados para T_2.

Inclusão	**Precisão**	**Eficiência (*ns*)**		**Potencial de Redução**	**Densidade de Faltas**
		Inexperiente	Experiente		
$92,85\%$	$85,13\%$	45831890	$1,01 \cdot 10^8$	$12,06\%$	$38,23\%$

Durante cada projeto experimental são realizadas análises para verificar se a configuração da técnica afeta significativamente os resultados obtidos. A partir desta análise, verificamos que o nível de experiência do testador influencia, significativamente, apenas a eficiência da técnica, i.e. o tempo de execução (medido em nanossegundos) da técnica. Esta diferença é observada pois os valores especificados pelos testadores experientes apresentam maior variância, o que exige maior tempo de processamento.

Grande parte do tempo de execução da técnica (cerca de 95% do tempo total de T_2) é utilizado durante a etapa de configuração do algoritmo. Esta etapa contempla: a atribuição de custo aos casos de teste de regressão, definição da severidade das faltas de regressão e a construção da matriz de risco. Para diminuir o tempo de execução, é recomendado estruturar melhor os dados, já que o acesso e busca em uma matriz pode ser custoso, em especial quando o tamanho da matriz cresce de acordo com a quantidade de casos de teste de regressão. Devemos lembrar que uma suíte de testes de regressão tende a crescer bastante durante o ciclo de vida do software [Korel et al. 2002], o que pode aumentar significativamente o tempo de execução da técnica, prejudicando, portanto, sua eficiência.

Apesar de possuir uma baixa eficiência, T_2 apresenta bons resultados de inclusão. Selecionando cerca de 95% dos casos de teste que capturam faltas de regressão, o algoritmo de T_2 é capaz de capturar 9 das 10 faltas de regressão. A falta que não foi capturada estava em um dos elementos da nova versão do software, e uma vez que T_2 selecionou apenas casos de testes da versão base do software, não foi possível obter uma inclusão de 100%, característica das denominadas *técnicas seguras*.

Ainda sob esta perspectiva, T_2 seleciona alguns casos de testes obsoletos, o que pode dificultar a execução dos casos de teste [Rothermel and Harrold 1997]. A principal característica que contribui para a alta inclusão da técnica foi a alta quantidade de casos de teste na suíte reduzida. O algoritmo reduz a suíte de testes de 58 casos de teste para 51 casos de teste, ou seja, é reduzida apenas 12% da suíte de regressão.

A técnica poderia ter apresentado 100% de inclusão se fosse executada na suíte obtida a partir da versão delta do sistema, como é realizado por demais técnicas de re-teste seletivo baseado em especificação [Korel et al. 2002, Subramaniam et al. 2009]. Enquanto que o bom resultado de inclusão de T_2 é atribuído ao baixo potencial de redução da téncica, os resultados de precisão, por sua vez, contribuem para o bom desempenho geral da técnica. Apesar de remover poucos casos de teste, a técnica remove apenas os casos de teste que não cobrem faltas de regressão.

Para caracterizar o desempenho da inclusão e precisão da técnica com o potencial de redução, é investigada a densidade de faltas. Como observado nos resultados obtidos, o baixo potencial de redução faz com que a técnica apresente a pior densidade de faltas dentre as técnicas analisadas. Isto indica que, apesar de capturar muitas faltas de regressão, a técnica não reduz significativamente a suíte de testes de regressão. Dessa forma, é importante que o potencial de redução da técnica seja melhorado para que a suíte reduzida possa ser executada com menos custos.

Diante dos aspectos apresentados, podemos observar as seguintes características na técnica:

- Vantagens:
  - A técnica captura muitos casos de teste que revelam falta de regressão;
  - A análise de risco permite que casos de testes com faltas críticas sejam selecionados;
  - Poucos casos de teste que não revelam faltas de regressão são selecionados, ou seja, a técnica possui uma alta precisão.
- Desvantagens:
  - Não é recomendada quando os recursos disponíveis para a realização do teste de regressão são poucos;
  - A técnica é pouco eficiente, e o seu tempo de execução pode crescer significativamente ao longo do crescimento da suíte de regressão;
  - É necessário realizar manutenção na suíte reduzida (analisar e identificar os casos de teste obsoletos);

– O baixo potencial de redução seleciona muitos casos de teste para a execução.

Portanto, podemos concluir que a técnica de Seleção baseada em Análise de Riscos é adequada para um cenário onde pouca redução da suíte de testes de regressão é desejada; por exemplo, aplicações críticas, como as de sistemas de tempo real, ou aplicações médicas. Geralmente, nessas aplicações, a captura de faltas é uma prioridade, e portanto, é necessário executar muitos casos de teste.

Os resultados mostram que T_2 é capaz de reduzir, mesmo que pouco, a quantidade de casos de teste. Dessa forma, com esta técnica, podemos executar menos casos de teste de regressão e ainda assim manter uma alta cobertura de faltas de regressão. Além disto, é possível observar que o algoritmo de identificação das modificações pode ser aplicado em modelos nos quais seja possível comparar elementos como transições e vértices. Exemplos destes tipos de modelos seriam máquinas de estados, sistemas de transições rotuladas, dentre outros. Diante disto, é possível aproveitar a técnica para outros modelos além do Diagrama de Atividades.

7.7.3 *Weighted Similarity Approach for Regression Testing* (WSA-RT) – T_3

WSA-RT procura identificar os casos de teste da versão modificada que exercitam trechos menos similares aos trechos exercitados pelos casos de teste da versão base. O perfil de uso utilizado na técnica auxilia na seleção de suítes com maior credibilidade, i.e. casos de teste que exercitam funcionalidades que o usuário executará mais freqüentemente [Kaner 2003]. Um resumo dos resultados apresentados por T_3 pode ser observado na Tabela 7.4.

Tabela 7.4: Resumo dos resultados (média aritmética) observados para T_3.

Inclusão		**Precisão**		**Eficiência (*ns*)**	
Inexperiente	Experiente	Inexperiente	Experiente	Inexperiente	Experiente
$34,30\%$	$35,85\%$	$29,78$	$25,78\%$	14588457	14831671

Potencial de Redução		**Densidade de Faltas**
Inexperiente	Experiente	
$70,10\%$	$68,47\%$	$49,38\%$

Um dos primeiros aspectos verificados é se o nível de experiência do testador que fornece a configuração do perfil de uso afeta, significativamente, o desempenho da técnica, para cada variável dependente analisada (inclusão, precisão, eficiência, potencial de redução e densidade de faltas). Após esta análise, é observado que o nível de experiência do testador influencia na inclusão, precisão, eficiência e potencial de redução da técnica. A densidade de faltas da técnica não apresenta diferença estatisticamente significativa para os dois níveis de experiência do testador analisados (testador experiente, e testador inexperiente).

A partir dos resultados obtidos na execução de WSA-RT durante o experimento, é possível observar que a baixa inclusão e precisão da técnica podem ser atribuídos ao seu alto potencial de redução. Ao reduzir a quantidade de casos de teste, a suíte perde muitos casos de teste que revelam faltas de regressão. Investigando os dados, verificamos que WSA-RT não seleciona muitos casos de teste similares aos casos de teste da versão base.

No entanto, WSA-RT também mantém muitos casos de teste da suíte da versão delta que são similares entre si. Dessa forma, a técnica possui uma baixa cobertura das transições do modelo, o que diminui, de forma significativa, sua inclusão e precisão. Diante disto, é necessário refinar o algoritmo de seleção para o contexto de Teste de Regressão, de forma que a suíte reduzida mantenha casos de teste menos similares aos casos de teste da versão base, e também, menos similares entre si.

Sob a perspectiva do potencial de redução, WSA-RT apresenta os melhores resultados dentre as técnicas analisadas. O algoritmo é capaz de reduzir a suíte de testes de regressão de 65 para cerca de 20 casos de teste, o que caracteriza um potencial de redução de aproximadamente 70% da suíte de testes de regressão. Portanto, WSA-RT é recomendada quando poucos recursos são disponibilizados, e é necessário reduzir bastante, a suíte de testes de regressão.

O alto potencial de redução da técnica contribui para o bom desempenho da técnica na variável de densidade de faltas, de forma que, em média, 49% (e desvio-padrão de $1,8$) dos casos de teste da suíte selecionada capturam faltas de regressão. Apesar disto, a técnica captura em média 5 (com desvio-padrão de $0,321$) faltas de um total de 10 faltas presentes na suíte de regressão. Assim como observado na inclusão, a suíte reduzida manteve casos de teste muito similares entre si, o que causa uma cobertura das mesmas faltas de regressão.

Diante dos aspectos analisados, são observadas as seguintes características para WSA-

RT:

- Vantagens:
  - A técnica apresenta um alto potencial de redução (cerca de 70% da suíte de regressão);
  - A técnica apresenta uma alta densidade de faltas.
  - É adequada para cenários onde os recursos para a execução dos casos de teste são poucos, e uma grande redução da suíte de testes é desejada.
  - O perfil de uso utilizado pela técnica permite capturar faltas de regressão que seriam exercitadas por um usuário.
  - Necessita apenas da configuração do testador e das suítes de testes das versões base e modificada para executar.
- Desvantagens:
  - A técnica apresenta baixa precisão e inclusão.
  - A suíte selecionada mantém casos de teste muito similares entre si.
  - Captura poucas faltas de regressão, uma vez que seleciona casos de teste similares entre si.
  - O desempenho da técnica é dependente do nível do testador, o que indica a necessidade de treinamento, ou contratação de testadores experientes.

A partir dos resultados obtidos para a técnica WSA-RT, podemos observar que melhorias podem ser realizadas no algoritmo da técnica para aumentar a inclusão e precisão desta. O primeiro passo é refinar o algoritmo para manter casos de teste menos similares entre si na suíte selecionada. Esta característica deve ser incorporada no algoritmo atual, onde casos de teste menos similares aos da versão anterior são selecionados para a suíte reduzida.

Dessa forma, é possível obter uma suíte reduzida com os casos menos similares aos da versão anterior (i.e. que exercitam os trechos modificados do sistema) e também menos similares entre si (i.e. que exercitam diversas transições do modelo da versão delta). A partir deste melhoramento, é possível manter o potencial de redução (já que a suíte resultante no

algoritmo atual apresenta muita redundância) e aumentar a inclusão e precisão da técnica. No entanto, é necessário submeter estes aspectos a um outro estudo experimental, a partir do qual, fosse possível observar a melhoria no desempenho da técnica.

7.7.4 Técnica de Seleção baseada em *Clusters* – T_4

A técnica de seleção baseada em *clusters* (T_4) é utilizada no experimento por apresentar bons desempenhos de inclusão e precisão quando utilizada no re-teste seletivo baseado em código [Rothermel and Harrold 1996]. O algoritmo de identificação das modificações proposto pela técnica foi adaptado e aplicado ao contexto de especificação na expectativa de obter um desempenho similar ao encontrado no contexto de código. A Tabela 7.5 apresenta um resumo com o desempenho de T_4.

Tabela 7.5: Resumo dos resultados (média aritmética) observados para T_4.

Inclusão	**Precisão**	**Eficiência (*ns*)**	**Potencial de Redução**	**Densidade de Faltas**
$78,57\%$	$35,13\%$	13502	$46,15\%$	$62,85\%$

Sob esta perspectiva, são realizadas as análises das características da técnica baseada em *clusters* com relação às variáveis dependentes escolhidas para o estudo experimental. Observando os resultados de inclusão, verificamos que a técnica é capaz de selecionar $78,57\%$ dos casos de teste que capturam faltas de regressão. Considerando que as faltas de regressão estão nos *cluster* com modificações, a técnica é capaz de selecionar os casos de teste que exercitam estas faltas.

Essa característica, no entanto, prejudica a precisão da técnica, uma vez que, após executar a técnica, são obtidos alguns *clusters* grandes, i.e. *clusters* que possuem muitas transições do modelo. Dessa forma, são selecionados muitos casos de teste, de um mesmo *cluster*, que não capturam faltas de regressão.

Por outro lado, também observamos a presença de muitos *clusters* pequenos, o que possibilita um grande potencial de redução de casos de teste (cerca de 46% da suíte de regressão), já que os pequenos *clusters* não possuem transições modificadas. Outro aspecto positivo da técnica é a eficiência, pois ela é capaz de realizar uma seleção com uma alta inclusão, e em pouco tempo de execução (em média $13502,29\ ns$ com desvio-padrão de $7393,62\ ns$). No entanto, é necessário verificar a escalabilidade do algoritmo, considerando que modelos

maiores necessitariam de muito mais tempo para executar, já que a técnica caminha em dois modelos (os modelos da versão delta e da versão base) simultaneamente.

A inclusão e o potencial de redução da técnica beneficiam, de forma significativa, a densidade de faltas obtida na suíte reduzida. Mesmo com uma precisão baixa, a quantidade de casos de teste que revelam faltas caracterizou, em média, $62,85\%$ da suíte reduzida. Dessa forma, a técnica fornece confiabilidade na capacidade de redução, permitindo que menos casos de teste sejam executados, porém obtendo uma alta cobertura de faltas de regressão. Podemos, então, observar as seguintes características na técnica:

- Vantagens:
  - A técnica possui um alto potencial de redução (cerca de 46% da suíte de regressão);
  - A divisão do modelo em *clusters* facilita a identificação de faltas;
  - É possível manter uma boa cobertura de casos de teste que revelam faltas de regressão após reduzir a suíte;
  - A técnica apresenta bons resultados de eficiência e densidade de galtas.
- Desvantagens:
  - A seleção da técnica foi pouco precisa, i.e. selecionou muitos casos de teste que não revelam faltas;
  - O modelo utilizado deve representar fluxos da aplicação, em alguns cenários com interrupções a técnica pode não ser aplicável ou adequada.

É importante lembrar que estes resultados são obtidos a partir dos elementos utilizados neste experimento (i.e. o modelo de faltas e o objeto do experimento). Para generalizar os resultados, é aconselhável que este experimento seja repetido aumentando a quantidade de objetos e modelos de faltas utilizados. Apesar disto, a quantidade de execuções da técnica foi maior que o tamanho mínimo da amostra calculado para obter significância estatística, o que reforça a confiabilidade dos resultados obtidos.

7.7.5 Técnica de Seleção Aleatória de Casos de Teste – T_5

Escolhemos a técnica de seleção aleatória de casos de teste (T_5), para o estudo experimental, por dois motivos. O primeiro motivo é que esta técnica é geralmente utilizada na indústria, onde o próprio testador, seleciona, de forma aleatória ou *ad hoc*, alguns casos de teste para serem executados. O segundo motivo é que esta técnica é amplamente utilizada em estudos comparativos com outras técnicas de seleção de casos de testes [Graves et al. 2001, Graves et al. 1998, Cartaxo et al. 2007, Cartaxo et al. 2009]. Em estudos experimentais, o desempenho da técnica aleatória pode variar de acordo com a cobertura de casos de testes especificada, i.e. a quantidade de casos de teste que devem ser selecionados na suíte de testes de regressão, e também de acordo com a quantidade de repetições realizadas. Os resultados obtidos com a técnica T_5 são apresentados na Tabela 7.6.

Tabela 7.6: Resumo dos resultados (média aritmética) observados para T_5.

Inclusão			**Precisão**			**Eficiência** (ns)		
25%	50%	75%	25%	50%	75%	25%	50%	75%
23,75%	48,89%	73,35%	25,27%	49,48%	74,21%	13502,89	9722,25	6813,68

Potencial de Redução			**Densidade de Faltas**
25%	50%	75%	
25%	50%	75%	42,37%

O único parâmetro de configuração necessário para utilizar a técnica é o percentual de cobertura. Este percentual define a quantidade de casos de teste selecionados. Os valores de cobertura 25%, 50% e 75%, escolhidos para este experimento, fornecem uma visão da suíte reduzida ao serem realizadas uma grande, uma média e uma pequena redução, respectivamente. Para realizar esta análise, é verificado se este parâmetro de cobertura causa uma diferença significativa nos resultados de cada variável dependente. Como resultado, é observado que o percentual de cobertura possui uma relação direta com 4 de 5 variáveis analisadas. São elas: inclusão, precisão, eficiência e o potencial de redução.

A densidade de faltas, no entanto, não foi afetada, significativamente, pelo percentual de cobertura. Este resultado era esperado, pois, para cada percentual de cobertura, um valor grande no tamanho da suíte reduzida era ponderado por um valor também grande na quantidade de casos de teste com faltas de regressão. Os percentuais de cobertura de 25%, 50% e

75% apresentam, respectivamente e em média, $41,56\%$, $42,78\%$, $42,79\%$ dos casos de teste selecionados revelando faltas de regressão. Estes valores refletem bem que não há diferença, estatisticamente significativa, entre os diferentes percentuais de cobertura. No entanto, o desvio-padrão obtido para cada percentual de cobertura foi, respectivamente: $11,07\%$, $6,08\%$ e $3,70\%$. O alto desvio-padrão obtido nos resultados é atribuído à natureza aleatória da técnica.

Os resultados obtidos para a inclusão, precisão e potencial de redução das técnicas de seleção aleatória (cada uma com seu respectivo percentual de cobertura), correspondem aos resultados esperados, nos quais a inclusão foi próxima do percentual de cobertura de cada técnica, a precisão foi próxima, do valor de inclusão, mesmo que algumas unidades inferior (uma vez que 100% de precisão é inatingível [Rothermel and Harrold 1997]). O potencial de redução da técnica corresponde precisamente ao percentual de cobertura especificado, já que o testador especifica a quantidade da suíte de regressão que ele ou ela deseja selecionar.

Com relação à variável de eficiência, a técnica apresenta um desempenho muito bom, independente do percentual de cobertura escolhido. Este resultado também era esperado, pois a técnica não realiza nenhuma análise no modelo. São realizadas, apenas, iterações sobre a suíte de testes, selecionando aleatoriamente os casos de teste até que a cobertura especificada seja atingida. Sob esta perspectiva, a técnica é muito eficiente, pois seu custo de execução é baixo ao mesmo tempo que o seu potencial de redução é definido pelo testador.

A partir dos aspectos apresentados, podemos observar as seguintes características a respeito da técnica:

- Vantagens:
  - A técnica é computacionalmente eficiente;
  - O custo de desenvolvimento da técnica é muito baixo, considerando que não envolve análises complexas no modelo, apenas a iteração e seleção de casos de teste em uma suíte de testes de regressão.
  - O testador define o percentual de cobertura de acordo com os recursos disponíveis, aumentando a versatilidade da técnica aos diversos cenários de custo do processo de teste.
- Desvantagens:

  - A ausência de um critério de seleção, ou análise de modificações prejudicam, significativamente, a confiança na qualidade da suíte reduzida.
  - O alto desvio-padrão observado nos dados obtidos da técnica indica uma baixa confiabilidade na cobertura de faltas de regressão.

Uma das principais vantagens da técnica de seleção aleatória é a sua eficiência e versatilidade. Esta técnica não necessita de formatos específicos de modelos, como máquinas de estados, grafos, matrizes, dentre outros. É necessário apenas uma suíte de testes de regressão. A opção de definir o percentual de cobertura torna a técnica versátil, uma vez que a disponibilidade dos recursos determina a quantidade de casos de teste que devem ser selecionados.

Apesar destas vantagens, a técnica apresenta pouca confiabilidade nos seus resultados, uma vez que nenhuma análise é realizada na suíte de testes, ou no modelo da aplicação, relacionando as faltas de regressão com as modificações ou aos elementos afetados por estas. Diante disto, ainda é recomendado utilizar as demais técnicas, quando o tempo necessário para a execução das técnicas, ou os recursos, não são tão restritos. Algumas técnicas apresentaram um potencial de redução semelhante à técnica de seleção aleatória (e.g. WSA-RT e a seleção baseada em análise de dependência em máquinas de estados), e mantendo uma média de densidade de faltas maior.

7.8 Análise sobre a Generalidade das Técnicas

A propriedade de generalidade das técnicas de re-teste seletivo caracteriza a habilidade destas serem aplicáveis em uma ampla variedade de situações [Rothermel and Harrold 1996]. Esta habilidade envolve dependências e limitações ao utilizar a técnica, como, por exemplo, uma dependência ferramental, uma restrição no formato do modelo, dentre outros. Sob esta perspectiva, as dependências e características da execução das técnicas foram analisadas e os resultados são apresentados nesta seção.

Para obter resultados acerca da generalidade das técnicas alguns aspectos de cada técnica foram verificados, dentre eles:

- **Aspectos do modelo:** São observados aspectos como a linguagem utilizada pelo mo-

delo (e.g. UML, STR, GFC, dentre outros), e a complexidade de construção e uso deste.

- **Natureza da Análise realizada:** As técnicas são verificadas de acordo com a análise realizada, ou seja, se realizam uma análise de dependência, uma cobertura de modificações, uma análise de riscos, dentre outras.

- **Natureza das modificações:** É verificada a versatilidade da técnica em lidar com os diversos tipos de modificações como a adição e/ou remoção de funcionalidades, a modificação de funcionalidades, i.e. a modificação de rótulos no modelo, ou de estruturas (e.g. fluxos, estados, dentre outros).

- **Dependências:** Para cada técnica analisada é investigada a necessidade de um suporte ferramental ou operacional para o processo de seleção dos casos de teste.

7.8.1 Generalidade da Técnica de Análise de Dependência em Máquinas de Estados

Observando a técnica de análise de dependência em máquinas de estados (T_1) sob os critérios de generalidade, podemos concluir os seguintes aspectos a respeito da técnica. Inicialmente, são utilizados dois modelos pela técnica: a máquina de estados e o grafo estático de dependência (GED). Sob a perspectiva do testador, a utilização de uma máquina de estados (que pode ser representada por um Diagrama de Máquina de Estados UML) é conveniente, pois os estados do sistema são facilmente representados e visualizados no modelo.

Por outro lado, grande parte do processamento da técnica está no GED, e apesar de ser transparente para o testador, o desenvolvimento da técnica pode envolver muito custo e esforço. Considerando que a técnica necessita do GED para executar, a generalidade de T_1 é prejudicada, pois para aplicar a técnica é necessário que os seus passos e a análise de dependência sejam entendidos. Este entendimento não é trivial, e envolve muitos aspectos estruturais de uma máquina de estados inviabilizando o uso da técnica (sem nenhuma alteração inicial em seu algoritmo) com outros modelos como, por exemplo, o Diagrama de Atividades ou o Diagrama de Componentes UML.

Apesar da generalidade da técnica ser prejudicada pelos modelos que esta utiliza, a

técnica realiza diversas análises nestes modelos. Em um primeiro passo são identificadas as modificações realizadas no modelo, para que então, a técnica identifique as dependências associadas a cada modificação realizada. Diante disto, a técnica trata diversos aspectos de análise, o que contribui com a sua generalidade.

Os mesmos aspectos podem ser ditos com relação à natureza das modificações tratadas por T_1. A técnica descreve, precisamente, como lidar com a adição, remoção e modificação de transições (que representam funcionalidades) no modelo. Além de fornecer esta perspectiva para cada natureza da modificação, a técnica provê a perspectiva das entidades afetadas tanto no modelo e no trecho modificado quanto nos efeitos colaterais da modificação. Dessa forma, T_1 possui uma alta generalidade com relação às naturezas de modificações tratadas pela técnica.

Não foram observadas dependências ferramentais ou operacionais por T_1. Apesar de ser fortemente acoplada ao modelo de entrada (a máquina de estados) é possível que este seja representado por diversos formatos, desde um arquivo XMI de um Diagrama de Máquina de Estados UML a formatos específicos definidos pelo testador. Ainda sob esta perspectiva, o testador não necessita, a princípio, configurar nenhum parâmetro na técnica, de forma que é necessário apenas fornecer o modelo de entrada, as modificações e a suíte de testes para a técnica, que esta executa automaticamente.

Diante disto, podemos concluir que a técnica apresenta limitações relacionadas, apenas, com os aspectos dos modelos utilizados, em especial a complexidade de entendimento e desenvolvimento da técnica, com o objetivo de sua automatização. Os demais aspectos investigados em T_1, i.e. as naturezas das modificações e da análise, assim como as dependências da técnica, contribuem de forma positiva para a sua generalidade.

7.8.2 Generalidade da Seleção baseada em Análise de Riscos

A técnica de seleção baseada em análise de riscos (T_2) foi investigada sob os critérios de generalidade e os seguintes aspectos foram concluídos. A primeira perspectiva observada foi o modelo utilizado pela técnica. No trabalho em que descrevem a técnica [Chen et al. 2002], os autores argumentam que a técnica pode ser usada em diversos modelos, desde que seja possível representar as funcionalidades e fluxos da ferramenta neste modelo.

Os autores ilustram essa versatilidade da técnica executando o mesmo algoritmo em um

Grafo de Fluxo de Controle (GFC) e em um Diagrama de Atividades. Para este experimento utilizamos o algoritmo em um Diagrama de Atividades, e foi observado que os passos do algoritmo podiam ser executados em um modelo semelhante, ou seja, os passos do algoritmo não estão restritos aos elementos do Diagrama de Atividades, o que contribui para a generalidade da técnica.

Considerando a natureza da análise realizada pela técnica, é observado que, em uma primeira etapa o algoritmo identifica as modificações realizadas no modelo. Ao identificá-las, o algoritmo seleciona todos os casos de teste que exercitam esta modificação. Além disto, a técnica não diferencia os tipos de modificação, apenas identifica elementos diferentes entre o modelo da versão base e o modelo da versão modificada do sistema.

Sob esta perspectiva, a técnica é prejudicada em generalidade, pois a seleção poderia ser mais criteriosa, permitindo um maior potencial de redução. No entanto, uma característica positiva da técnica é a análise de riscos que é realizada durante a seleção de casos de teste. Esta análise aumenta a confiança na suíte selecionada, pois esta é especificada pelo testador e considera diversos aspectos de custos relacionados aos casos de teste e às faltas que estes podem revelar.

É possível observar que a técnica apresenta uma dependência operacional. Esta dependência é o testador, necessário para a configuração da técnica. Os autores da técnica argumentam que a etapa da análise de riscos é optativa, portanto, esta dependência não afeta significativamente a generalidade de T_2. Além disto, durante o estudo experimental, é verificado que o nível de experiência do testador não representa um fator significativo no desempenho da técnica. Ainda que não seja possível generalizar esta conclusão a respeito do nível de experiência do testador, a quantidade de execuções da técnica no experimento aumentam a credibilidade estatística da análise e, portanto, aumenta a confiança das conslusões.

7.8.3 Generalidade de *Weighted Similarity Approach for Regression Testing*

Assim como as demais técnicas analisadas no experimento, WSA-RT foi verificada com relação a sua generalidade. O primeiro aspecto observado é que a técnica não é executada em nenhum formato específico de modelo. WSA-RT investiga a suíte de testes de regressão com

uma suíte obtida a partir da versão delta do sistema. O algoritmo então compara os passos dos casos de teste identificando as similaridades e selecionando aqueles menos similares entre as duas suítes de teste. Portanto, a suíte pode ter sido gerada a partir de qualquer modelo. Desde que seja possível comparar os passos das duas suítes, a técnica pode ser executada.

Por sua vez, a natureza da análise realizada por WSA-RT possui dois aspectos. O primeiro é a análise de similaridade entre as duas suítes de teste, de forma que apenas os passos entre os casos de teste são verificados, e portanto, nenhuma análise de dependência, ou identificação de modificações, específica, é realizada pela técnica. O segundo aspecto analisado pela técnica é o perfil de uso especificado pelo testador. O modelo do sistema é configurado com valores de probabilidade que representam a probabilidade de um usuário executar os respectivos fluxos da ferramenta. Estes valores de probabilidade são verificados, de forma que os casos de teste que exercitam os fluxos com maiores valores de probabilidade são selecionados para a suíte reduzida.

Apesar de não realizar uma análise de dependência, a técnica realiza uma análise para lidar com remoções de funcionalidades. Ao identificar a remoção de uma funcionalidade, a técnica procura selecionar um caso de teste mais similar a um caso de teste da versão base que possua a funcionalidade removida. Dessa forma, a técnica seleciona casos de teste que cobrem mais estados que possam apresentar uma falta de regressão devido à remoção da funcionalidade [Korel et al. 2002]. Os demais tipos de modificações (adição e modificação de funcionalidades) não são tratadas de forma específica pela técnica, apesar desta selecionar casos de teste que exercitam funcionalidades adicionadas ou modificadas.

Assim como em T_2, WSA-RT apresenta uma dependência operacional, na qual é necessário um testador para realizar a configuração da técnica, ou seja, especificar o perfil de uso no modelo do sistema. A dependência, sob a perspectiva de WSA-RT é maior que em T_2 pois é verificado que o nível de experiência do testador está relacionado com o desempenho da técnica. Esta diferença foi considerada estatisticamente significativa, a partir dos testes realizados (estes testes estão descritos no Apêndice B).

Diante dos aspectos observados, podemos concluir que a técnica apresenta uma boa generalidade quanto aos aspectos do modelo, mas a natureza da análise, das modificações e as dependências da técnica prejudicam a generalidade da técnica. Dessa forma, é aconselhável o refinamento da análise da técnica, para que a sua generalidade compense as limitações na

dependência operacional, isto é a necessidade de um testador experiente para configurar a técnica.

7.8.4 Generalidade da Seleção baseada em *Clusters*

Algumas análises de generalidade foram realizadas acerca da técnica de seleção baseada em *clusters* (T_4) para o contexto de re-teste seletivo baseado em código [Rothermel and Harrold 1996]. Porém, neste estudo experimental, é investigada a generalidade da técnica para o contexto de especificação.

O primeiro aspecto analisado é o modelo utilizado pela técnica. Ao propor a técnica, os autores utilizaram um GFC para especificar os fluxos no código do sistema, utilizando uma perspectiva procedural. No contexto deste estudo experimental, os fluxos de execução, modelados na especificação da ferramenta, são estruturados como um GFC, permitindo a aplicação da técnica. Um processo semelhante é descrito por Chen et. al, durante a descrição da técnica de seleção baseada em análise de riscos [Chen et al. 2002].

Uma limitação no desempenho da técnica para o contexto de especificação é a criação de diversos *clusters*. É adequado que os *clusters* englobem os fluxos comuns da aplicação, e para o nível de especificação, diversos fluxos se dividiam em fluxos principais (fluxos comuns da aplicação) e fluxos alternativos (fluxos com exceções, e tratamento de erros realizados pela ferramenta especificada). Portanto, para o contexto de especificação o uso deste modelo pode caracterizar um risco na execução da técnica, pois o desempenho pode ser prejudicado quando os diversos fluxos que o sistema pode seguir se conectam.

Com relação à natureza da análise realizada pela técnica, foi verificado que T_4 utiliza os *clusters* no modelo para identificar modificações e a propagação destas modificações, o que pode caracterizar uma pequena análise de dependência. Sob esta perspectiva, a análise de dependência da técnica é restrita pelos aspectos estruturais do modelo, como vértices e transições do GFC. Ao identificar as modificações a técnica é capaz de identificar a natureza da modificação, ou seja, a técnica é capaz de identificar se uma funcionalidade é removida, adicionada ou modificada no modelo, porém o algoritmo não realiza um tratamento específico para cada um destes tipos de modificações.

A técnica não apresenta dependências operacionais ou ferramentais, o que contribui com a generalidade da técnica, e com a sua automatização. Apesar de não apresentar muito

esforço durante a execução, a técnica necessita de muito esforço de implementação. O algoritmo para identificar e criar os *clusters*, assim como o caminhamento nos GFC da versão base e versão delta do sistema é difícil e necessita de muito tempo para ser implementado. No entanto, uma vez automatizada a execução da técnica é rápida e simples, o que contribui para a sua generalidade.

7.8.5 Generalidade da Seleção Aleatória de Casos de Teste

Diversos aspectos dificultam uma análise precisa da generalidade da técnica de seleção aleatória de casos de teste (T_5). A sua natureza aleatória impede a caracterização de sua análise no contexto de teste de regressão, pois as modificações não são observadas, assim como nenhuma análise de dependência é realizada pela técnica. Sob a perspectiva de técnicas para seleção de casos de teste no contexto geral do teste de software, esta técnica apresenta uma boa generalidade, pois pode ser executada em qualquer contexto. Porém, a confiabilidade da suíte resultante é prejudicada pela ausência de critérios que caracterizem a seleção de casos de teste.

Os aspectos do modelo necessários para a execução desta técnica contribuem para sua generalidade em teste, uma vez que não é necessário gerar os casos de teste a partir de um formato de modelo específico, pois a técnica executa, apenas, na suíte de testes de regressão. Dentre as dependências da técnica se destacam, o percentual de cobertura, que estabelece o critério de parada do algoritmo, e a utilização do gerador de números aleatórios presente no ambiente em que a técnica é executada. Sob a perspectiva de execução, é necessário que o testador insira o percentual de cobertura desejado, uma vez que a técnica não executa independente desta configuração.

Considerando, a implementação da técnica, o desempenho obtido nesta possui uma dependência com o componente de implementação que define os números aleatórios utilizados na seleção dos casos de teste. Um gerador de números aleatórios utiliza seqüências de números (e.g. bits) para determinar um número aleatório. Diante disto, é recomendável que a técnica utilize um gerador de números aleatórios capaz de gerar uma grande quantidade de seqüências diferentes.

7.9 Ameaças à Validade

Nesta seção são discutidas as ameaças à validade encontradas durante a execução e análise do experimento. Esta discussão envolve as ameaças discutidas no Capítulo 5, e outras ameaças que não foram levantadas durante o planejamento e foram, porém, identificadas durante a etapa operacional e de análise do experimento. Esta discussão é importante para fornecer a visão de como melhorar o estudo experimental. Ou seja, a partir dessa discussão, é possível fornecer informações a outros pesquisadores que desejam reproduzir o estudo experimental, procurando avaliá-lo ou expandi-lo.

De acordo com o que foi discutido ao levantar as ameaças, durante o planejamento, a utilização de apenas um objeto se mostrou uma ameaça à validade dos resultados do experimento. Durante a execução, foram observados que alguns aspectos (algumas junções de fluxos e laços) do modelo afetam o tempo de execução de algumas técnicas. Além disso, estes aspectos estruturais fazem parte da análise de dependência, e na identificação das mudanças no modelo, quando comparamos duas versões de uma especificação. Esta ameaça foi encontrada durante a fase de instrumentação da etapa Operacional, ao realizar as primeiras execuções e os testes de implementação das técnicas. A familiaridade do investigador com relação ao objeto facilitou o tratamento desta ameaça.

Outro aspecto que deve ser observado na obtenção dos resultados é a utilização do modelo de faltas. Como explicado no Capítulo 6, um modelo de faltas pode ser utilizado para estabelecer faltas em passos de um modelo ou código, quando é inviável a execução dos casos de teste, ou a análise no código e/ou especificação para identificar as faltas reais [Binder 1999]. Essas faltas, por sua vez, são utilizadas para identificar os casos de teste que revelam faltas de regressão, e portanto definem os resultados de inclusão, precisão e densidade de faltas. Diante disto, ao utilizar um modelo de faltas, é importante que as faltas sejam cuidadosamente posicionadas no modelo, para evitar um viés nos resultados.

Sob esta perspectiva o modelo de faltas pode ser considerado uma ameaça à validade. Para tratar esta ameaça, as faltas são posicionadas de acordo com um histórico de faltas da própria ferramenta. Estas faltas foram encontradas durante um processo de reengenharia [Oliveira Neto and Machado 2008]. Portanto, uma vez que as faltas estão relacionadas com elementos anteriores à definição do experimento, e elas caracterizam faltas devido a um pro-

cesso de modificação, o risco de que o modelo de faltas caracterize um viés do experimento é minimizado.

Durante a implementação das técnicas, na etapa de instrumentação, observamos uma grande quantidade de casos de testes gerados automaticamente pela máquina de estados de T_1. Esta quantidade é grande, pois as modificações por remoções adicionam auto-laços no modelo. Esta característica aumenta bastante a quantidade de caminhos na máquina de estados. É comum, em Teste Baseado em Modelos, obter uma suíte de testes muito grande, quando esta é gerada automaticamente a partir de um modelo, como uma máquina de estados [Fraser and Wotawa 2007].

Enquanto que os demais modelos geram 65 casos de teste, a máquina de estados de T_1, após as modificações, gera 8079 casos de teste. A maioria destes 8079 casos de teste cobrem as mesmas transições, porém em ordens diferentes, devido aos auto-laços no modelo. Sob esta perspectiva não seria justo comparar esta suíte de testes de regressão com as demais, pois as demais suítes de regressão (utilizadas pelas outras técnicas) não apresentam tanta redundância quanto a suíte de testes da máquina de estados.

Portanto, para evitar um viés no experimento, a suíte de 8079 casos de teste foi reduzida para uma suíte com 68 casos de teste, removendo cerca de 99% desta suíte. Mesmo que a redução seja muito grande, foi utilizada uma técnica automática de seleção de casos de teste baseada em similaridade [Cartaxo et al. 2009]. Através desta técnica, os casos de testes redundantes (i.e. que cobrem os mesmos passos) foram removidos. Inicialmente, foi levantada a preocupação desta estratégia caracterizar uma ameaça à validade interna do experimento. Porém, é importante lembrar que o processo de geração de casos de teste não faz parte da técnica, uma vez que os próprios autores da técnica declaram que qualquer algoritmo de geração pode ser utilizado na máquina de estados [Korel et al. 2002, Chen et al. 2007].

Uma das ameaças identificadas durante o planejamento, está relacionada com a análise estatística realizada. De acordo com o design experimental, o método adequado para analisar os resultados de um experimento de fator único e diversos níveis é ANOVA, que fornece uma forte credibilidade estatística para a análise [Jain 1991]. No entanto, os dados obtidos não respeitam as premissas necessárias para aplicar ANOVA.

Isto é observado devido à natureza determinística de alguns algoritmos. Dessa forma, ao fornecer uma entrada, a técnica fornece uma saída que, em diversos casos, se repete mesmo

com várias replicações do experimento. Uma vez que é fornecida apenas uma saída, os dados não apresentam variância. Dessa forma, não é observado uma distribuição específica dos dados (e.g. distribuição normal), o que inviabiliza a utilização de testes paramétricos, como por exemplo, ANOVA ou teste t de Student. Diante disto, é necessário utilizar o teste não-paramétrico correspondente a ANOVA, ou seja, o teste de Kruskal-Wallis.

Por ser um teste não-paramétrico, o Kruskal-Wallis não possui o poder estatístico que ANOVA possui, sendo considerado, por alguns autores, como um teste de fraco poder estatístico. Siegel e Castellan Junior, e Kvam e Vidakovic, [Siegel and Junior 1988, Kvam and Vidakovic 2007] argumentam que os pesquisadores não deviam considerar os testes não-paramétricos como 'fracos', pois muitas vezes (em especial nas ciências humanas, e biológicas) estes testes são os principais recursos para a realização de análise.

A utilização de um teste paramétrico (e.g. ANOVA) sem a investigação adequada de suas premissas é caracterizado como um risco maior que a utilização dos testes não-paramétricos [Siegel and Junior 1988, Wohlin et al. 2000]. O teste é fundamentado em suas premissas, e usá-lo indiscriminadamente prejudica, significativamente, a credibilidade da análise realizada. Diante disto, a análise das premissas de ANOVA foi cuidadosamente realizada neste estudo experimental e é descrita nos apêndices através dos testes de normalidade dos dados (Apêndice A) e pelos testes das próprias premissas de ANOVA (Apêndice C).

As medidas estabelecidas para tratar as ameaças encontradas neste estudo experimental estão descritas nesta seção e na Seção 5.11 do Capítulo de Definição e Planejamento. É importante descrever as ameaças em um estudo experimental, para que os pesquisadores interessados em utilizar, reproduzir ou modificar, os resultados do experimento, para sua própria pesquisa, estejam cientes das condições encontradas para a obtenção dos resultados. Portanto, estas ameaças caracterizam também aspectos de trabalhos futuros do experimento, em que estas ameaças podem ser eliminadas, por exemplo, inserindo mais fatores, reduzindo a quantidade de níveis, ou utilizando mais objetos e sujeitos, para que novos resultados sejam obtidos e comparados com os apresentados neste trabalho.

7.10 Considerações Finais do Capítulo

Os aspectos discutidos neste capítulo contemplam os resultados obtidos e a análise realizada no estudo experimental. Algumas análises estatísticas realizadas na etapa de análise são apresentadas nos Apêndices. Utilizando esta estratégia, este capítulo apresenta um aspecto mais experimental do que estatístico, permitindo que o leitor possa focar nas conclusões obtidas a partir da análise estatística. Os elementos da análise estão relacionados com elementos do planejamento (e.g. ameaças à validade, projetos experimentais, dentre outros) apresentados no Capítulo 5.

A partir da execução das técnicas, e dos respectivos dados, foram obtidas as características das técnicas, como as limitações (apresentadas nos elementos de generalidade da técnica), vantagens e desvantagens de sua utilização. Uma das principais verificações realizadas na análise do experimento diz respeito às premissas necessárias para a realização de ANOVA (Apêndice C). Com esta verificação evitamos a utilização indevida do teste, o que poderia comprometer, significativamente, as conclusões atingidas. Os resultados do teste de hipóteses de cada variável dependente são resumidos na Tabela 7.7.

Tabela 7.7: Resumo dos resultados dos testes de hipóteses.

Hipóteses	$H0_1$	$H0_2$	$H0_3$	$H0_4$	$H0_5$
$p - valor$	$0,0001$	$0,0001$	$0,0001$	$0,0001$	$0,0001$
status	rejeitar	rejeitar	rejeitar	rejeitar	rejeitar

Além dos resultados dos testes de hipóteses, foi realizada uma análise adicional para caracterizar o desempenho comparativo das técnicas, uma vez que o resultado do teste de hipóteses fornece apenas a perspectiva se as técnicas são semelhantes, ou diferentes. Os aspectos estatísticos desta análise podem ser encontrados no Apêndice D. Um resumo destes resultados são apresentados na Tabela 7.8.

A técnica de seleção aleatória (T_5) apresenta um bom desempenho, considerando alguns de seus valores de cobertura (e.g. inclusão de $T_{5-75\%}$, potencial de redução de $T_{5-25\%}$). No entanto, é importante lembrar que esta técnica não apresenta um critério para a seleção, e os valores de cobertura são determinados pelo testador. Portanto, apesar de apresentar bons resultados, a técnica não é confiável, pois estes resultados são consequências de diversas

Tabela 7.8: Resultado das análises de desempenho realizada para cada variável dependente.

Variável Dependente	**Desempenho observado**
Inclusão	$T_2 > T_4 > T_{5-75\%} > T_1 = T_{5-50\%} > T_{3e} > T_{3i} > T_{5-25\%}$.
Precisão	$T_2 > T_{5-75\%} > T_{5-50\%} > T_4 > T_{3i} > T_1 > T_{3e} = T_{5-25\%}$
Eficiência	$T_{5-75\%} > T_{5-50\%} > T_{5-25\%} = T_4 > T_{3i} > T_{3e} > T_{2e} > T_{2i} > T_1$
Potencial de Redução	$T_{5-25\%} > T_{3i} > T_{3e} > T_1 > T_{5-50\%} > T_4 > T_{5-75\%} > T_2$
Densidade de Faltas	$T_4 > T_3 > T_1 = T_5 > T_2$

repetições da execução da técnica. O principal objetivo na utilização da técnica de seleção aleatória é comparar o desempenho das demais técnicas, com um processo de seleção em que não há um critério específico (seleção aleatória), e que reflete uma redução grande, média e pequena (25%, 50% e 75% de cobertura, respectivamente) da suíte de testes de regressão.

No capítulo seguinte são apresentadas propostas de trabalhos futuros e as considerações finais. As propostas de trabalhos futuros envolvem aspectos para a reprodução deste estudo experimental, destacando estratégias para lidar com as ameaças encontradas. Considerando que novas técnicas foram propostas durante a realização deste trabalho [Naslavsky et al. 2009, Subramaniam et al. 2009], o conteúdo deste capítulo fornece elementos que facilitam a análise destas técnicas, e demais técnicas que venham a ser propostas.

Capítulo 8

Considerações Finais

Neste trabalho nós realizamos uma análise e investigação experimental de 5 técnicas de re-teste seletivo baseado em especificação. A partir da análise estatística, e utilizando um nível de confiança de 95%, rejeitamos a hipótese de que todas as técnicas apresentam comportamentos semelhantes de inclusão, precisão, eficiência, potencial de redução e densidade de faltas. Dessa forma, foi possível confirmar a hipótese geral estabelecida para este trabalho: "*As técnicas de re-teste seletivo, analisadas neste experimento, são diferentes quanto à inclusão, precisão, eficiência, potencial de redução e densidade de faltas*".

Através da análise comparativa entre as técnicas conseguimos identificar as técnicas com os melhores desempenhos para as diferentes propriedades analisadas (inclusão, precisão, eficiência, potencial de redução e densidade de faltas). A investigação realizada em cada técnica revelou informações a respeito das vantagens, desvantagens e generalidade das técnicas. Estas informações facilitam a escolha de uma técnica específica em um processo de teste de regressão.

Dentre os aspectos investigados, foi verificado que a técnica de seleção baseada em análise de risco, apresenta uma alta inclusão e precisão da técnica devido ao baixo potencial de redução da técnica. Este resultado é comprovado pelos dados de densidade de faltas da técnica. Considerando a técnica de seleção baseada em *clusters*, verificamos que a precisão da técnica é afetada pelo tamanho e quantidade dos *clusters*. Foi observado que os *clusters* com muitas transições tendem a diminuir significativamente a precisão da técnica.

Além disto, verificamos também que o nível de experiência do testador afeta o desempenho de WSA-RT para as variáveis de inclusão, precisão, eficiência e potencial de redução.

Por outro lado, o nível de experiência do testador afeta apenas a eficiência da técnica de seleção baseada em análise de riscos. Dessa forma, quando utilizada em um cenário real, não é necessário alocar um testador com muita experiência, para configurar essa técnica.

Também foi possível observar o desempenho comparativo entre as técnicas. Por exemplo, podemos observar que as técnicas que utilizam apenas o modelo da especificação, para identificar as modificações (por exemplo, a seleção baseada em *clusters*), executam mais rápido que técnicas que realizam a seleção através de operações em matrizes (e.g. WSA-RT e a seleção baseada em análise de riscos). Essas e as demais conclusões obtidas, assim como, as considerações a respeito da generalidade e aplicabilidade das técnicas no processo de teste de software, são encontrados no Capítulo 7.

Além do estudo experimental, também foi proposta uma técnica de re-teste seletivo baseado em especificação. O principal objetivo ao desenvolver esta técnica, WSA-RT, foi icorporar elementos de uma abordagem baseada em valores no processo de seleção dos casos de teste de regressão. Diante disto, WSA-RT é capaz de selecionar os casos de teste que cobrem as modificações realizadas, assim como, os casos de teste considerados como "importantes" (i.e. que apresentam um alto valor) para o usuário da aplicação.

Ao analisar o desempenho de WSA-RT no estudo experimental, observamos uma redução de cerca de 70% da suíte de testes de regressão, de forma que, 50% desta suíte, é composta por casos de teste que capturam faltas de regressão. Estes resultados foram observados através da densidade de faltas da técnica. Além disto, a técnica proposta, e as demais utilizadas neste trabalho, estão implementadas na ferramenta LTS-BT. Através da ferramenta, é possível executar automaticamente a geração e a seleção da suíte de testes de regressão utilizando algumas das técnicas utilizadas no estudo experimental.

Uma vez que os casos de teste, em WSA-RT, são gerados a partir do modelo da versão delta do software, a suíte de regressão não apresenta casos de teste obsoletos. Além disto, a utilização do perfil de uso na seleção, também permite que as funcionalidades mais críticas, para o usuário, sejam testadas. Em alguns casos, WSA-RT seleciona casos de teste que não exercitam trechos modificados do modelo (e.g. casos de testes muito similares aos casos de teste obsoletos). Esta característica é importante, pois estes trechos podem apresentar as faltas de regressão por efeito colateral.

Apesar das diversas vantagens, WSA-RT não exibiu bons resultados de inclusão e pre-

cisão da técnica. Além disto, foram selecionadas poucas faltas de regressão, ou seja, muitos casos de teste da suíte selecionada cobriam as mesmas faltas. Dessa forma, é necessário melhorar o algoritmo de seleção da técnica.

Uma das principais dificuldades encontradas foi a etapa de instrumentação do estudo experimental. A maior parte do tempo do trabalho foi utilizada para esta etapa, em especial na implementação das técnicas, e na modelagem da especificação (das duas versões e nos diversos formatos de entrada). Apesar das dificuldades, o suporte ferramental ao re-teste seletivo baseado em especificação é uma das principais contribuições do trabalho.

Através da ferramenta LTS-BT, é possível gerar e selecionar, automaticamente, os casos de teste de regressão, a partir de diversos formatos de modelo (GFC, STR, diagramas de atividades, diagramas de máquinas de estado). Além disto, o código da implementação foi desenvolvido com o objetivo de facilitar a adição de outras técnicas, de re-teste seletivo, à ferramenta.

8.1 Trabalhos relacionados

Diversos trabalhos com investigação experimental têm sido propostos na literatura de teste de regressão [Rothermel and Harrold 1996, Graves et al. 1998, Graves et al. 2001, Kim et al. 2000, Elbaum et al. 2002, Leon and Podgurski 2003, Do and Rothermel 2006, Do et al. 2008]. Porém, a maioria deles abordam técnicas de priorização de casos de teste. Técnicas de priorização procuram ordenar a execução dos casos de teste de regressão a partir de algum critério específico (e.g. cobertura de faltas, quantidade de funcionalidades cobertas, dentre outros), de forma que a redução no tamanho da suíte de regressão não é observada.

As técnicas consideradas em nosso estudo experimental procuram reduzir a quantidade de casos de teste que devem ser executados, caracterizando a redução no tamanho da suíte de regressão. Sob esta perspectiva, os estudos experimentais de técnicas de priorização não foram considerados neste trabalho, pois estes investigam perspectivas diferentes das que desejamos observar.

Grande parte das investigações experimentais realizadas em técnicas de re-teste seletivo se concentram no contexto baseado em código [Korel et al. 2002]. Dentre eles, Rothermel e Harrold apresentam um *framework* para a análise de técnicas de re-teste seletivo baseado

em código [Rothermel and Harrold 1996]. O principal elemento deste *framework*, utilizado no nosso estudo experimental, são as propriedades (inclusão, precisão, eficiência e generalidade) analisadas nesse *framework*. Enquanto que Rothermel e Harrold observam elementos do código (e.g. chamadas de funções e procedimentos), nós observamos elementos da especificação (casos de uso, e cenários da aplicação).

Apesar de serem propostas para a análise de técnicas baseadas em código, essas propriedades são utilizadas em outros trabalhos realizados no contexto de especificação. Em seu livro, Binder descreve aspectos de inclusão, precisão, eficiência e generalidade de algumas técnicas, tanto para o contexto de código, como de especificação [Binder 1999].

Em seu trabalho, Mahdian et al. realizam um *survey* sobre o teste de regressão realizado com diagramas UML [Mahdian et al. 2009]. Um dos principais aspectos abordados no trabalho são as técnicas de re-teste seletivo que utilizam modelos UML para a seleção de casos de teste. São apresentadas técnicas que utilizam diagramas de classes, casos de uso, diagramas de seqüência, dentre outros modelos. Dessa forma, são avaliadas técnicas tanto para o contexto de código como o de especificação. Assim como no nosso estudo, essas técnicas são avaliadas utilizando as mesmas propriedades (inclusão, precisão, eficiência e generalidade).

Outras propriedades, também avaliadas por Mahdian et al., são a cobertura e a segurança das técnicas [Mahdian et al. 2009]. Ou seja, é verificado se a cobertura realizada pela técnica satisfaz os requisitos de testes especificados (e.g. modificações, funcionalidades críticas, dentre outros), e se as técnicas são seguras (i.e. apresentam 100% de inclusão). A segurança não é considerada neste estudo experimental, pois não foi observada 100% de inclusão em nenhuma das técnicas analisadas.

Assim como a generalidade, a propriedade de cobertura é qualitativa, dificultando a análise estatística de seus resultados. A propriedade de cobertura não é considerada em nosso estudo, pois, é indicada para a análise de técnicas de redução de suítes de teste, onde o principal objetivo é a cobertura dos requisitos de testes, e não a redução no tamanho da suíte de regressão. Apesar de avaliar mais propriedades, Mahdian et al. não realizam uma investigação experimental para obter conclusões estatisticamente significativas a respeito das técnicas avaliadas em seu *survey*.

Graves et al. realizam um estudo empírico de técnicas de re-teste seletivo baseado em

código [Graves et al. 2001]. O estudo é realizado com técnicas seguras, de minimização, e de seleção aleatória. Os autores utilizam 9 aplicações reais como objetos do experimento, e as variáveis dependentes analisadas são a eficiência e a detecção de faltas de cada técnica. Assim como Graves et al., nosso experimento utiliza uma técnica de seleção aleatória e uma aplicação real como objeto. Porém, a principal diferença entre os trabalho é o contexto, pois nós observamos elementos da especificação (casos de uso, cenários de execução, e requisitos), enquanto que Graves et al. observam faltas de regressão presentes no código das aplicações.

Enquanto que os estudos experimentais sobre re-teste seletivo baseado em especificação ainda são poucos, a quantidade de técnicas propostas têm aumentado [Engström et al. 2008]. Grande parte dos trabalhos propostos para abordagens baseadas em modelos ainda estão relacionados com o nível de código ou de componentes do software, através de, por exemplo, diagramas de classes e diagramas de sequências. Abaixo, descrevemos algumas destas técnicas e as comparamos com WSA-RT.

A técnica de teste baseada em *firewall* utiliza diagramas de classes para realizar a seleção dos casos de teste de regressão [Binder 1999]. A partir das modificações, são identificadas as dependências e relacionamentos (heranças, composições, dentre outros) das entidades afetadas. Após identificar as modificações, são estabelecidos alguns perímetros, denominados *firewall*, com o objetivo de determinar a propagação destas mudanças nos elementos do modelo.

Para determinar os *firewalls*, a técnica investiga os pares de classes. Duas classes podem estar relacionadas por modificações no contrato ou na implementação, onde o primeiro considera a visão externa de uma classe, como a disponibilidade de métodos, argumentos e assinaturas. As modificações de implementação consideram aspectos internos à classe, ou seja, os elementos dentro do encapsulamento e escopo de classe, como por exemplo, os métodos privados e os trechos de códigos dentro dos métodos (quando considerando também o nível de código).

Ao contrário de WSA-RT, a técnica de *firewall* é utilizada para um contexto de especificação e código, pois são utilizados elementos da especificação (relacionamentos entre as classes, ou pacotes) relacionados com o código [Leung and White 1990]. Por outro lado, os casos de teste de regressão ainda são especificados a partir do código, e não a partir dos

requisitos, ou casos de uso da aplicação. Dessa forma, o custo e esforço de identificar as modificações, nos casos de teste, a partir das mudanças nos requisitos é maior.

Subramaniam et al. propõem uma técnica de re-teste seletivo que utiliza provadores de teoremas para identificar transições afetadas pelas modificações em Máquinas de Estados Finitas e Estendidas (MEFE) [Subramaniam et al. 2009]. Este trabalho provê um suporte maior à execução da MEFE, pois apresenta suporte a um rico conjunto de tipos de dados (e.g. booleanos, arrays, filas, números e registros). Diante disto, os autores procuram preencher a lacuna existente entre a facilidade de execução das abordagens baseadas em código e o nível de abstração das abordagens baseadas em especificação.

Diferentemente de TBM, em que os casos de testes são expressos através de uma sequência de transições do modelo, a técnica de Subramaniam et al. representa os casos de teste através de sequências de dados de entrada (atribuídos às variáveis no modelo), saídas esperadas e um veredito. Esta característica representa a principal diferença entre a técnica e WSA-RT, pois não consideramos variáveis e dados da especificação durante a seleção, exceto pelas informações do perfil de uso. No entanto, WSA-RT não é dependente da utilização de uma MEFE. Diante disto, diversos formatos de modelos (e.g. MEFE, STR, GFC, dentre outros) podem ser utilizados para a geração das suítes de teste de regressão em WSA-RT.

Naslavsky et al. apresentam uma técnica de re-teste seletivo baseada em modelos, para realizar a seleção de casos de teste de regressão para testar a implementação, observando as entidades modificadas. Para isto, são utilizados diagramas de classes e diagramas de sequências UML [Naslavsky et al. 2009]. As modificações são identificadas a partir de comparações realizadas entre os modelos de diferentes versões. Ao contrário de WSA-RT, a técnica não identifica as modificações dos casos de uso e cenários da aplicação, apenas nos métodos, classes e demais elementos do código da aplicação.

Algumas técnicas de re-teste seletivo baseado em especificação, são aplicadas ao nível de componentes de um software [Sajeev and Wibowo 2003, Mao and Lu 2005]. Estas técnicas observam as interfaces e as respectivas interações entre os componentes, com o objetivo de capturar as faltas de regressão por efeito colateral das modificações. Estas técnicas não foram utilizadas no estudo experimental, pois foi considerado apenas o nível de sistemas (e.g. casos de uso e cenários da aplicação).

Um resumo dos trabalhos relacionados acerca de estudos experimentais na área, e as

técnicas de re-teste seletivo, são apresentados nas Tabelas 8.1 e 8.2, respectivamente.

Tabela 8.1: Resumo dos trabalhos relacionados envolvendo estudos experimentais.

Referência Bibliográfica	**Tipo de Estudo**	**Objetos de Estudo**
[Rothermel and Harrold 1996]	Análise Experimental	Técnicas de re-teste seletivo baseado em código
[Mahdian et al. 2009]	Survey	Técnicas de re-teste seletivo que usam modelos UML
[Graves et al. 1998]	Estudo Empírico	Grupos de técnicas de re-teste seletivo
[Kim et al. 2000]	Estudo Empírico	Técnicas de priorização de casos de teste
[Do and Rothermel 2006]	Estudo Empírico	Técnicas de priorização de casos de teste
[Do et al. 2008]	Estudo Empírico	Técnicas de priorização de casos de teste
[Elbaum et al. 2002]	Revisão Sistemática	Técnicas de priorização de casos de teste

Tabela 8.2: Resumo dos trabalhos relacionados envolvendo técnicas de re-teste seletivo.

Referência Bibliográfica	**Semelhanças com WSA-RT**	**Diferenças com WSA-RT**
[Binder 1999]	Utiliza modelos; Identifica modificações	Analisa métodos e atributos do código; Realiza análise de dependência
[Subramaniam et al. 2009]	Utiliza modelos	Necessita de uma MEFE ; Casos de teste são sequências de dados de entrada
[Naslavsky et al. 2009]	Utiliza modelos; Identifica modificações; Casos de teste são caminhos do modelo	Necessita de diagramas de classe e sequência em UML; Utiliza classes e métodos do código
[Sajeev and Wibowo 2003]	Utiliza modelos	Considera apenas o nível de componente de um software.
[Mao and Lu 2005]	Utiliza modelos	Considera apenas o nível de componente de um software.

Como podemos observar, diversas técnicas de re-teste seletivo têm sido proposta na literatura. No entanto, apesar da utilização de modelos, o código ainda é muito abordado pelas técnicas, devido à facilidade de execução dos casos de teste. Além disto, a maior parte dos estudos empíricos realizados, com as técnicas existentes na literatura, são estudos de casos

[Engström et al. 2008]. Dessa forma, é necessária a realização de mais estudos experimentais procurando observar e comparar o desempenho das técnicas propostas na área.

8.2 Trabalhos futuros

Os nossos trabalhos futuros são: a realização de mais estudos experimentais, e o melhoramento da técnica WSA-RT. A seguir descrevemos como planejamos alcançar esses objetivos.

8.2.1 Realização de mais estudos experimentais

O principal objetivo na realização de mais estudos experimentais com as técnicas, é lidar com as limitações encontradas. A primeira destas limitações é a quantidade de objetos utilizados. É necessário utilizar aplicações com diferentes modelos de especificação (e.g. diagramas de atividades, casos de uso, dentre outros) para possibilitar a generalização do comportamento das técnicas.

Ainda sob esta perspectiva, é importante utilizar os históricos de faltas de regressão das aplicações utilizadas no experimento. Quando não houver disponibilidade deste histórico, é recomendado a utilização de um, ou mais, modelos de faltas para cada aplicação. Dessa forma, é possível obter uma maior variância nos resultados de inclusão, precisão e densidade de faltas. Com esta variância, é possível encontrar uma distribuição normal, e portanto, utilizar os testes paramétricos para a análise estatística das hipóteses.

A variável de eficiência, considerada neste estudo, investiga apenas o tempo de execução de cada técnica analisada. Outros modelos, que consideram o tempo de execução da suíte selecionada, podem ser utilizados para obter resultados mais precisos dessa variável. Uma das maiores dificuldades de utilização destes modelos, é o tempo necessário para executar os casos de teste gerados a partir da especificação. Na maioria das vezes, estes casos de teste não são executáveis automaticamente, e portanto, devem ser executados manualmente.

Além da modificação no modelo de eficiência, é desejado analisar mais variáveis dependentes. Uma variável amplamente utilizada em técnicas de priorização de casos de teste é a Porcentagem Média de Faltas Detectadas (*Average Percentage of Faults Detected* - APFD). Esta variável fornece a perspectiva das faltas de regressão descobertas durante a execução das suítes de testes. Apesar de ser utilizada no contexto de priorização, é necessário investi-

gar se é possível utilizá-la na análise de técnicas de re-teste seletivo, sem causar um viés nos resultados (e.g. erros de medições).

8.2.2 Melhoramento de WSA-RT

A análise estatística realizada, mostra que WSA-RT apresenta bons resultados de potencial de redução e densidade de faltas. No entanto, o seu algoritmo deve ser melhorado, com o objetivo de aumentar a inclusão e precisão da técnica, uma vez que esta apresentou os piores desempenhos nestas propriedades. O principal aspecto que deve ser tratado é a redundância observada na suíte selecionada.

Um dos elementos da execução de WSA-RT, é a seleção de casos de teste da versão delta, menos similares, aos casos de teste da versão base. Nos resultados do experimento, observamos que a suíte selecionada apresenta muitos casos de teste (da versão delta) similares entre si. Dessa forma, é necessário modificar a seleção, e o cálculo dos valores da matriz, com o objetivo de incorporar a similaridade entre os próprios casos de teste da versão delta.

Além disto devem ser investigados os elementos de implementação, em especial, a construção e busca na matriz de similaridades. Dessa forma, é possível melhorar a eficiência da técnica, pois os resultados mostram que grande parte do processamento de WSA-RT está nas operações com a matriz. Também desejamos investigar as características um bom perfil de uso, para estabelecer diretrizes, para a especificação dos valores de probabilidade.

Os resultados obtidos pelas demais técnicas investigadas no estudo experimental também podem contribuir para o melhoramento de WSA-RT. Esta contribuição se dá através da combinação de elementos das demais técnicas para o processo de seleção de WSA-RT. Um exemplo seria incorporar a seleção baseada em *clusters* para identificar *clusters* entre casos de teste de regressão e incorporar os aspectos da função de similaridade (Equação 4.1) para selecionar casos de teste menos similares entre diferentes *clusters*. No entanto, é necessário investigar esta e demais possíveis combinações de técnicas através de outro estudo experimental, para verificar se há melhora no desempenho de WSA-RT.

Referências Bibliográficas

[Agrawal et al. 1993] Agrawal, H., Horgan, J. R., Krauser, E. W., and London, S. (1993). Incremental regression testing. In *ICSM '93: Proceedings of the Conference on Software Maintenance*, pages 348–357, Washington, DC, USA. IEEE Computer Society.

[Amland 2000] Amland, S. (2000). Risk-based testing: risk analysis fundamentals and metrics for software testing including a financial application case study. *J. Syst. Softw.*, 53(3):287–295.

[Babbie 1990] Babbie, E. R. (1990). *Survey research methods*. Wadsworth.

[Barbosa et al. 2004] Barbosa, D. L., Andrade, W. L., Machado, P. D. L., and Abrantes, J. C. F. (2004). Spaces - uma ferramenta para teste funcional de componentes. In *XI Sessão de Ferramentas - SBES 2004*, pages 55–61, Porto Alegre, Brasil. Sociedade Brasileira de Computação.

[Barbosa et al. 2007] Barbosa, D. L., Lima, H. S., Machado, P. D. L., de Figueiredo, J. C. A., Jucá, M. A., and de L. Andrade, W. (2007). Automating functional testing of components from uml specifications. *International Journal of Software Engineering and Knowledge Engineering*, 17(3):339–358.

[Beizer 1990] Beizer, B. (1990). *Software testing techniques (2nd ed.)*. Van Nostrand Reinhold Co., New York, NY, USA.

[Beizer 1995] Beizer, B. (1995). *Black-box testing: techniques for functional testing of software and systems*. John Wiley & Sons, Inc., New York, NY, USA.

[Bertolino et al. 2008] Bertolino, A., Cartaxo, E., Machado, P., and Marchetti, E. (2008). Weighting influence of user behavior in software validation. In *19th International Con-*

ference on Database and Expert Systems Application - DEXA 2008 Workshops, pages 495–500. IEEE Computer Society.

[Binder 1999] Binder, R. V. (1999). *Testing object-oriented systems: models, patterns, and tools*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Bo 2005] Bo, X. (2005). Requirement-based regression test suite reduction using dependence analysis. Master's thesis, University of Ottawa.

[Boehm 2006] Boehm, B. (2006). Value-based software engineering: Overview and agenda. In *In book*, pages 3–14. Springer Verlag.

[Boehm 1981] Boehm, B. W. (1981). *Software Engineering Economics*. Prentice Hall.

[Briand et al. 2009] Briand, L., Labiche, Y., and He, S. (2009). Automating regression test selection based on uml designs. *Information and Software Technology*, 51(1):16 – 30. Special Section - Most Cited Articles in 2002 and Regular Research Papers.

[Briand and Labiche 2001] Briand, L. C. and Labiche, Y. (2001). A uml-based approach to system testing. In *Proceedings of the 4th International Conference on The Unified Modeling Language, Modeling Languages, Concepts, and Tools*, pages 194–208, London, UK. Springer-Verlag.

[Cabral and Sampaio 2008] Cabral, G. and Sampaio, A. (2008). Formal specification generation from requirement documents. *Electron. Notes Theor. Comput. Sci.*, 195:171–188.

[Cartaxo et al. 2008] Cartaxo, E. G., Andrade, W. L., Neto, F. G. O., and Machado, P. D. L. (2008). LTS-BT: a tool to generate and select functional test cases for embedded systems. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 1540–1544, New York, NY, USA. ACM.

[Cartaxo et al. 2007] Cartaxo, E. G., de Oliveira Neto, F. G., and Machado, P. D. L. (2007). Automated test case selection based on a similarity function. In *Proceedings of MOTES07 - Model-based Testing - Workshop in conjunction with the 37th Annual Congress of the Gesellschaft fuer Informatik*, volume 110 of *Lecture Notes in Informatics (LNI)*, pages 381–386.

[Cartaxo et al. 2009] Cartaxo, E. G., Machado, P. D. L., and Neto, F. G. O. (2009). On the use of a similarity function for test case selection in the context of model-based testing. *STVR Journal of Software Testing, Verification, and Reliability*.

[Chen et al. 2002] Chen, Y., Probert, R. L., and Sims, D. P. (2002). Specification-based regression test selection with risk analysis. In *CASCON '02: Proceedings of the 2002 conference of the Centre for Advanced Studies on Collaborative research*. IBM Press.

[Chen et al. 2007] Chen, Y., Probert, R. L., and Ural, H. (2007). Regression test suite reduction using extended dependence analysis. In *SOQUA '07: Fourth international workshop on Software quality assurance*, pages 62–69, New York, NY, USA. ACM.

[Cook and Campbell 1979] Cook, T. D. and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin Company.

[Dalal et al. 1999] Dalal, S. R., Jain, A., Karunanithi, N., Leaton, J. M., Lott, C. M., Patton, G. C., and Horowitz, B. M. (1999). Model-based testing in practice. In *ICSE '99: Proceedings of the 21st international conference on Software engineering*, pages 285–294, New York, NY, USA. ACM.

[Do et al. 2005] Do, H., Elbaum, S., and Rothermel, G. (2005). Supporting controlled experimentation with testing techniques: An infrastructure and its potential impact. *Empirical Software Engineering*, 10:405–435. 10.1007/s10664-005-3861-2.

[Do et al. 2008] Do, H., Mirarab, S., Tahvildari, L., and Rothermel, G. (2008). An empirical study of the effect of time constraints on the cost-benefits of regression testing. In *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering*, SIGSOFT '08/FSE-16, pages 71–82, New York, NY, USA. ACM.

[Do and Rothermel 2006] Do, H. and Rothermel, G. (2006). An empirical study of regression testing techniques incorporating context and lifetime factors and improved cost-benefit models. In *Proceedings of the 14th ACM SIGSOFT international symposium on Foundations of software engineering*, SIGSOFT '06/FSE-14, pages 141–151, New York, NY, USA. ACM.

[El-Far 2001] El-Far, I. K. (2001). Enjoying the perks of model-based testing. In *In Proceedings of the Software Testing, Analysis, and Review Conference*.

[Elbaum et al. 2002] Elbaum, S., Malishevsky, A., and Rothermel, G. (2002). Test case prioritization: A family of empirical studies. *IEEE Transactions on Software Engineering*, 28:159–182.

[Engström et al. 2008] Engström, E., Skoglund, M., and Runeson, P. (2008). Empirical evaluations of regression test selection techniques: a systematic review. In *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement*, ESEM '08, pages 22–31, New York, NY, USA. ACM.

[Feitelson and Russell 2006] Feitelson, D. G. and Russell, B. (2006). Experimental computer science: The need for a cultural change.

[Fraser and Wotawa 2007] Fraser, G. and Wotawa, F. (2007). Redundancy based test-suite reduction. In *FASE'07: Proceedings of the 10th international conference on Fundamental approaches to software engineering*, pages 291–305, Berlin, Heidelberg. Springer-Verlag.

[Gamma et al. 1994] Gamma, E., Helm, R., Johnson, R., and Vlissides, J. M. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1 edition.

[Gimenes et al. 1999] Gimenes, M. S., Weis, G. M., and Huzita, E. H. M. (1999). Um padrão para definição de um gerenciador de processos de software. In *Proceedings of the 2nd Workshop IberoAmericano de Engenharia de Requisitos Y Ambientes Software*, pages 30–46.

[Graves et al. 1998] Graves, T., Harrold, M. J., Kim, J.-M., Porter, A., and Rothermel, G. (1998). An empirical study of regression test selection techniques. In *Proceedings of theInternational Conference on Software Engineering (ICSE 1998)*, pages 188–197, Kyoto, Japan.

[Graves et al. 2001] Graves, T. L., Harrold, M. J., Kim, J.-M., Porter, A., and Rothermel, G. (2001). An empirical study of regression test selection techniques. *ACM Trans. Softw. Eng. Methodol.*, 10(2):184–208.

[Harrold and Orso 2008] Harrold, M. J. and Orso, A. (2008). Retesting software during development and maintenance. In *Frontiers of Software Maintenance (FoSM 2008)*, pages 99–108, Beijing, China.

[Jain 1991] Jain, R. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation and Modeling*. John Wiley.

[Jorgensen 1995] Jorgensen, P. C. (1995). *Software Testing: A Craftsman's Approach*. CRC Press, Inc., Boca Raton, FL, USA.

[Kaner 2003] Kaner, C. (2003). What is a good test case.

[Kim et al. 2000] Kim, J., Porter, A., and Rothermel, G. (2000). An empirical study of regression test application frequency. In *In Proceedings of the 22nd International Conference on Software Engineering*, pages 126–135. ACM Press.

[Korel and Koutsogiannakis 2009] Korel, B. and Koutsogiannakis, G. (2009). Experimental comparison of code-based and model-based test prioritization. In *ICSTW '09: Proceedings of the IEEE International Conference on Software Testing, Verification, and Validation Workshops*, pages 77–84, Washington, DC, USA. IEEE Computer Society.

[Korel et al. 2005] Korel, B., Tahat, L. H., and Harman, M. (2005). Test prioritization using system models. *Software Maintenance, IEEE International Conference on*, 0:559–568.

[Korel et al. 2002] Korel, B., Tahat, L. H., and Vaysburg, B. (2002). Model based regression test reduction using dependence analysis. In *ICSM '02: Proceedings of the International Conference on Software Maintenance (ICSM'02)*, Washington, DC, USA. IEEE Computer Society.

[Kvam and Vidakovic 2007] Kvam, P. H. and Vidakovic, B. (2007). *Nonparametric Statistics with Applications to Science and Engineering (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

[Laski and Szermer 1992] Laski, J. and Szermer, W. (1992). Identification of program modifications and its applications in software maintenance. In *ICSM '92: Proceedings of the Conference on Software Maintenance*, pages 282–290. IEEE Computer Society.

[Leon and Podgurski 2003] Leon, D. and Podgurski, A. (2003). A comparison of coverage-based and distribution-based techniques for filtering and prioritizing test cases. In *Proceedings of the 14th International Symposium on Software Reliability Engineering*, ISSRE '03, pages 442–, Washington, DC, USA. IEEE Computer Society.

[Leung and White 1990] Leung, H. K. N. and White, L. (1990). A study of integration testing and software regression at the integration level. In *Proc. Conf. Software Maintenance*, pages 290–300.

[Leung and White 1991] Leung, H. K. N. and White, L. (1991). A cost model to compare regression test strategies. In *Proc. Conf. Software Maintenance*, pages 201–208.

[Lilja 2000] Lilja, D. J. (2000). *Measuring Computer Performance: A Practitioner's Guide*. Cambridge University Press.

[Ma et al. 2005] Ma, X.-y., He, Z.-f., Sheng, B.-k., and Ye, C.-q. (2005). A genetic algorithm for test-suite reduction. In *IEEE International Conference on System, Man and Cybernetics*, pages 133–139.

[Mahdian et al. 2009] Mahdian, A., Andrews, A. A., and Pilskalns, O. J. (2009). Regression testing with uml software designs: A survey. *J. Softw. Maint. Evol.*, 21(4):253–286.

[Mao and Lu 2005] Mao, C. and Lu, Y. (2005). Regression testing for component-based software systems by enhancing change information. In *Proceedings of the 12th Asia-Pacific Software Engineering Conference*, pages 611–618, Washington, DC, USA. IEEE Computer Society.

[McGregor and Sykes 2001] McGregor, J. D. and Sykes, D. A. (2001). *A Practical Guide to Testing Object-Oriented Software*. Addison-Wesley Object Technology Series. Addison-Wesley Professional.

[Musa 1998] Musa, J. (1998). *Software Reliability Engineered Testing*. McGraw-Hill, Inc., New York, NY, USA.

[Naslavsky et al. 2009] Naslavsky, L., Ziv, H., and Richardson, D. J. (2009). A model-based regression test selection technique. In *ICSM*, pages 515–518.

[Nogueira et al. 2007] Nogueira, S., Cartaxo, E., Torres, D., Aranha, E., and Marques, R. (2007). Model based test generation: An industrial experience. In *1st Brazilian Workshop on Systematic and Automated Software Testing*.

[Offutt and Abdurazik 1999] Offutt, J. and Abdurazik, A. (1999). Generating tests from UML specifications. In France, R. and Rumpe, B., editors, *UML'99 - The Unified Modeling Language. Beyond the Standard. Second International Conference, Fort Collins, CO, USA, October 28-30. 1999, Proceedings*, volume 1723, pages 416–429. Springer.

[Oliveira Neto and Machado 2008] Oliveira Neto, F. G. and Machado, P. D. L. (2008). Reengenharia da ferramenta LTS-BT. Technical report, Departamento de Sistemas e Computação da Universidade Federal de Campina Grande.

[Rothermel and Harrold 1996] Rothermel, G. and Harrold, M. J. (1996). Analyzing regression test selection techniques. *IEEE Transactions on Software Engineering*, 22:529–551.

[Rothermel and Harrold 1997] Rothermel, G. and Harrold, M. J. (1997). A safe, efficient regression test selection technique. *ACM Trans. Softw. Eng. Methodol.*, 6(2):173–210.

[Sajeev and Wibowo 2003] Sajeev, A. S. M. and Wibowo, B. (2003). Regression test selection based on version changes of components. In *Proceedings of the Tenth Asia-Pacific Software Engineering Conference Software Engineering Conference*, APSEC '03, Washington, DC, USA. IEEE Computer Society.

[Siegel and Junior 1988] Siegel, S. and Junior, N. J. C. (1988). *Nonparametric Statistics for The Behavioral Sciences*. McGraw-Hill.

[Sommerville 2010] Sommerville, I. (2010). *Software Engineering*. Addison-Wesley, Harlow, England, 9. edition.

[Subramaniam et al. 2009] Subramaniam, M., Xiao, L., Guo, B., and Pap, Z. (2009). An approach for test selection for efsms using a theorem prover. In *TESTCOM '09/FATES '09: Proceedings of the 21st IFIP WG 6.1 International Conference on Testing of Software and Communication Systems and 9th International FATES Workshop*, pages 146–162, Berlin, Heidelberg. Springer-Verlag.

[Tretmans and Brinksman 2002] Tretmans, G. J. and Brinksman, H. (2002). Côte de resyste – automated model based testing. In *Proceedings of Progress 2002 - 3rd Workshop on Embedded Systems*, pages 246–255.

[White et al. 2008] White, L. J., Jaber, K., Robinson, B., and Rajlich, V. (2008). Extended firewall for regression testing: an experience report. *Journal of Software Maintenance*, 20(6):419–433.

[Wohlin et al. 2000] Wohlin, C., Runeson, P., Host, M., Ohlsson, C., Regnell, B., and Wesslén, A. (2000). *Experimentation in Software Engineering: an Introduction*. Kluver Academic Publishers.

[Yin 1994] Yin, R. K. (1994). *Case Study Research : Design and Methods*. SAGE Publications.

Apêndice A

Teste de Normalidade

Para verificar a normalidade dos dados, uma das premissas de ANOVA, o teste estatístico de Anderson-Darling foi executado para cada técnica sob os aspectos de inclusão, precisão e eficiência. O resultado da execução deste teste para as entradas especificadas é mostrado nas subseções a seguir, considerando o nível de significância de $\alpha = 0,05$.

A.1 Inclusão

O teste Anderson-Darling realizado sobre os dados de inclusão não pode ser realizado nas técnicas T_1, T_2 e T_4, pois os valores obtidos de cada uma destas técnicas possuía variância nula, o que já caracteriza que estes dados não possuem distribuição normal.

Para as demais técnicas, em todos os casos o p foi inferior a $0,05$, que é menor que o valor de α definido previamente. Portanto, de acordo com as características do teste de Anderson-Darling, nenhuma das técnicas possui distribuição normal. A Figura A.1 ilustra os gráficos resultantes destes testes.

A.2 Precisão

Para observar a normalidade dos dados de precisão, o teste Anderson-Darling foi realizado sobre os dados obtidos para esta variável dependente. Este teste não pôde ser realizado para as técnicas T_2 e T_4, pois os valores obtidos de cada uma destas técnicas possuía variância nula, o que já caracteriza que estes dados não possuem uma distribuição normal.

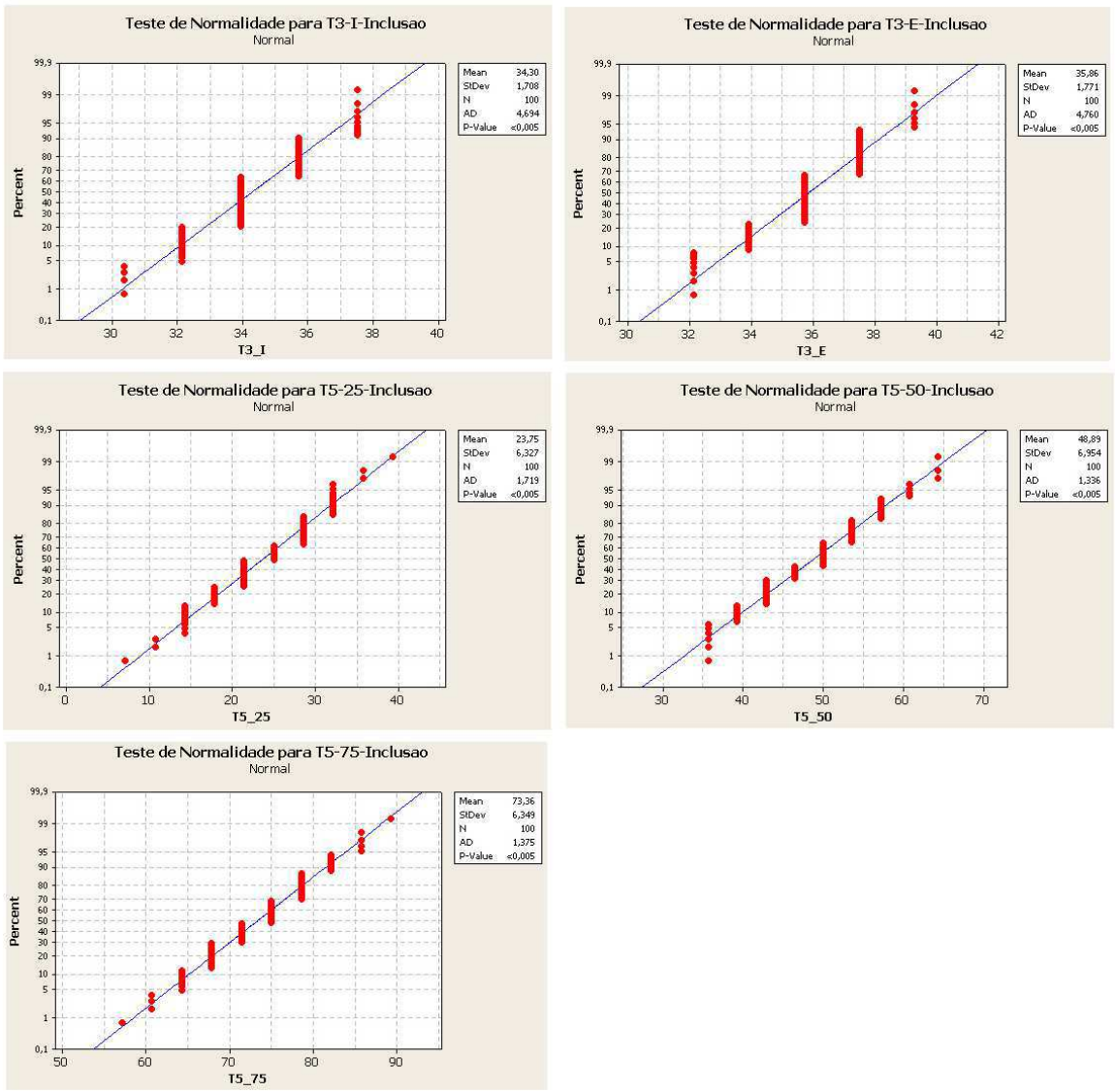

Figura A.1: Resultado do teste Anderson-Darling para T_{3i},T_{3e}, $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$ no tocante à propriedade de inclusão.

As demais técnicas, por outro lado, foram submetidas ao testes. Para todos os casos o p foi inferior a $0,005$, que é menor que o valor de α definido previamente. Isto indica que os dados não apresentam uma distribuição normal. A Figura A.2 ilustra os gráficos resultantes destes testes.

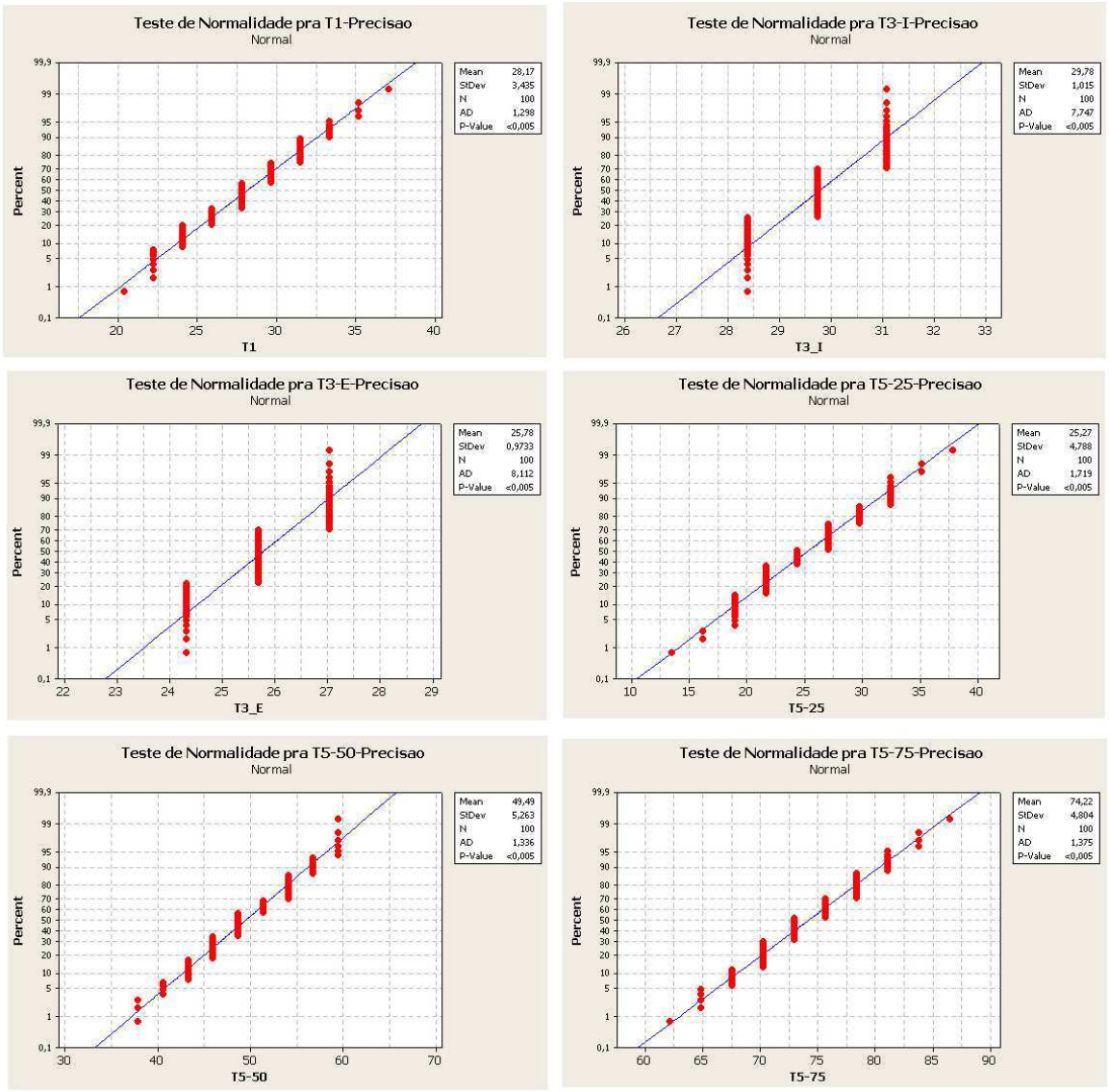

Figura A.2: Resultado do teste Anderson-Darling para T_1, T_{3i},T_{3e}, $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$ no tocante à propriedade de precisão.

A.3 Eficiência

O teste Anderson-Darling, considerando os dados de eficiência, foi realizado para todas as técnicas, pois nenhuma delas possuiu variância nula. Em todos os casos o p foi inferior a $0,005$, que é menor que o valor de α definido previamente. Diante disto, não podemos afirmar que as técnicas apresentam uma distribuição normal. A Figura A.3 ilustra os gráficos resultantes destes testes.

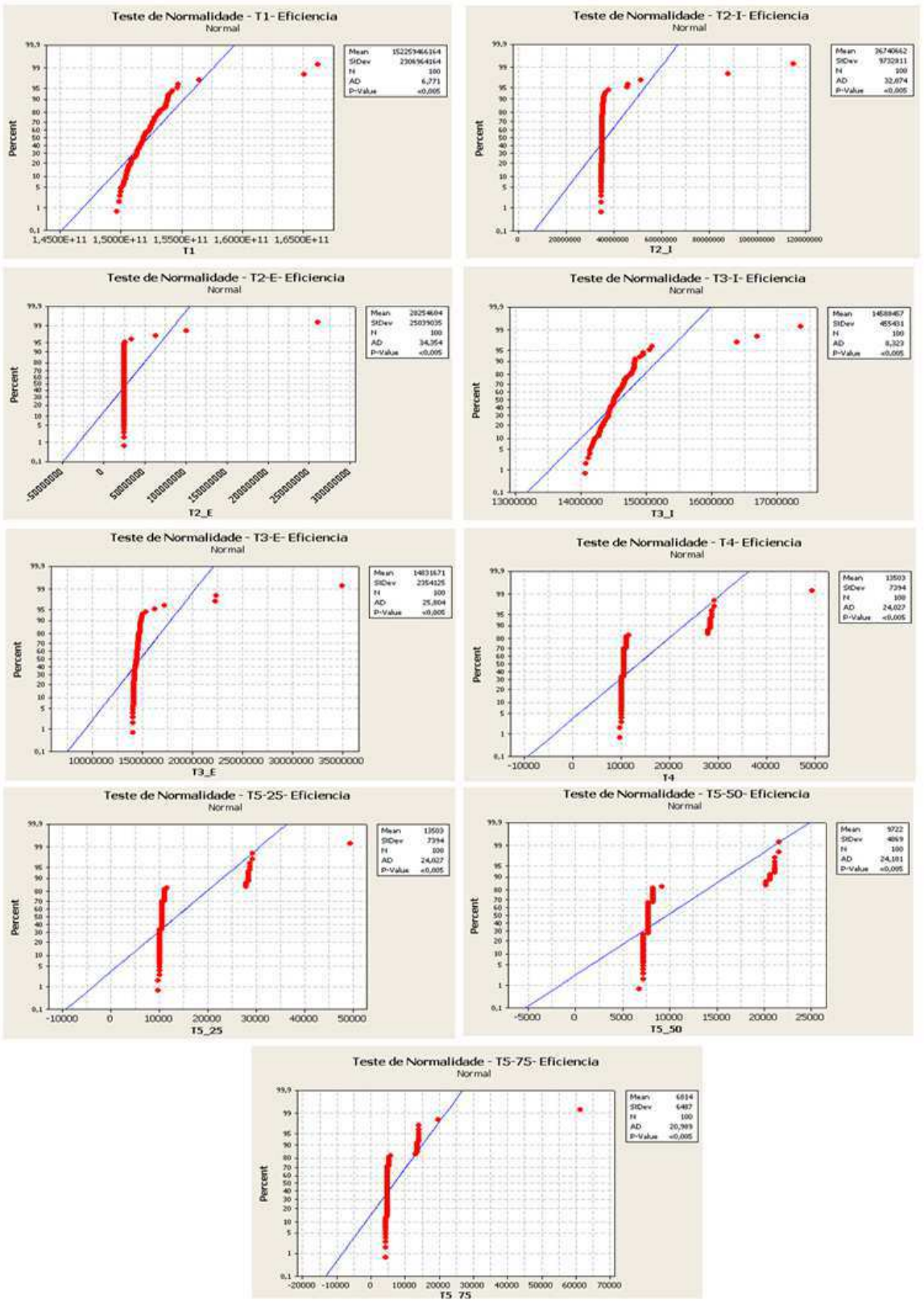

Figura A.3: Resultado do teste Anderson-Darling para todas as técnicas no tocante à propriedade de eficiência.

A.4 Potencial de Redução das técnicas

Os dados de potencial de redução das técnicas foram submetidos ao teste de Anderson-Darling para verificar se estes possuem uma distribuição normal. As técnicas T_2, T_4, $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$ apresentaram variância nula, portanto, já é possível observar que os dados de potencial de redução destas técnicas não apresentam distribuição normal.

Aplicando o teste nas demais técnicas (T_1, $T3i$ e ${}_T3e$), foi observado que o p é menor que $0,005$ para todas estas amostras. Como o p é menor que o nível de significância ($\alpha = 0,05$), não podemos afirmar que estes dados possuem uma distribuição normal. Os resultados do teste podem ser verificados na Figura A.4

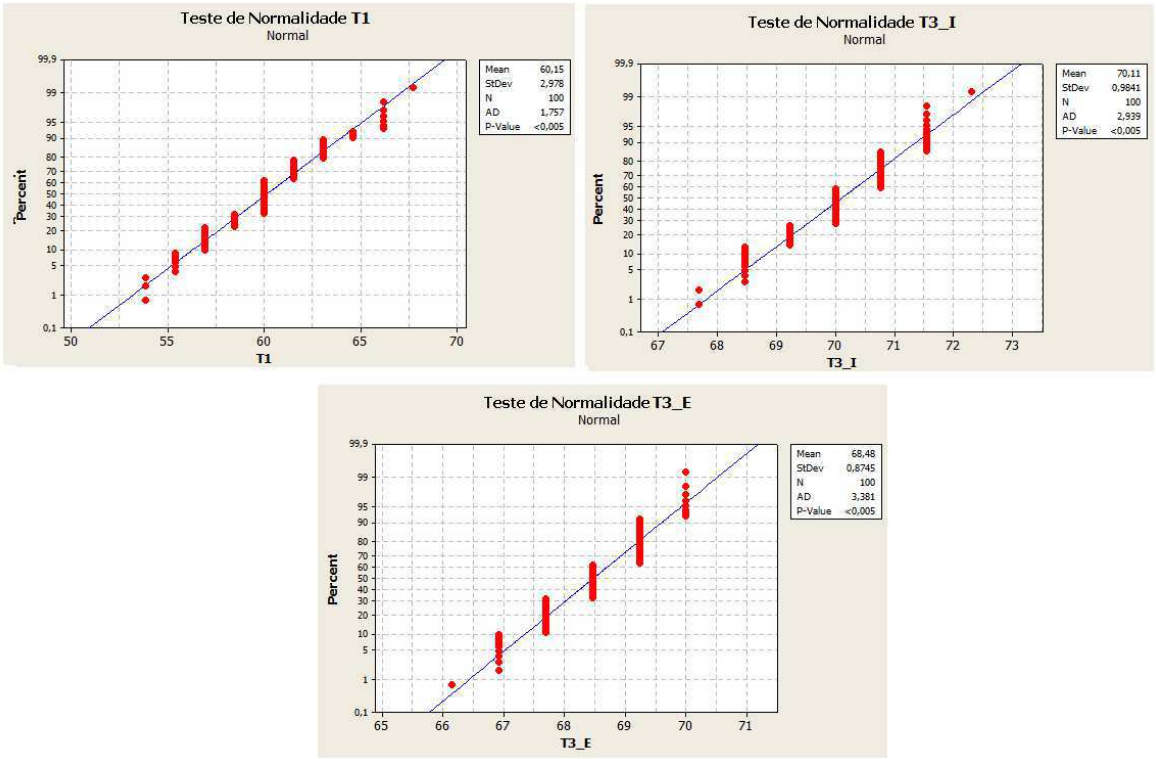

Figura A.4: Resultado do teste Anderson-Darling para todas as técnicas no tocante à propriedade de potencial de redução.

A.5 Densidade de Faltas

O teste de normalidade de Anderson-Darling foi aplicado aos dados de densidade de faltas das técnicas, para verificar se os dados seguem uma distribuição normal. Observando

os dados, verificamos que as técnicas T_2 e T_4 apresentam variância nula, e portanto não é necessário realizar o teste de normalidade pois a variância nula já indica que os dados não seguem uma distribuição normal. O teste de Anderson-Darling foi aplicado às demais técnicas (T_1, T_3 e T_5), e os resultados podem ser observados nos gráficos da Figura A.5.

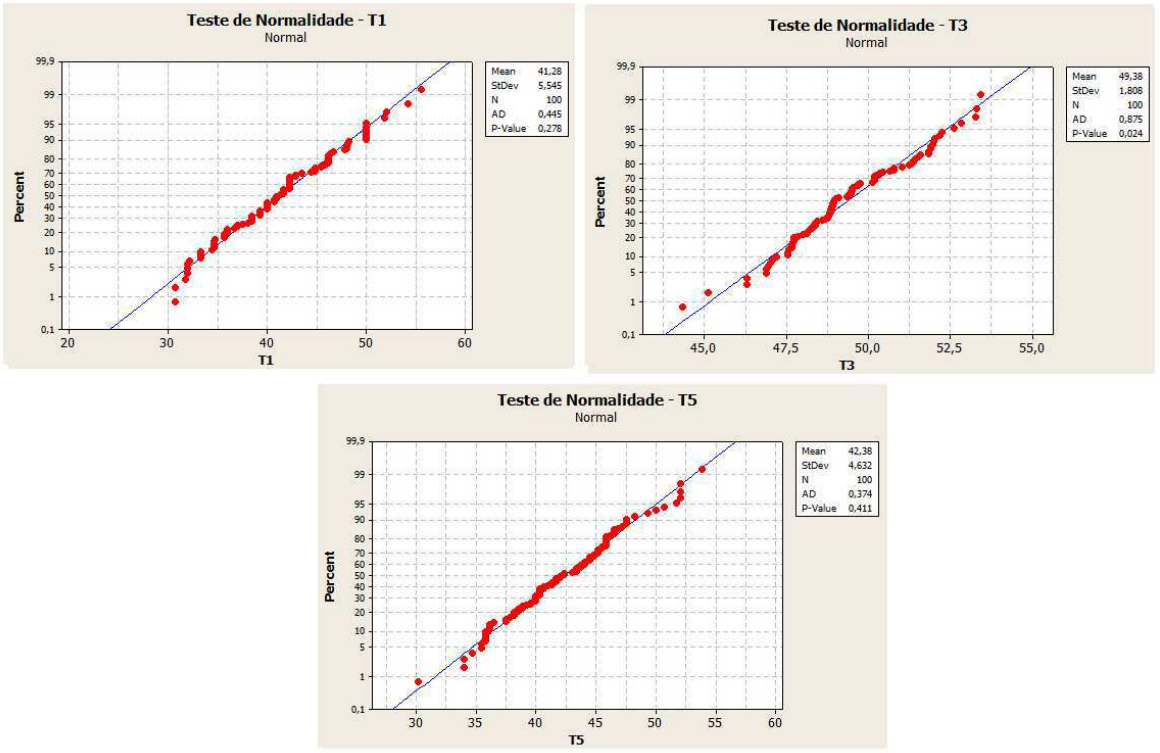

Figura A.5: Resultado do teste Anderson-Darling para todas as técnicas no tocante à propriedade de densidade de faltas.

De acordo com os resultados do teste, as técnicas T_1 e T_5 apresentaram um p de $0,278$ e $0,411$, respectivamente. Este p é menor que o nível de significância ($\alpha = 0,05$) o que indica que os dados destas técnicas apresentam uma distribuição normal. O p referente aos dados T_3 ($p = 0,024$), foi menor que o nível de significância, que, pelo teste de Anderson-Darling, indica que não é possível afirmar que os dados seguem uma distribuição normal.

Apêndice B

Comparativo entre as técnicas com diferentes características

No projeto experimental foi considerado que cada técnica seria auto-contida, ou seja, assumiu-se que as mesmas possuíam todos os parâmetros para o seu funcionamento adequado. Porém, é possível configurar alguns destes parâmetros de forma diferente e nem todos os parâmetros constituem uma mesma entrada para todas as técnicas. Em algumas situações, vale ressaltar, é de interesse do investigador analisar o comportamento de uma mesma técnica sob diferentes condições de inicialização, verificando se esta mudança caracteriza, por exemplo, em um desempenho melhor sob um determinado aspecto (e.g. uma variável dependente).

Neste contexto, foram utilizados diferentes parâmetros de inicialização das técnicas T_2, T_3 e T_5. Para tratar estas situações, foi feito um teste de hipóteses para analisar quando os parâmetros de configuração promoveram uma distinção entre o comportamento da técnica. Quando esta distinção era observada, as diferentes configurações foram tratadas separadamente, ou seja, caracterizando um novo nível do fator do experimento.

As seções a seguir mostram os testes de hipóteses realizados para cada uma das técnicas e as conclusões alcançadas com os mesmos. Fixa-se ainda o nível de signficância em $\alpha = 0.05$. Os resultados reportados aqui foram respeitados ao longo do estudo experimental.

B.1 Técnica T_2 – Técnica de Análise Baseada em Riscos

A técnica T_2 fazia uso de dois tipos de sujeitos: experientes e inexperientes na definição dos valores de custos e riscos dos casos de teste de regressão. O testador deve especificar um valor de custo caracterizado neste experimento como um valor real de 1 a 10, (sendo 10 o maior custo) que represente o custo de manutenção na ferramenta se o respectivo caso de teste apresentar faltas de regressão. Por sua vez, o valor de risco (inteiros de 1 a 5, sendo 5 o maior valor de risco), representa o risco de um caso de teste expor faltas de regressão durante a versão funcional da ferramenta.

Estes valores foram especificados pelos sujeitos do experimento e incorporados à técnica T_2 durante a sua configuração e execução. Considera-se que T_{2e} é a técnica configurada por um testador experiente e T_{2i} é a técnica configurada por um testador inexperiente. Sejam $I(x)$, $P(x)$, $E(x)$, $R(x)$ e $D(x)$ a inclusão, precisão, eficiência, potencial de redução e densidade de faltas de uma técnica x, respectivamente. O teste de hipóteses realizado para a técnica T_2 foi:

Hipóteses Nulas (H_0)

- A técnica T_2 é independente do sujeito no tocante à inclusão, ou seja, $I(T_{2e}) = I(T_{2i})$;
- A técnica T_2 é independente do sujeito no tocante à precisão, ou seja, $P(T_{2e}) = P(T_{2i})$;
- A técnica T_2 é independente do sujeito no tocante à eficiência, ou seja, $E(T_{2e}) = E(T_{2i})$.
- A técnica T_2 é independente do sujeito no tocante ao potencial de redução, ou seja, $R(T_{2e}) = R(T_{2i})$.
- A técnica T_2 é independente do sujeito no tocante à densidade de faltas, ou seja, $D(T_{2e}) = D(T_{2i})$.

Hipóteses Alternativas (H_1)

- A técnica T_2 é dependente do sujeito no tocante à inclusão, ou seja, $I(T_{2e}) \neq I(T_{2i})$;

- A técnica T_2 é dependente do sujeito no tocante à precisão, ou seja, $P(T_{2e}) \neq P(T_{2i})$;
- A técnica T_2 é dependente do sujeito no tocante à eficiência, ou seja, $E(T_{2e}) \neq E(T_{2i})$.
- A técnica T_2 é dependente do sujeito no tocante ao potencial de redução, ou seja, $R(T_{2e}) \neq R(T_{2i})$.
- A técnica T_2 é dependente do sujeito no tocante à densidade de faltas, ou seja, $D(T_{2e}) \neq D(T_{2i})$.

Para todas as variáveis dependentes analisadas, foi verificado, inicialmente se os dados apresentavam uma distribuição normal. No caso positivo, era utilizado um teste t de Student, caso contrário era utilizado o teste não-paramétrico de Mann-Whitney para obter conclusões acerca dos testes de hipóteses. Os testes foram realizados na ferramenta Minitab utilizando um nível de confiança de 95% para a construção dos intervalos de confiança dos testes. A verificação das hipóteses levantadas para a técnica T_2 é apresentada a seguir.

B.1.1 Inclusão

Para decidir qual o teste estatístico seria adequado utilizar, inicialmente foi efetuado o teste de normalidade de Anderson-Darling para verificar se os dados seguiam uma distribuição normal. Os resultados destes testes são exibidos na Figura B.1. Uma vez que em ambos os casos o p foi menor que α, não foi possível aceitar a hipótese de que os dados seguiam distribuição normal. Parte-se, portanto, para o uso do teste não-paramétrico Mann-Whitney.

O resultado do teste Mann-Whitney, obtido pela ferramenta Minitab, é mostrado a seguir:

```
                              N  Median
INEXPERIENTE - INCLUSAO - T2  200  92,857
EXPERIENTE - INCLUSAO - T2    200  92,857


Point estimate for ETA1-ETA2 is -0,0001
95,0 Percent CI for ETA1-ETA2 is (-0,0001;0,0001)
W = 40100,0
```

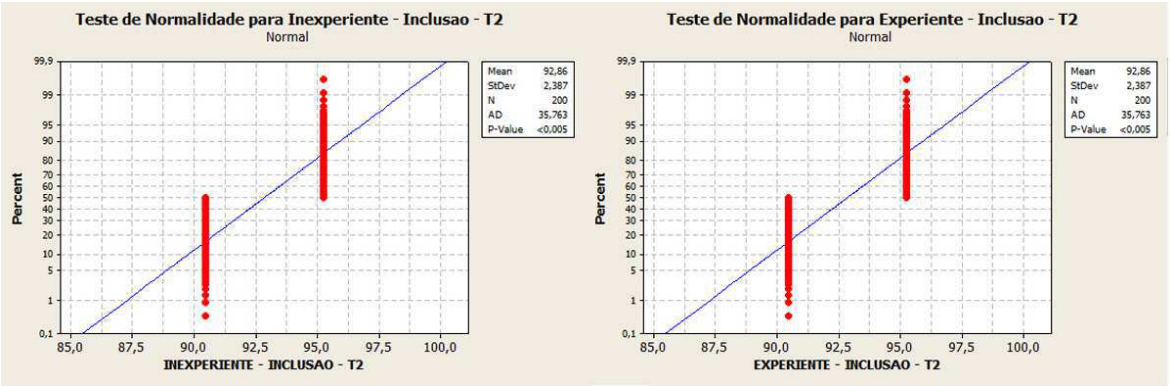

Figura B.1: Resultados dos testes Anderson-Darling para T_{2i} e T_{2e} no tocante à propriedade de inclusão.

```
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 1,0000
The test is significant at 1,0000 (adjusted for ties)
```

O p obtido foi igual a 1, que é maior que α. Com isto, não foi possível rejeitar a hipótese nula de que as populações são diferentes. Portanto, para a propriedade de inclusão ao nível de confiança de 95%, , não há diferença entre T_{2i} e T_{2e}. Conseqüentemente, é possível afirmar que a técnica T_2 é independente de sujeito para a propriedade de inclusão.

B.1.2 Precisão

Inicialmente foi verificado se os dados seguiam uma distribuição normal. Esta verificação foi realizada para decidir que teste estatístico utilizar para o teste de hióteses. Os resultados dos testes Anderson-Darling realizados são exibidos na Figura B.2. Como o p obtido nas duas distribuições foi menor que α, não foi possível aceitar a hipótese de que os dados seguiam distribuição normal. Parte-se, portanto, para o uso do teste não-paramétrico Mann-Whitney.

O resultado do teste Mann-Whitney da ferramenta Minitab é mostrado a seguir:

```
                                N  Median
INEXPERIENTE - PRECISAO - T2  200  85,135
EXPERIENTE - PRECISAO - T2    200  85,135
```

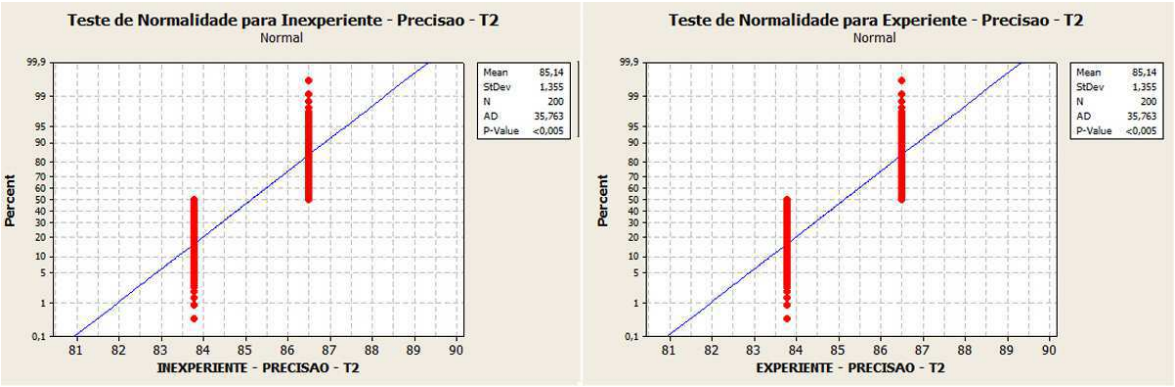

Figura B.2: Resultados dos testes Anderson-Darling para T_{2i} e T_{2e} no tocante à propriedade de precisão.

```
Point estimate for ETA1-ETA2 is -0,0001
95,0 Percent CI for ETA1-ETA2 is (-0,0001;0,0001)
W = 40100,0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 1,0000
The test is significant at 1,0000 (adjusted for ties)
```

O p resultante foi 1, que é maior que o nível de significância ($\alpha = 0,05$). Com isto, não foi possível rejeitar a hipótese nula de que as populações são diferentes. Portanto, para a propriedade de precisão, não há diferença, ao nível de confiança de 95%, entre T_{2i} e T_{2e}. Conseqüentemente, é possível afirmar que a técnica T_2 é independente de sujeito para a variável dependente de precisão.

B.1.3 Eficiência

O primeiro aspecto verificado foi a distribuição dos dados de eficiência obtidos com as duas configurações. Para isto, foi efetuado o teste de normalidade de Anderson-Darling para verificar se os dados seguiam uma distribuição normal. Os resultados destes testes são exibidos na Figura B.3. O p observado em ambos os casos foi menor que α, indicando que não é possível aceitar a hipótese de que os dados são normalmente distribuídos. Parte-se, portanto, para o uso do teste não-paramétrico Mann-Whitney.

O resultado do teste Mann-Whitney é mostrado a seguir:

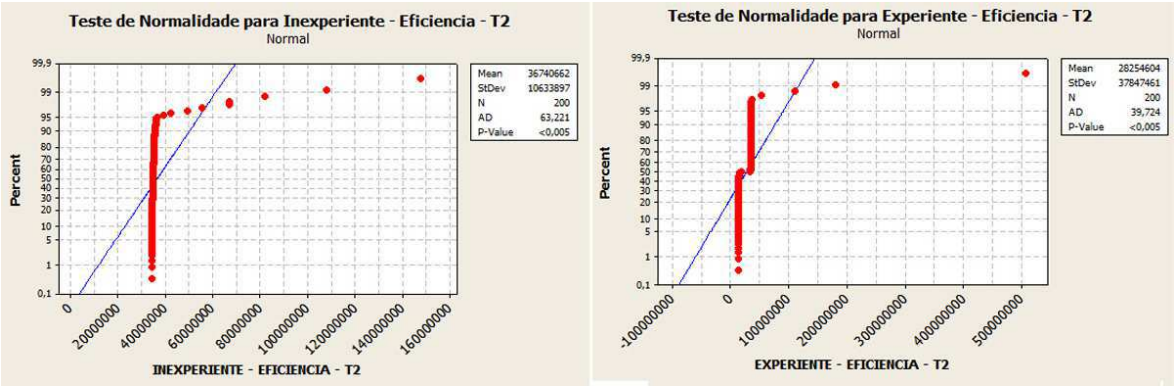

Figura B.3: Resultados dos testes Anderson-Darling para T_{2i} e T_{2e} no tocante à propriedade de eficiência.

```
                                N  Median
INEXPERIENTE - EFICIENCIA - T2  200  35036487
EXPERIENTE - EFICIENCIA - T2    200  27293141


Point estimate for ETA1-ETA2 is 19458507
95,0 Percent CI for ETA1-ETA2 is (771454;20157636)
W = 50270,0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0,00001
The test is significant at 0,00001 (adjusted for ties)
```

O p obtido foi igual a $0,00001$, que é menor que α. Diante deste resultado, é possível rejeitar a hipótese nula de que as populações são diferentes. Ou seja, para a propriedade de eficiência, ao nível de confiança de 95%, há diferença entre T_{2i} e T_{2e}. Conseqüentemente, é possível afirmar que a técnica T_2 é dependente de sujeito para a propriedade de eficiência.

B.1.4 Potencial de Redução

A partir dos dados das 100 execuções da técnica utilizando as duas configurações, foi possível observar que o potencial de redução da técnica foi o mesmo (com média $12,06$ e variância zero). Ou seja, para as duas configurações, a técnica apresentou o mesmo potencial de redução, indicando que não há diferença estatisticamente significativa entre a utilização de um testador experiente, ou inexperiente. Diante disto, não foi necessário realizar testes de

normalidade, ou testes estatísticos para obter esta conclusão.

B.1.5 Densidade de Faltas

Com o objetivo de definir o teste estatístico a ser utilizado, foi realizado um teste de normalidade de Anderson-Darling com os dados de densidade de faltas da técnica com as diferentes configurações. Os resultados deste teste são apresentados na Figura B.4. Uma vez que o p obtido é menor que o nível de significância utilizado no teste, não podemos afirmar que os dados seguem uma dsitribuição normal.

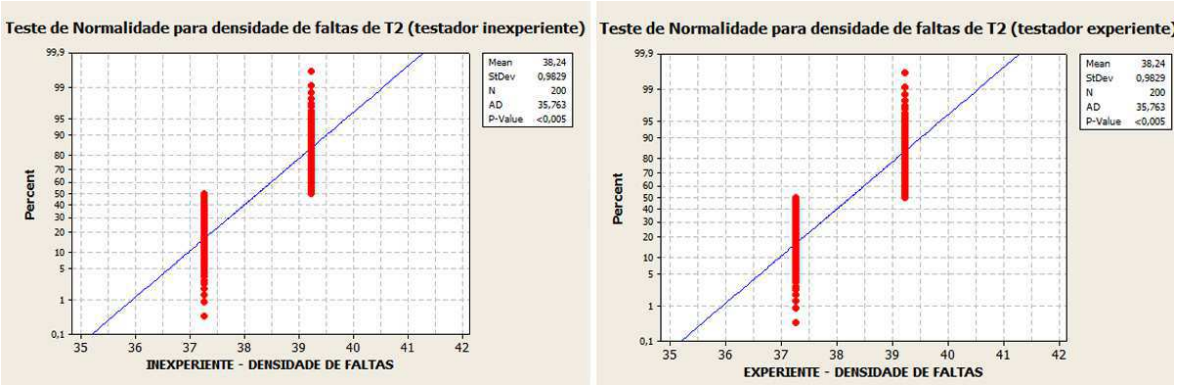

Figura B.4: Resultados dos testes Anderson-Darling para T_{2i} e T_{2e} no tocante à densidade de faltas.

Portanto, devemos utilizar o teste não paramétrico de Mann-Whitney para verificar se as diferentes configurações afetam significativamente os resultados do potencial de redução das técnicas. Os resultados deste teste são apresentados a seguir.

```
                                    N  Median
INEXPERIENTE - DENSIDADE DE FALTA  200  38,235
EXPERIENTE   - DENSIDADE DE FALTA  200  38,235

Point estimate for ETA1-ETA2 is -0,000
95,0 Percent CI for ETA1-ETA2 is (-0,000;0,000)
W = 40100,0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 1,0000
The test is significant at 1,0000 (adjusted for ties)
```

A partir do teste de Mann-Whitney realizado, podemos observar que o $p = 1,000$ é menor que o nível de significância utilizado pelo teste ($\alpha = 0,05$). Diante disto não podemos rejeitar a hipótese nula em favor da hipótese alternativa. Portanto, sob um nível de confiança de 95%, não podemos afirmar que há diferença estatisticamente significativa entre os resultados da técnica T_2 com relação ao nível de experiência do testador.

B.2 Técnica T_3 – WSA para Teste de Regressão

A técnica T_3 faz uso de dois tipos de sujeitos: experientes e inexperientes na definição dos perfis de uso da ferramenta. A técnica T_3 necessita que o testador atribua valores de probabilidade para cada transição saindo de um nó de decisão no modelo. Estes valores de probabilidade são utilizados durante a seleção dos casos de teste com o objetivo de selecionar os casos de teste que exercitam os fluxos mais executados por um usuário ao utilizar a ferramenta. Diante disto, os casos de teste que exercitam as funcionalidades mais utilizadas na ferramenta são executados.

O nível de experiência do testador, pode auxiliar a atribuição destes valores de probabilidade provendo cenários mais reais dos testes executados e possibilitando uma maior cobertura de faltas pela suíte reduzida. Os sujeitos do experimento definiram os perfis de uso da ferramenta, i.e. os valores de probabilidade que representam os fluxos mais executados por um usuário, para a configuração e execução da técnica T_3. Para a verificação do impacto do nível do testador na configuração da técnica, iremos considerar que T_{3e} é a técnica com uso de um testador experiente e T_{3i} é a técnica com uso de um testador inexperiente.

Sejam $I(x)$, $P(x)$, $E(x)$, $R(x)$ e $D(x)$ a inclusão, precisão, eficiência, potencial de redução e densidade de faltas de uma técnica x, respectivamente. O teste de hipóteses realizado para a técnica T_3 foi:

Hipóteses Nulas (H_0)

- A técnica T_3 é independente do sujeito no tocante à inclusão, ou seja, $I(T_{3e}) = I(T_{3i})$;
- A técnica T_3 é independente do sujeito no tocante à precisão, ou seja, $P(T_{3e}) = P(T_{3i})$;

- A técnica T_3 é independente do sujeito no tocante à eficiência, ou seja, $E(T_{3e}) = E(T_{3i})$.
- A técnica T_3 é independente do sujeito no tocante ao potencial de redução, ou seja, $R(T_{3e}) = R(T_{3i})$.
- A técnica T_3 é independente do sujeito no tocante à densidade de faltas, ou seja, $D(T_{3e}) = D(T_{3i})$.

Hipóteses Alternativas (H_1)

- A técnica T_3 é dependente do sujeito no tocante à inclusão, ou seja, $I(T_{3e}) \neq I(T_{3i})$;
- A técnica T_3 é dependente do sujeito no tocante à precisão, ou seja, $P(T_{3e}) \neq P(T_{3i})$;
- A técnica T_3 é dependente do sujeito no tocante à eficiência, ou seja, $E(T_{3e}) \neq E(T_{3i})$.
- A técnica T_3 é dependente do sujeito no tocante ao potencial de redução, ou seja, $R(T_{3e}) \neq R(T_{3i})$.
- A técnica T_3 é dependente do sujeito no tocante à densidade de faltas, ou seja, $D(T_{3e}) \neq D(T_{3i})$.

A análise desta dependência entre a técnica e a sua configuração foi realizada para cada variável dependente do estudo experimental. O primeiro aspecto a ser investigado é a normalidade dos dados, através de um teste de Anderson-Darling. Os dados que apresentam uma distribuição normal são submetidos a um teste t de Student, caso contrário deve ser aplicado o teste não-paramétrico de Mann-Whitney para obter conclusões acerca dos testes de hipóteses. Os testes foram realizados na ferramenta Minitab utilizando um nível de confiança de 95% para a construção dos intervalos de confiança dos testes. A verificação das hipóteses levantadas para a técnica T_3 é apresentada a seguir.

B.2.1 Inclusão

Para decidir qual o teste estatístico seria adequado utilizar, inicialmente foi efetuado o teste de Anderson-Darling para verificar se os dados seguiam uma distribuição normal. Os resul-

tados dos testes Anderson-Darling realizados são exibidos na Figura B.5.

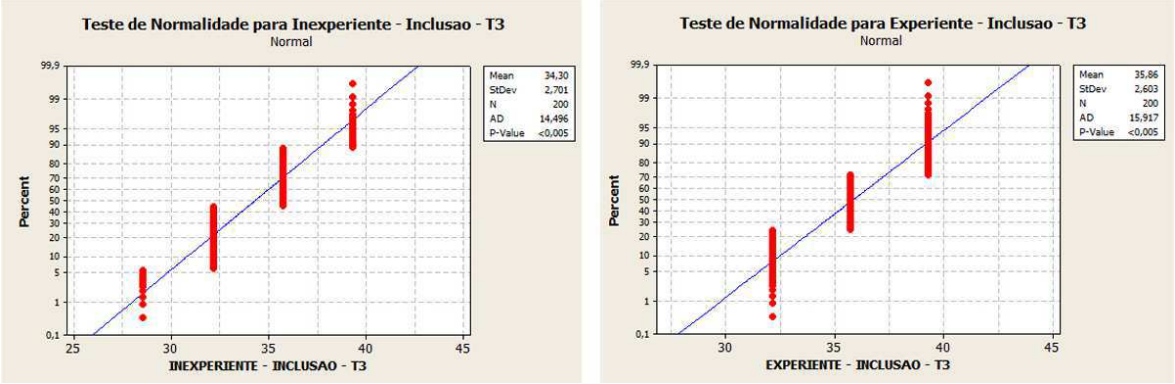

Figura B.5: Resultados dos testes Anderson-Darling para T_{3i} e T_{3e} no tocante à propriedade de inclusão.

Uma vez que em ambos os casos o p foi menor que α, não foi possível aceitar a hipótese de que os dados possuem uma distribuição normal. Parte-se, portanto, para o uso do teste não-paramétrico Mann-Whitney. O resultado deste teste, utilizando a ferramenta Minitab, é mostrado a seguir:

```
                                N  Median
INEXPERIENTE - INCLUSAO - T3  200  35,714
EXPERIENTE   - INCLUSAO - T3  200  35,714

Point estimate for ETA1-ETA2 is 0,0001
95,0 Percent CI for ETA1-ETA2 is (-3,571;-0,0001)
W = 34256,5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0,00001
The test is significant at 0,00001 (adjusted for ties)
```

O p obtido foi igual a $0,00001$, que é menor que α. Com isto, é possível rejeitar a hipótese nula de que as populações são diferentes. Portanto, para a propriedade de inclusão, há diferença, ao nível de confiança de 95%, entre T_{3i} e T_{3e}. Conseqüentemente, é possível afirmar que a técnica T_3 é dependente de sujeito para a propriedade de inclusão.

B.2.2 Precisão

Com o objetivo de decidir o teste estatístico adequado para obter os resultados de precisão, com relação às duas configurações de T_3, foi efetuado o teste de Anderson-Darling para verificar se os dados seguiam uma distribuição normal. Os resultados dos testes Anderson-Darling realizados são exibidos na Figura B.6.

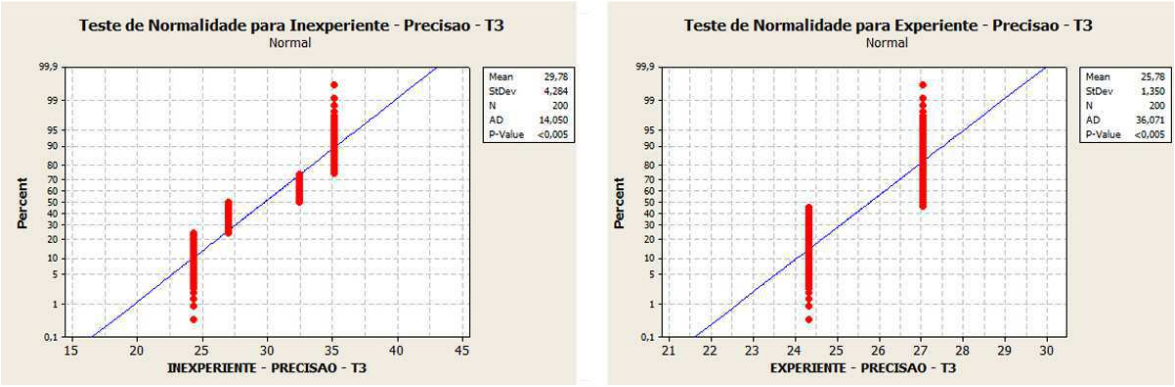

Figura B.6: Resultados dos testes Anderson-Darling para T_{3i} e T_{3e} no tocante à propriedade de precisão.

Uma vez que em ambos os casos o p foi menor que o nível de significância ($\alpha = 0,05$), não foi possível aceitar a hipótese de que os dados seguiam uma distribuição normal. Devemos utilizar, então, o teste de Mann-Whitney. Os dados foram submetidos à ferramenta Minitab para a realização deste teste. O resultado é mostrado a seguir:

```
                                N  Median
INEXPERIENTE - PRECISAO - T3  200  29,730
EXPERIENTE - PRECISAO - T3    200  27,027

Point estimate for ETA1-ETA2 is 4,054
95,0 Percent CI for ETA1-ETA2 is (2,703;5,406)
W = 49900,0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0,00001
The test is significant at 0,00001 (adjusted for ties)
```

Como resultado do teste, foi obtido um p igual a $0,00001$. Este p é menor que α, o que possibilita rejeitar a hipótese nula de que as populações são iguais. Portanto, para a pro-

priedade de precisão, ao nível de confiança de 95%, há diferença entre T_{3i} e T_{3e}. Conseqüentemente, a técnica T_3 é dependente da configuração fornecida pelo sujeito para a propriedade de precisão.

B.2.3 Eficiência

Para decidir qual o teste estatístico seria adequado utilizar, inicialmente foi efetuado o teste de Anderson-Darling para verificar se os dados de eficiência de T_3 seguiam uma distribuição normal. Os resultados destes testes são apresentados na Figura B.7.

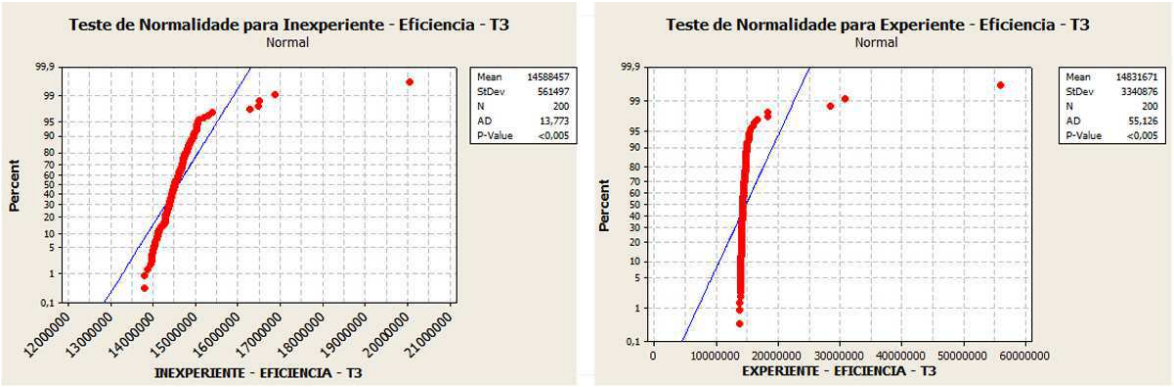

Figura B.7: Resultados dos testes Anderson-Darling para T_{3i} e T_{3e} no tocante à propriedade de eficiência.

A partir do resultado dos testes de normalidade, observamos que, em ambos os casos, o p foi menor que o nível de significância. Diante disto, não podemos aceitar a hipótese de que estes dados seguem uma distribuição normal. Dessa forma, devemos utilizar o teste não-paramétrico de Mann-Whitney.

O resultado do teste Mann-Whitney é mostrado a seguir:

```
                                N    Median
INEXPERIENTE - EFICIENCIA - T3  200  14511587
EXPERIENTE - EFICIENCIA - T3    200  14338847


Point estimate for ETA1-ETA2 is 148544
95,0 Percent CI for ETA1-ETA2 is (78579;214184)
W = 44771,0
```

```
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0,0001
The test is significant at 0,0001 (adjusted for ties)
```

A partir do resultado do teste, comparamos o p com o nível de significância ($\alpha = 0,05$). Como o p obtido foi $0,00001$, i.e. $p < \alpha$, é possível rejeitarmos a hipótese nula de que as populações são iguais. Portanto, para a propriedade de eficiência, há diferença, ao nível de confiança de 95%, entre T_{3i} e T_{3e}. Conseqüentemente, é possível afirmar que a técnica T_3 é dependente do nível de experiência do testador (sujeito) para a variável dependente de eficiência.

B.2.4 Potencial de Redução

Para descobrir se as diferentes configurações de T_3 baseada nos diferentes níveis de experiência do testador, causam diferenças estatisticamente significativas nos resultados das técnicas, devemos realizar um teste de hipótese. Para saber qual teste estatístico é adequado para os dados obtidos, é necessário realizar um teste de normalidade. O teste de normalidade de Anderson-Darling permite observar se uma amostra possui uma distribuição normal. Os resultados obtidos das diferentes configurações foram submetidos a este teste e são apresentados na Figura B.8

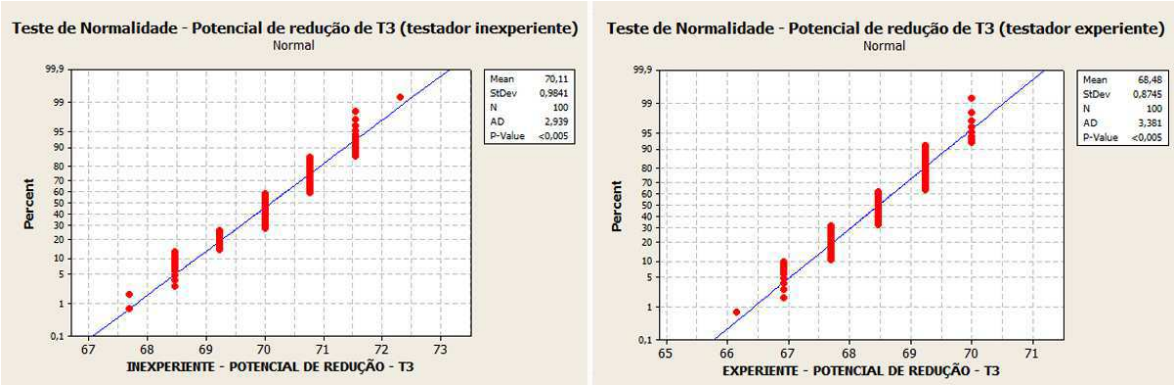

Figura B.8: Resultados dos testes Anderson-Darling para T_{3i} e T_{3e} no tocante ao potencial de redução.

Observando os gráficos da Figura B.8, verificamos que ambos os p são menores que o nível de significância utilizado no teste. Portanto, não podemos afirmar que as amostras

seguem uma distribuição normal. Diante disto não podemos utilizar o teste paramétrico t de Student, e devemos partir para a opção não-paramétrica correspondente, ou seja, o teste de Mann-Whitney. Estes dados foram então submetidos à ferramenta Minitab para a realização do teste de Mann-Whitney. O resultado deste teste é apresentado a seguir:

```
                              N  Median
INEXPERIENTE - REDUCAO - T3  100  70,000
EXPERIENTE - REDUCAO - T3    100  68,462


Point estimate for ETA1-ETA2 is 1,538
95,0 Percent CI for ETA1-ETA2 is (1,538;1,538)
W = 13871,5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0,00001
The test is significant at 0,00001 (adjusted for ties)
```

O teste de Mann-Whitney forneceu um p igual a $0,00001$. Uma vez que o p é menor que o nível de significância utilizado no teste, podemos então rejeitar a hipótese nula. Portanto, sob um nível de confiança de 95%, podemos rejeitar a hipótese de que a técnica T_3 é independente do nível de experiência do testador, com relação ao potencial de redução

B.2.5 Densidade de Faltas

.

Para verificar se as diferentes configurações afetavam a densidade de faltas da técnica, foram realizados, incialmente, testes de Anderson-Darling referentes às duas amostras obtidas na execução da técnica com cada configuração. Estes testes foram realizados para verificar qual o teste estatístico adequado para o teste das hipóteses levantadas no início desta subseção. Os resultados destes testes são apresentados na Figura B.9.

A partir dos gráficos da Figura B.9, podemos ver que o p, obtido nos dois teste, é menor que o nível de significância estabelecido. Portanto, não podemos aceitar a hipótese de que as amostras seguem uma distribuição normal. Dessa forma, o teste estatístico que devemos aplicar é o teste de Mann-Whitney. Este teste foi realizado utilizando a ferramenta Minitab, e os resultados são apresentados a seguir.

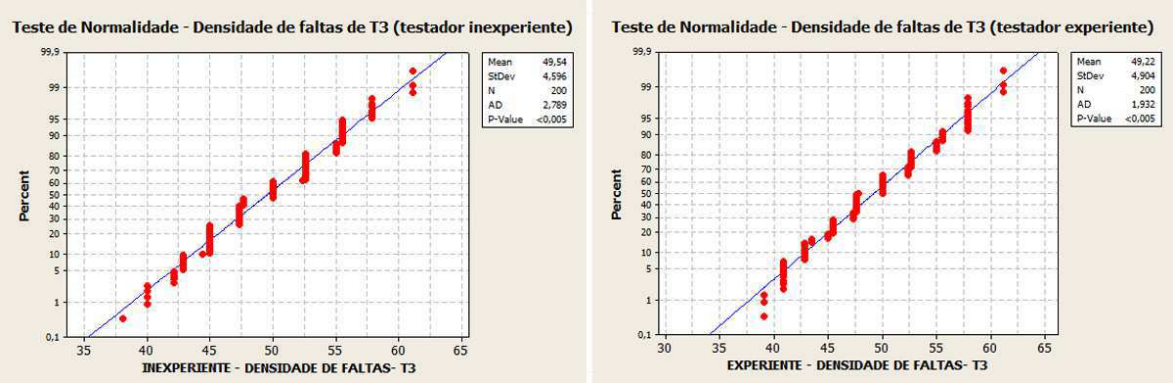

Figura B.9: Resultados dos testes Anderson-Darling para T_{3i} e T_{3e} no tocante à densidade de faltas.

```
                                N  Median
INEXPERIENTE - EFICACIA - T3  200  50,000
EXPERIENTE   - EFICACIA - T3  200  48,913

Point estimate for ETA1-ETA2 is 0,000
95,0 Percent CI for ETA1-ETA2 is (-0,250;1,914)
W = 40560,5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0,6907
The test is significant at 0,6891 (adjusted for ties)
```

O p obtido como resultado do teste de Mann-Whitney foi $0,6907$. Este p é maior que o nível de significância ($\alpha = 0,05$) estabelecido para o teste. Portanto, não podemos rejeitar a hipótese nula do teste. Dessa forma, sob um nível de confiança de 95%, não podemos afirmar que o nível de experiência do testador apresenta diferença estatisticamente significativa na densidade de faltas da técnica. A partir deste resultado, não é necessário considerar níveis distintos para as diferentes configurações dos sujeitos para T_3.

B.3 Técnica T_5 – Seleção Aleatória de Casos de Teste

A técnica T_5 fazia uso de três tipos de configuração de cobertura de casos de teste: 25%, 50% e 75%. Estes valores indicam a quantidade de casos de teste que serão selecionados para construir a suíte reduzida. Portanto, uma técnica de seleção aleatória de casos de teste,

com 75% de cobertura, seleciona 75% dos casos de teste da suíte de testes de regressão. A técnica T_5 foi executada com os valores de cobertura especificados para este experimento e a verificação do impacto desta cobertura no desempenho da técnica é descrita nesta Seção. Para esta verificação, considera-se que $T_{5-25\%}$ é a técnica com nível de cobertura de 25%, $T_{5-50\%}$ é a técnica com nível de cobertura de 50% e $T_{5-75\%}$ é a técnica com nível de cobertura de 75%.

Sejam $I(x)$, $P(x)$, $E(x)$, $R(x)$ e $D(x)$,a inclusão, precisão, eficiência, potencial de redução e densidade de faltas de uma técnica x, respectivamente. O teste de hipóteses realizado para a técnica T_5 foi:

Hipóteses Nulas (H_0)

- A técnica T_5 é independente do nível de cobertura no tocante à inclusão, ou seja, $I(T_{5-25\%}) = I(T_{5-50\%}) = I(T_{5-75\%})$;
- A técnica T_5 é independente do nível de cobertura no tocante à precisão, ou seja, $P(T_{5-25\%}) = P(T_{5-50\%}) = P(T_{5-75\%})$;
- A técnica T_5 é independente do nível de cobertura no tocante à eficiência, ou seja, $E(T_{5-25\%}) = E(T_{5-50\%}) = E(T_{5-75\%})$.
- A técnica T_5 é independente do nível de cobertura no tocante ao potencial de redução, ou seja, $R(T_{5-25\%}) = R(T_{5-50\%}) = R(T_{5-75\%})$;
- A técnica T_5 é independente do nível de cobertura no tocante à densidade de faltas, ou seja, $D(T_{5-25\%}) = D(T_{5-50\%}) = D(T_{5-75\%})$.

Hipóteses Alternativas (H_1)

- A técnica T_5 é dependente do nível de cobertura no tocante à inclusão, ou seja, $I(T_{5-25\%}) \neq I(T_{5-50\%}) \neq I(T_{5-75\%})$;
- A técnica T_5 é dependente do nível de cobertura no tocante à precisão, ou seja, $P(T_{5-25\%}) \neq P(T_{5-50\%}) \neq P(T_{5-75\%})$;
- A técnica T_5 é dependente do nível de cobertura no tocante à eficiência, ou seja, $E(T_{5-25\%}) \neq E(T_{5-50\%}) \neq E(T_{5-75\%})$.

- A técnica T_5 é dependente do nível de cobertura no tocante ao potencial de redução, ou seja, $R(T_{5-25\%}) \neq R(T_{5-50\%}) \neq R(T_{5-75\%})$;
- A técnica T_5 é dependente do nível de cobertura no tocante à densidade de faltas, ou seja, $D(T_{5-25\%}) \neq D(T_{5-50\%}) \neq D(T_{5-75\%})$.

Esta análise referente à dependência entre a técnica e a sua configuração foi realizada para cada variável dependente do estudo experimental. Para obter conclusões acerca das hipóteses é necessário realizar um teste estatístico. Pela configuração das hipóteses utilizadas, é recomendada a utilização do teste ANOVA. No entanto, para que este teste seja realizado, diversas premissas devem ser satisfeitas. A primeira delas é que as amostras investigadas possuam distribuição normal. Se as amostras não respeitam esta amostra, devemos partir para a utilização de um teste não-paramétrico[Siegel and Junior 1988]. Portanto, se os dados não seguirem uma distribuição normal, será utilizado o teste de Kruskal-Wallis.

Para a realização do teste estatístico, é estabelecido um nível de confiança de 95%, i.e. um nível de significância α igual a $0,05$. Os cálculos estatísticos são realizados pela ferramenta Minitab, cujos resultados são apresentados durante esta Seção. A verificação das hipóteses levantadas para a técnica T_5 é apresentada a seguir.

B.3.1 Inclusão

Para decidir qual o teste estatístico que seria adequado utilizar, foi efetuado o teste de normalidade de Anderson-Darling. Os resultados deste teste para cada amostra, são exibidos na Figura B.10.

Como, em todos os casos, o p foi menor que α, não foi possível aceitar a hipótese de que os dados seguiam distribuição normal. Parte-se, portanto, para o uso do teste não-paramétrico Kruskall-Walis. O resultado do teste Kruskall-Walis é mostrado a seguir:

```
Kruskal-Wallis Test on Resposta Inclusao

Fator
Inclusao    N  Median  Ave Rank       Z
T5_25     100   25,00      50,7  -14,10
T5_50     100   50,00     150,7    0,03
```

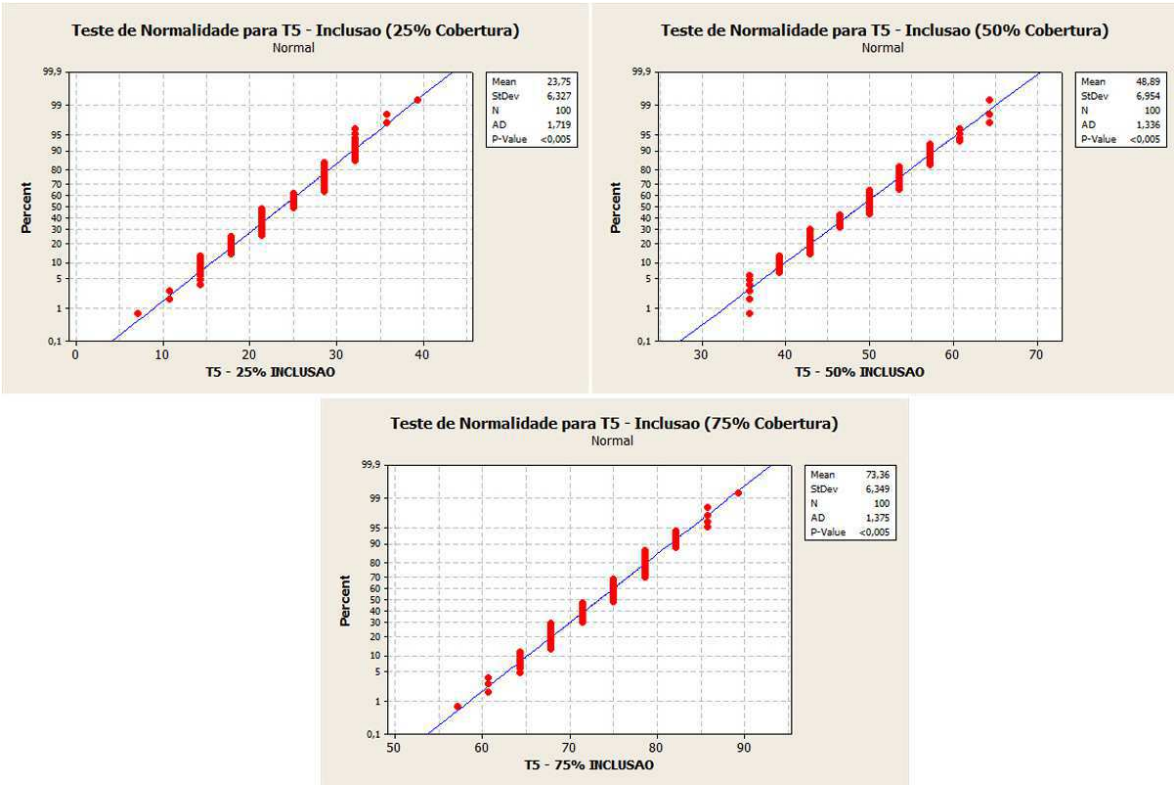

Figura B.10: Resultados dos testes Anderson-Darling para $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$ no tocante à propriedade de inclusão.

```
T5_75     100   75,00      250,1   14,06
Overall   300              150,5

H = 264,34  DF = 2  P = 0,0001
H = 265,20  DF = 2  P = 0,0001  (adjusted for ties)
```

O teste forneceu, como resultado um p igual a $0,0001$, que é menor que o nível de significância estabelecido para o teste. Dessa forma, é possível rejeitar a hipótese nula de que as populações são iguais. Portanto, para a propriedade de inclusão, há diferença, ao nível de confiança de 95%, entre $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$. Conseqüentemente, é possível afirmar que a técnica T_5 é dependente do nível de configuração para a propriedade de inclusão.

B.3.2 Precisão

Inicialmente, foi realizado um teste de normalidade de Anderson-Darling nas amostras. Os resultados destes testes são exibidos na Figura B.11, e são utilizados para decidir o teste estatístico que deve ser utilizado.

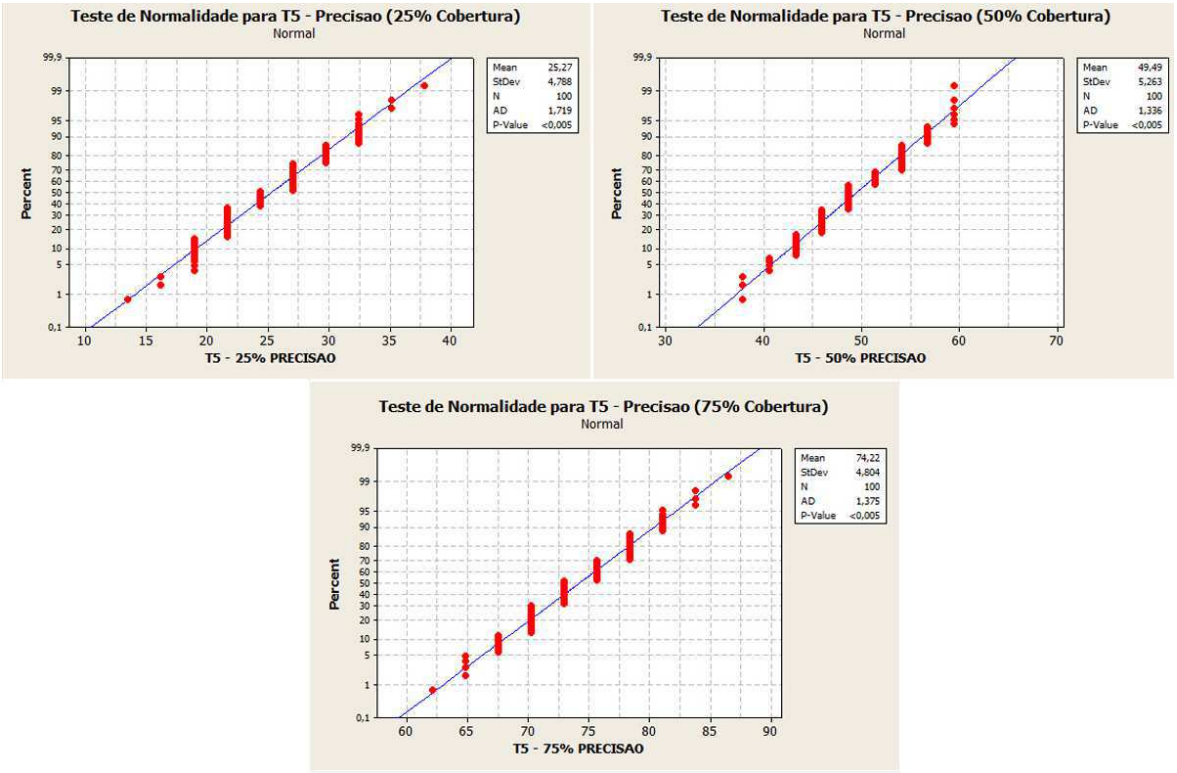

Figura B.11: Resultados dos testes Anderson-Darling para $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$ no tocante à propriedade de precisão.

Considerando que o p foi menor que α, em todos os casos, não podemos aceitar a hipótese de que os dados seguem uma distribuição normal. Diante disto, decidimos utilizar o teste não-paramétrico de Kruskall-Walis, para verificar as hipóteses. O resultado do teste Kruskall-Walis, realizado na ferramenta Minitab, é mostrado a seguir:

```
Kruskal-Wallis Test on Resposta Precisao

Fator
Precisao    N  Median  Ave Rank       Z
T5_25     100   24,32      50,5  -14,12
T5_50     100   48,65     150,5   -0,00
```

```
T5_75     100   72,97     250,5   14,12
Overall   300             150,5

H = 265,74  DF = 2  P = 0,0001
H = 266,59  DF = 2  P = 0,0001  (adjusted for ties)
```

O p obtido foi igual a $0,0001$, que é menor que α. Com isto, é possível rejeitar a hipótese nula. Portanto, para a propriedade de precisão, há diferença, ao nível de confiança de 95%, entre $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$. Conseqüentemente, é possível afirmar que a técnica T_5 é dependente do nível de configuração para a propriedade de precisão.

B.3.3 Eficiência

Assim como nas demais investigações desta Seção, o primeiro passo foi a realização de testes de normalidade de Anderson-Darling para verificar se os dados seguiam uma distribuição normal. A Figura B.12 apresenta os gráficos com os resultados desses testes.

Observando o p obtido, não foi possível aceitar a hipótese de que os dados seguiam distribuição normal. a partir deste resultado, decidimos, então, usar o teste estatístico de Kruskal-Wallis. Este teste foi aplicado nos dados de eficiência utilizando a ferramenta Minitab e os resultados são apresentados a seguir:

```
Kruskal-Wallis Test on Resposta Eficiencia

Fator
Eficiencia    N  Median  Ave Rank       Z
T5_25       100   10541     222,6   10,18
T5_50       100    7667     149,5   -0,14
T5_75       100    4792      79,4  -10,04
Overall     300             150,5

H = 136,35  DF = 2  P = 0,0001
H = 136,86  DF = 2  P = 0,0001  (adjusted for ties)
```

O p obtido foi igual a $0,0001$, que é menor que α, possibilitando a rejeição a hipótese nula. Portanto, para a propriedade de eficiência, há diferença, ao nível de confiança de 95%,

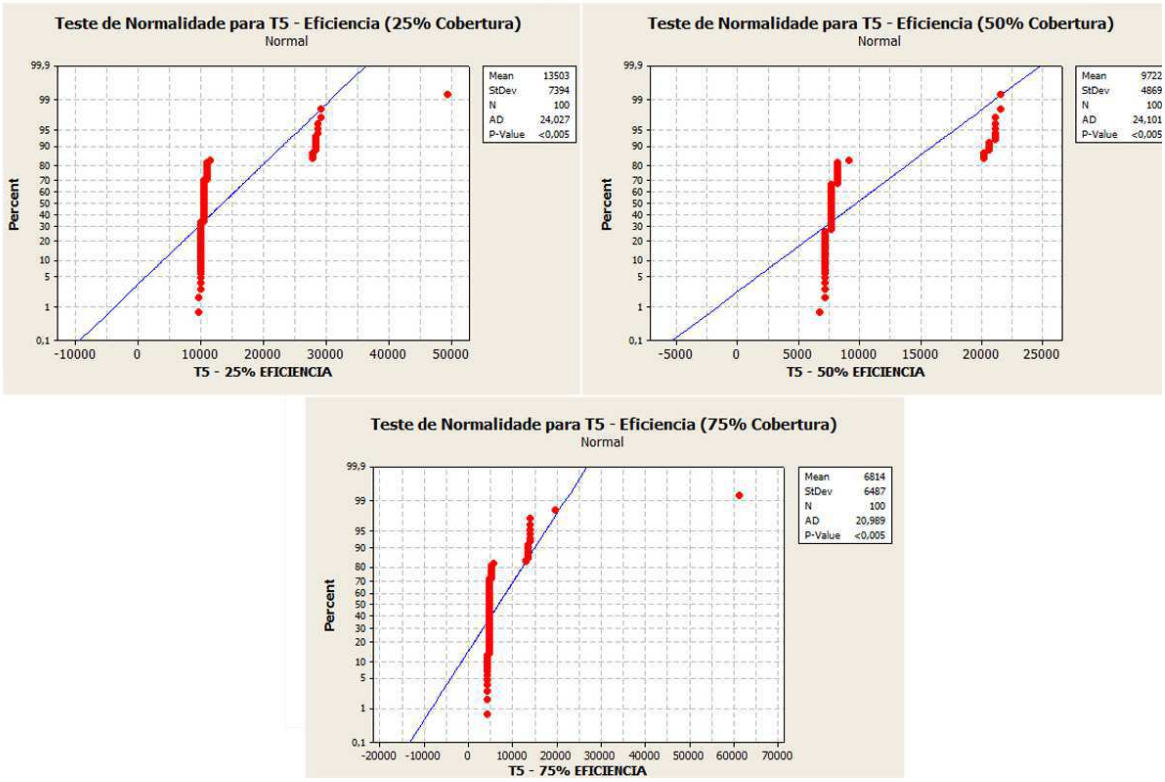

Figura B.12: Resultados dos testes Anderson-Darling para $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$ no tocante à propriedade de eficiência.

entre $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$. Conseqüentemente, é possível afirmar que a técnica T_5 é dependente do nível de configuração para a propriedade de eficiência.

B.3.4 Potencial de Redução

A variável de potencial de redução caracteriza, para esta técnica, o próprio parâmetro de configuração. Diante disto, nao foi necessário realizar análises para verificar se a técnica se comporta de forma diferente de acordo com o potencial de redução, pois o próprio parâmetro de cobertura caracteriza o potencial de redução da técnica. Portanto, podemos aceitar a hipótese de que a técnica apresenta resultados de potencial de redução diferentes de acordo com a configuração fornecida.

B.3.5 Densidade de Faltas

Os dados de densidade de faltas da técnica com as diferentes configurações de cobertura foram submetidos a um teste de normalidade de Anderson-Darling. Os gráficos resultantes desta análise são exibidos na Figura B.13.

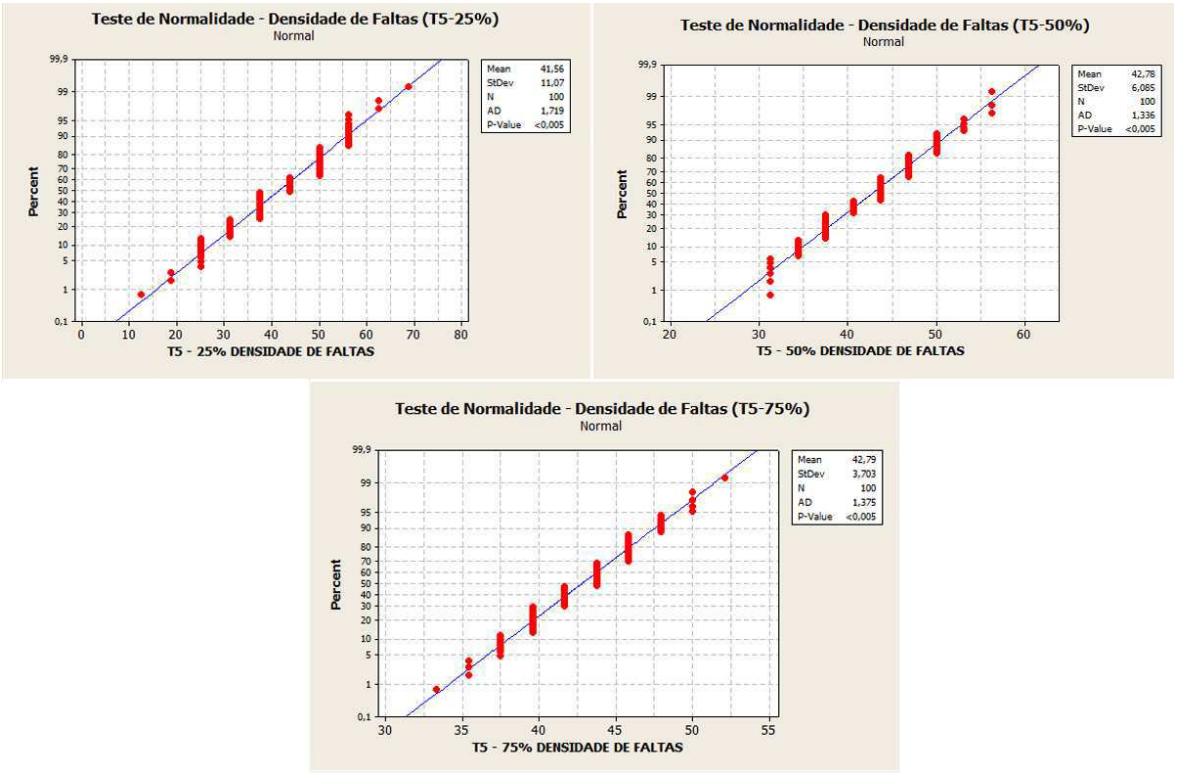

Figura B.13: Resultados dos testes Anderson-Darling para $T_{5-25\%}$, $T_{5-50\%}$ e $T_{5-75\%}$ no tocante à propriedade de densidade de faltas.

Observando o p resultante do teste, verificamos que não é possível aceitar a hipótese de que as amostras seguem uma distribuição normal. Dessa forma, não podemos utilizar ANOVA e partimos para o teste não-paramétrico correspondente, o teste de Kruskal-Wallis. Este teste foi aplicado utilizando ferramenta Minitab, e os resultados são apresentados a seguir:

```
Kruskal-Wallis Test on Densidade de Faltas.

Fator:
Reducao/Faltas    N  Median  Ave Rank      Z
```

```
T5_25           100   43,75     144,3  -0,87
T5_50           100   43,75     153,6   0,44
T5_75           100   43,75     153,5   0,43
Overall         300             150,5

H = 0,76  DF = 2  P = 0,684
H = 0,77  DF = 2  P = 0,680  (adjusted for ties)
```

Como resultado do teste, foi obtido um p igual a $0,684$. Uma vez que o p obtido é maior que o nível de significância estabelecido para o teste, não podemos rejeitar a hipótese nula. Portanto, não podemos afirmar que os diferentes percentuais de cobertura da técnica apresentam diferenças estatisticamente significativas na densidade de faltas da técnicas. Dessa forma, para o estudo experimental, será considerado apenas um nível T_5, constituído pela média aritmética dos dados de T_5 com 25%, 50% e 75% de cobertura.

Apêndice C

Investigação das Premissas de ANOVA

Considerando que os projetos experimentais descritos neste trabalho consistem de um único fator (técnica de re-teste seletivo) com vários níveis categóricos (cada uma das técnicas utilizadas no estudo experimental), o método indicado para verificar a semelhança entre estes níveis é a realização de uma Análise de Variância (ANOVA). No entanto, para aplicar ANOVA, é necessário que os dados obtidos respeitem algumas premissas. São elas:

- As amostras devem possuir distribuição normal;
- Os erros possuem um distribuição normal;
- A população amostrada é homoscedástica;
- Os erros são independentes dos níveis do fator;
- Os erros possuem uma mesma variância para os diversos níveis do fator [Jain 1991].

Os dados obtidos com a execução do experimento foram dispostos na ferramenta Minitab. A partir do uso dessa ferramenta, foi possível realizar a investigação acerca das premissas apresentadas. Todos os aspectos desta Seção foram investigados utilizado um nível de confiança de 95% ($\alpha = 0,05$). Os resultados desta investigação, organizados por projeto experimental, são apresentados a seguir.

C.1 Premissas de ANOVA - Projeto Experimental 1

Os primeiros dados analisados correspondem ao Projeto Experimental 1, que investiga a propriedade de *inclusão* de cada técnica de re-teste seletivo executada no experimento descrito na Seção 5.10.2. Foram realizadas 100 replicações do experimento, e os dados para os 8 níveis ($T_1, T_2, T_{3i}, T_{3e}, T_4, T_{5-25\%}, T_{5-50\%}$ e $T_{5-75\%}$) do fator foram armazenados. Estes dados foram submetidos à ferramenta Minitab, que ao realizar a análise das premissas da ANOVA apresentou os gráficos da Figura C.1.

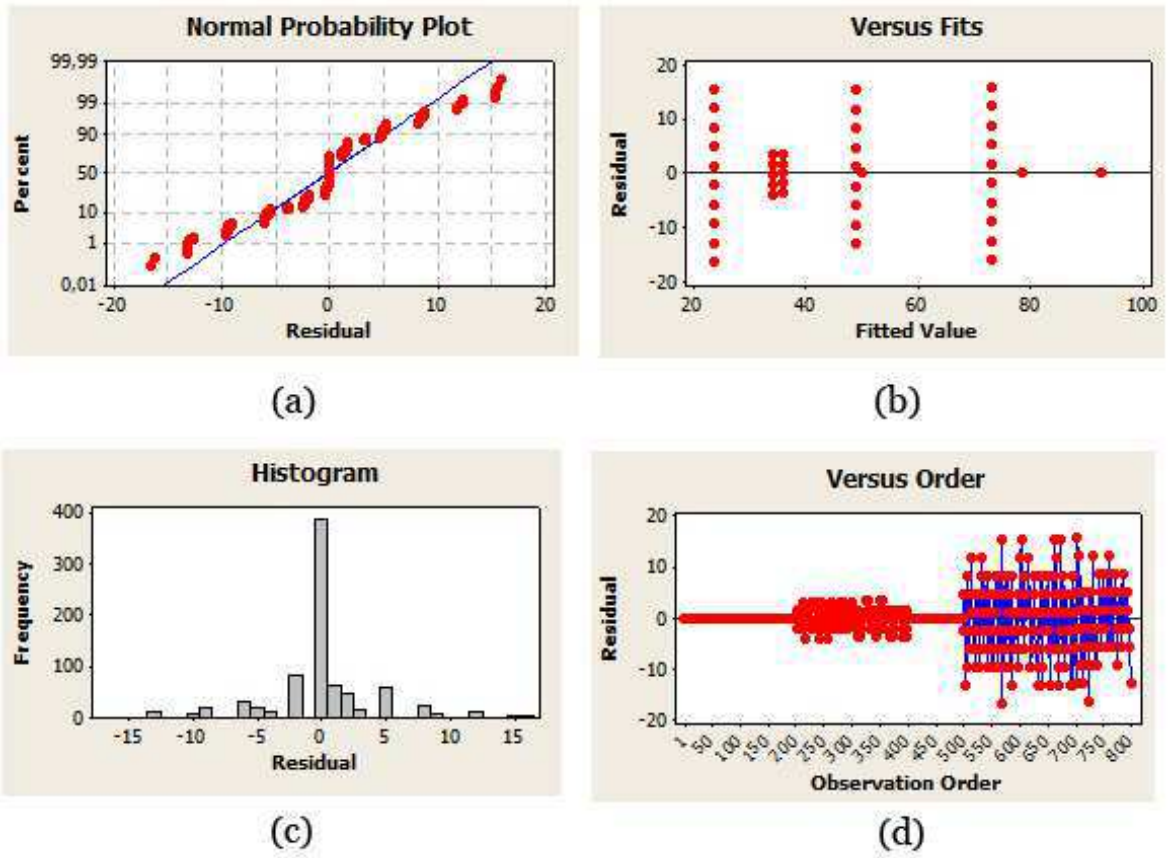

Figura C.1: Investigação da adequação dos dados residuais às premissas de ANOVA no tocante à inclusão.

A primeira premissa a ser obedecida é que os dados de cada amostra devem seguir uma distribuição normal. Realizando um teste de Anderson-Darling, foi possível observar, como apresentado no Apêndice A, que os dados não seguem uma distribuição normal, e portanto, ANOVA não poderia ser utilizado nesta análise. No entanto, para prover uma visão geral das

propriedades dos dados, as demais premissas foram também verificadas.

A partir dos gráficos da Figura C.1, é possível observar se as premissas apresentadas no início deste apêndice são respeitadas. O primeiro gráfico da figura, indica a distribuição dos resíduos dos dados. Pelo gráfico da Figura C.1 (a), os resíduos dos dados (pontos vermelhos no gráfico) deveriam estar dispostos sobre a linha azul (linha que indica a distribuição normal, neste gráfico). Porém, é possível observar que os resíduos não estão dispostos sobre esta linha, e portanto, os resíduos dos dados não seguem uma distribuição normal, o que viola uma das premissa de ANOVA.

Uma outra premissa está relacionada com a homoscedasticidade dos dados. O gráfico da Figura C.1 (b) apresenta o espalhamento dos resíduos dos dados. Para que os dados sejam homoscedásticos, os resíduos deste gráfico devem estar espalhados de forma semelhante. No entanto, no gráfico da Figura C.1 (b), este espalhamento é heterogêneo e indica que os dados analisados não são homoscedásticos, violando, portanto, outra premissa da ANOVA.

Na Figura C.1 (c), verificamos o histograma dos erros. Pelo histograma, podemos verificar que os erros possuem uma distribuição pouco próxima da distribuição normal. No entanto, alguns valores de erros são mais recorrentes que os demais (e.g. $erro = 0$), o que pode comprometer a premissa de que os erros são independentes dos níveis dos fatores. O gráfico da Figura C.1 (d), apresenta a ordem das observações. Este gráfico relaciona a independência e a homoscedasticidade dos resíduos. Para que as observações sejam independentes, este gráfico não deve apresentar tendências. No entanto, alguns trechos do gráfico não apresentam variância (o que evidencia uma tendência nos dados), ou seja, as premissas de homoscedasticidade e independência dos erros não foram respeitadas.

Como foi possível constatar, os dados do experimento não respeitam as premissas da ANOVA, portanto, não é adequado utilizá-la para a análise deste experimento. Diante disto, para obter o resultado dos testes de hipóteses, é necessária a utilização do teste não-paramétrico correspondente à ANOVA, o teste de Kruskal-Wallis.

C.2 Premissas de ANOVA - Projeto Experimental 2

Esta subseção apresenta a verificação realizada para observar se os dados obtidos sobre a *precisão* das técnicas respeitam as premissas da ANOVA. Foram realizadas 100 execuções

em cada um dos 8 níveis do fator ($T_1, T_2, T_{3i}, T_{3e}, T_4, T_{5-25\%}, T_{5-50\%}$ e $T_{5-75\%}$). Estes dados foram então submetidos à ferramenta Minitab, para realizar a verificação das premissas de ANOVA. Os resultados desta verificação podem ser observados na Figura C.2.

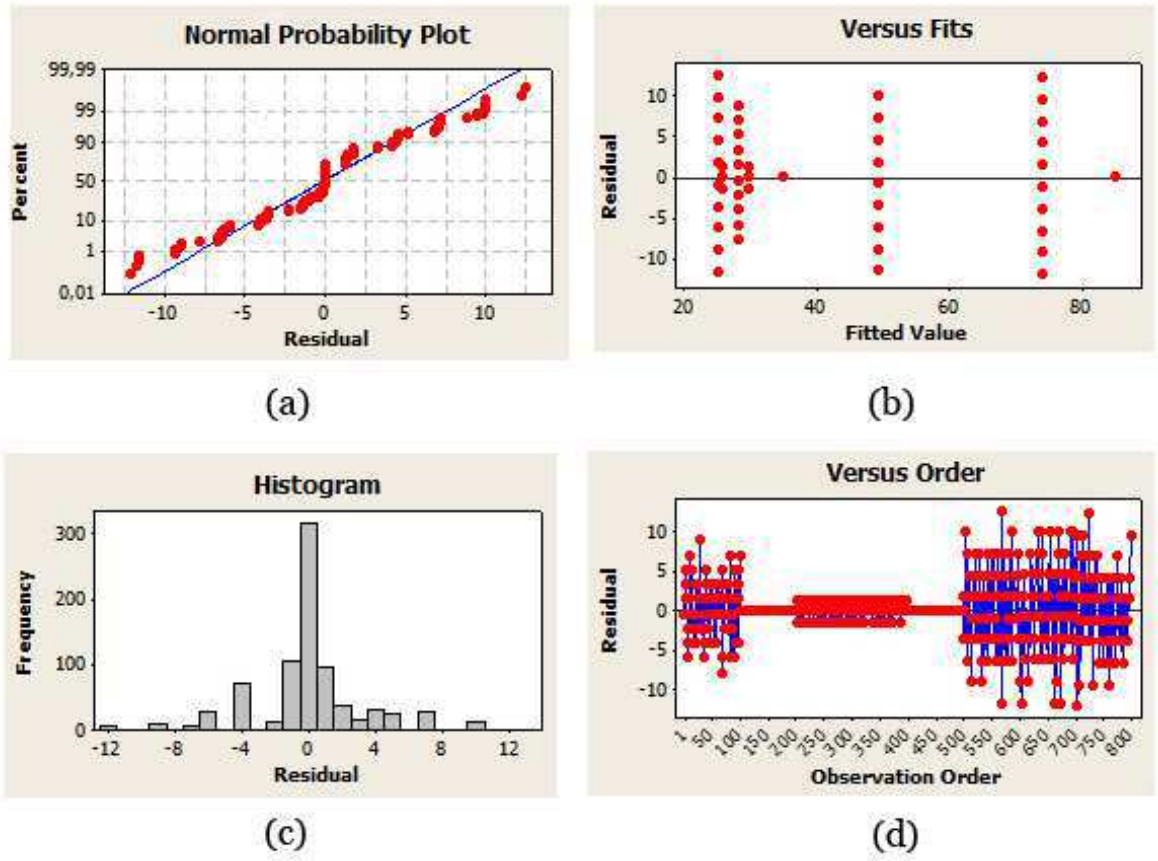

Figura C.2: Investigação da adequação dos dados residuais às premissas de ANOVA no tocante à precisão.

De acordo com as premissas de ANOVA, é necessário que os dados possuam uma distribuição normal. Como verificamos a partir dos testes de Anderson-Darling (Apêndice A), essa premissa não é respeitada, pois os dados não apresentaram distribuição normal. Verificando os gráficos da Figura C.2, é possível observar que outras premissas de ANOVA também não são respeitadas.

Na Figura C.2 (a), podemos observar que os resíduos (pontos vermelhos no gráfico) não seguem uma distribuição normal, pois não estão sobre a linha azul do gráfico. A homoscedasticidade dos erros é verificada no gráfico da Figura C.2 (b). De acordo com este gráfico, os

erros dos dados não apresentam homoscedasticidade (i.e. possuem um espalhamento heterogêneo), e portanto, outra premissa da ANOVA não é respeitada.

Na Figura C.2 (c), verificamos um histograma dos erros. Apesar de apresentar uma distribuição pouco próxima da distribuição normal, este histograma realça a frequência alta de alguns erros (e.g. zero), o que viola a premissa de que os erros são independentes dos níveis do fator. Observando o gráfico da Figura C.2 (d), podemos verificar a independência e a homoscedasticidade dos resíduos. Como alguns trechos deste gráfico apresentam algumas tendências, não podemos afirmar que os erros são homoscedásticos e independentes.

A partir dos aspectos analisados nos gráficos da Figura C.2, não podemos aplicar ANOVA no experimento de precisão, pois diversas das premissas necessárias para a realização do teste de ANOVA não foram respeitadas. Diante disto, decidimos aplicar o teste de Kruskal-Wallis, que é o teste não-paramétrico correspondente ao teste de ANOVA.

C.3 Premissas de ANOVA - Projeto Experimental 3

As premissas de ANOVA também foram verificadas para os dados de *eficiência*. Os dados das 100 repetições dos 9 níveis do fator ($T_1, T_{2i}, T_{2e}, T_{3i}, T_{3e}, T_4, T_{5-25\%}, T_{5-50\%}$ e $T_{5-75\%}$), foram submetidos à ferramenta Minitab para verificar as premissas de ANOVA. Os resultados gerados pelo Minitab são apresentados na Figura C.3.

Realizando um teste de Anderson-Darling para um teste de normalidade nas amostras de cada nível (Apêndice A, verificamos que nenhuma delas segue a distribuição normal. Desta forma, a primeira das premissas de ANOVA não é respeitada, inviabilizando a aplicação deste teste no Projeto Experimental 3. Verificando o gráfico da Figura C.3 (a), observamos que outra premissa da ANOVA é violada. Neste gráfico, é possível verificar que os erros das amostras não seguem uma distribuição normal.

Observando o gráfico da Figura C.3 (b), verificamos que não há homoscedasticidade nos erros obtidos a partir dos dados de eficiência. A ausência de homoscedasticidade é verificada pela heterogeneidade dos resíduos plotados no gráfico. Observando o histograma da Figura C.3 (c), verificamos que os erros não são independentes dos níveis do fator, pois alguns resíduos possuem uma frequência muito alta. Na Figura C.3 (d), o gráfico que exibe a ordem das observações apresenta algumas tendências em trechos do gráfico. Estas tendências

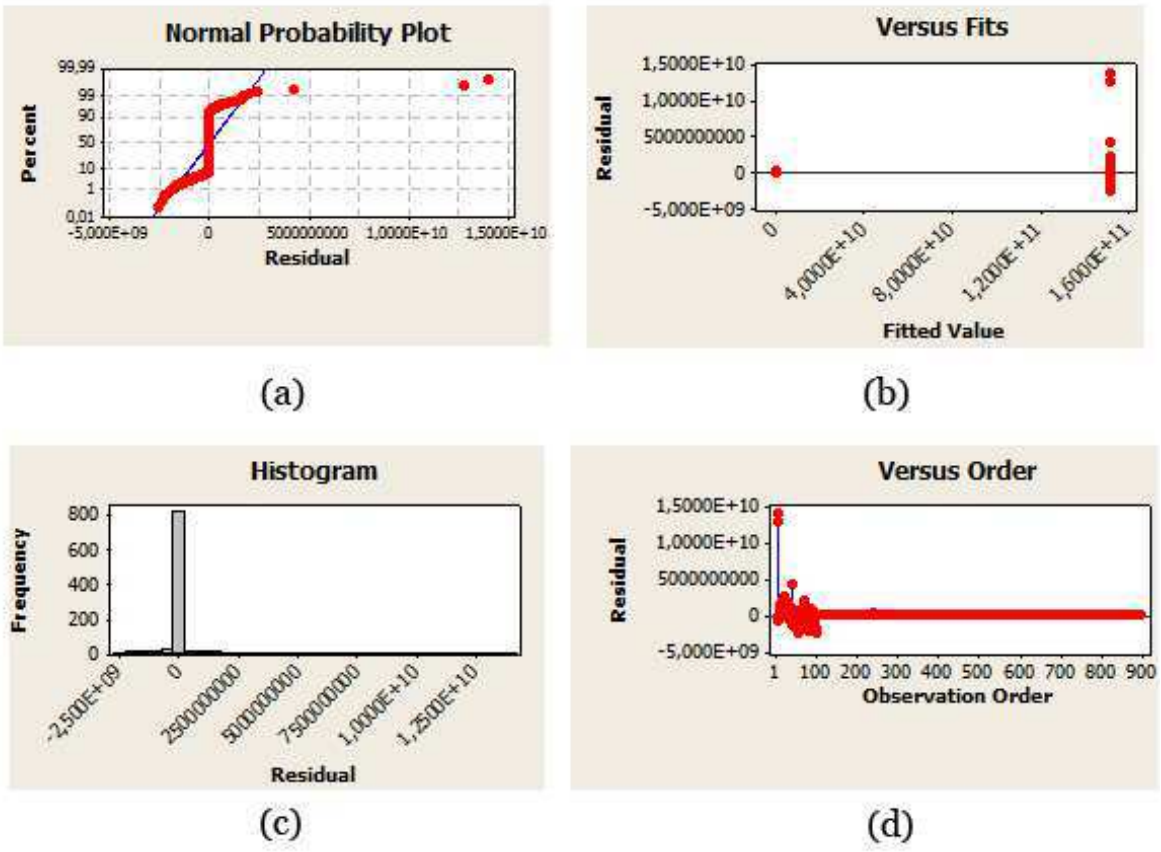

Figura C.3: Investigação da adequação dos dados residuais às premissas de ANOVA no tocante à eficiência.

indicam que os erros não são homoscedásticos nem independentes.

Portanto, após verificar os gráficos da Figura C.3, observamos que diversas premissas da ANOVA não são respeitadas pelos dados. Diante disto, não é adequado utilizarmos o teste de ANOVA para o Projeto Experimental 3. A partir destas observações decidimos utilizar o teste não-paramétrico de Kruskal-Wallis para a análise da eficiência das técnicas.

C.4 Premissas de ANOVA - Projeto Experimental 4

O dados do *potencial de redução* das técnicas foram investigados para verificar se estes violam as premissas de ANOVA. Estes dados correspondem aos 8 níveis do fator investigado ($T_1, T_2, T_{3i}, T_{3e}, T_4, T_{5-25\%}, T_{5-50\%}$ e $T_{5-75\%}$). Foram realizadas 100 execuções de

cada técnica, e os resultados obtidos foram submetidos à ferramenta Minitab. Os gráficos dos resíduos são apresentados na Figura C.4.

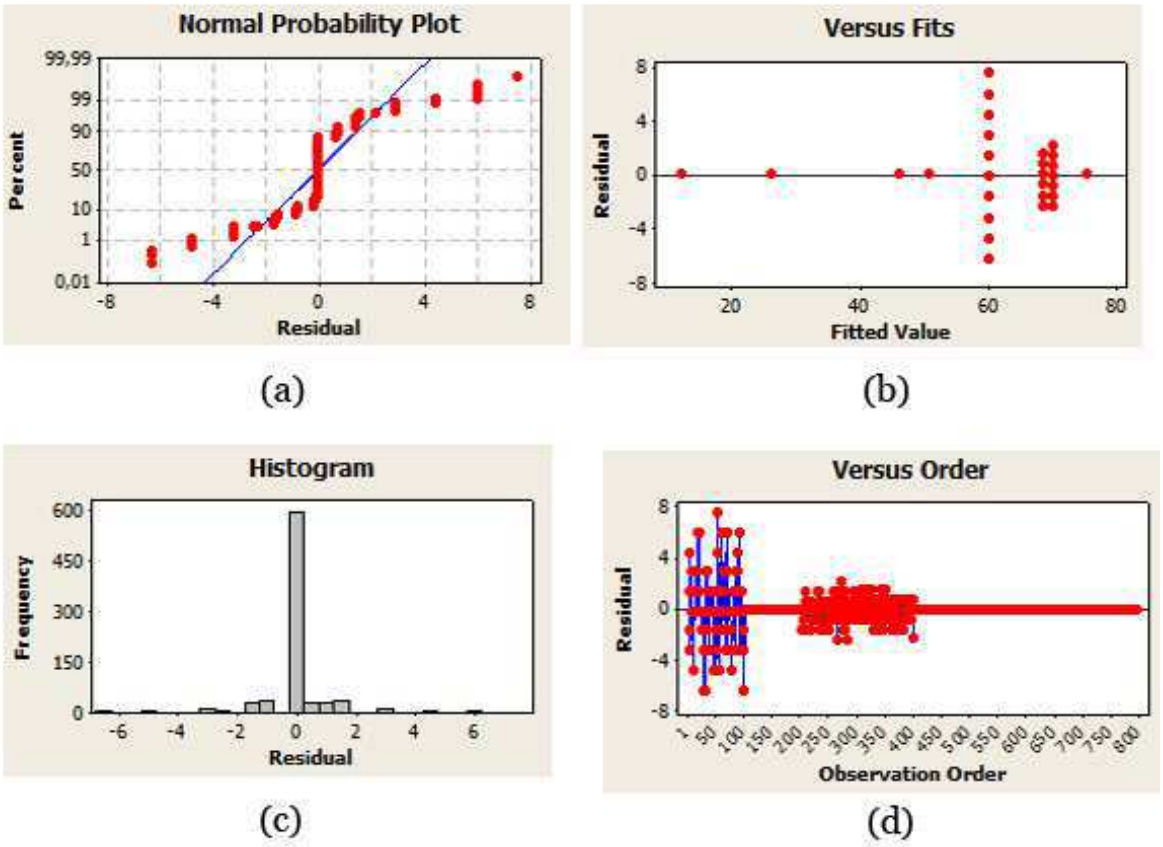

Figura C.4: Investigação da adequação dos dados residuais às premissas de ANOVA no tocante ao potencial de redução.

A partir do teste de normalidade de Anderson-Darling realizado nos dados (Apêndice A), verificamos que os dados não seguem uma distribuição normal, o que já viola uma das premissas de ANOVA. O próximo aspecto analizado é a distribuição dos resíduos dos dados. Esta dsitribuição pode ser observada no gráfico da Figura C.4 (a). Uma vez que os resíduos não estão dispostos sobre a linha azul, correspondente à distribuição normal, não podemos, então, afirmar que os resíduos são normalmente distribuídos. Ainda sob esta perspectiva, os resíduos dos dados não apresentam homoscedasticidade, pois estão dispostos de forma heterogênea no gráfico da Figura C.4 (b). Diante destes aspectos, duas premissas de ANOVA são violadas.

Na Figura C.4 (c), observamos o histogramas dos resíduos. Assim, como nos demais projetos experimentais observados, este histograma revela que alguns erros são mais frequentes que os demais, violando a premissa de independência entre os erros e os níveis do fator. Para concluir a análise das premissas de ANOVA no projeto experimental 4, observamos o gráfico da Figura C.4 (d), que apresenta os resíduos e a ordem das observações. As tendências presentes neste gráfico, indicam que não há homoscedasticidade nem independência nos erros, violando, portanto, outra das premissas de ANOVA.

A partir dos aspectos analisados, as premissas de ANOVA não foram respeitadas pelos dados de potencial de redução. Portanto, não é adequado utilizar ANOVA na etapa de análise dos dados. Esta característica indica que o investigador deve utilizar o teste não-paramétrico de Kruskal-Wallis, para obter os resultados apropriados dos testes de hipóteses.

C.5 Premissas de ANOVA - Projeto Experimental 5

O projeto Experimental 5 também foi analisado, para verificar se os dados de *densidade de faltas* das técnicas violam as premissas de ANOVA. Diante disto, os resíduos dos dados obtidos das 100 execuções dos 5 níveis do fator (T_1, T_2, T_3, T_4, e T_5) foram submetidos à ferrmanta Minitab, e os resultados são apresentados no gráfico da Figura C.5.

O primeiro aspecto observado é se os próprios dados seguem uma distribuição normal. Como verificado no Apêndice A, através dos testes de normalidade de Anderson-Darling, apenas os dados de densidade das técnicas T_1 e T_5 seguem uma distribuição normal. Porém, para que a premissa de ANOVA seja satisfeita, os dados de todos os níveis devem seguir uma distribuição normal, e como os dados de T_2, T_3 e T_4 não são normalmente distribuídos, esta premissa não é satisfeita.

Assim como foi realizado nos demais projetos experimentais, passamos, em seguida para a análise dos resíduos dos dados. Inicialmente, observamos o gráfico da Figura C.5 (a) para verificar se os resíduos seguem uma distribuição normal. Como estes não estão posicionados sobre a linha azul, podemos concluir que os resíduos dos dados não são normalmente distribuídos. O próximo passo é verificar a homoscedasticidade dos dados na Figura C.5 (b). Observando a disposição heterogênea dos resíduos no gráfico, não podemos afirmar que os dados são homoscedásticos. A partir destas duas análises iniciais, verificamos que duas

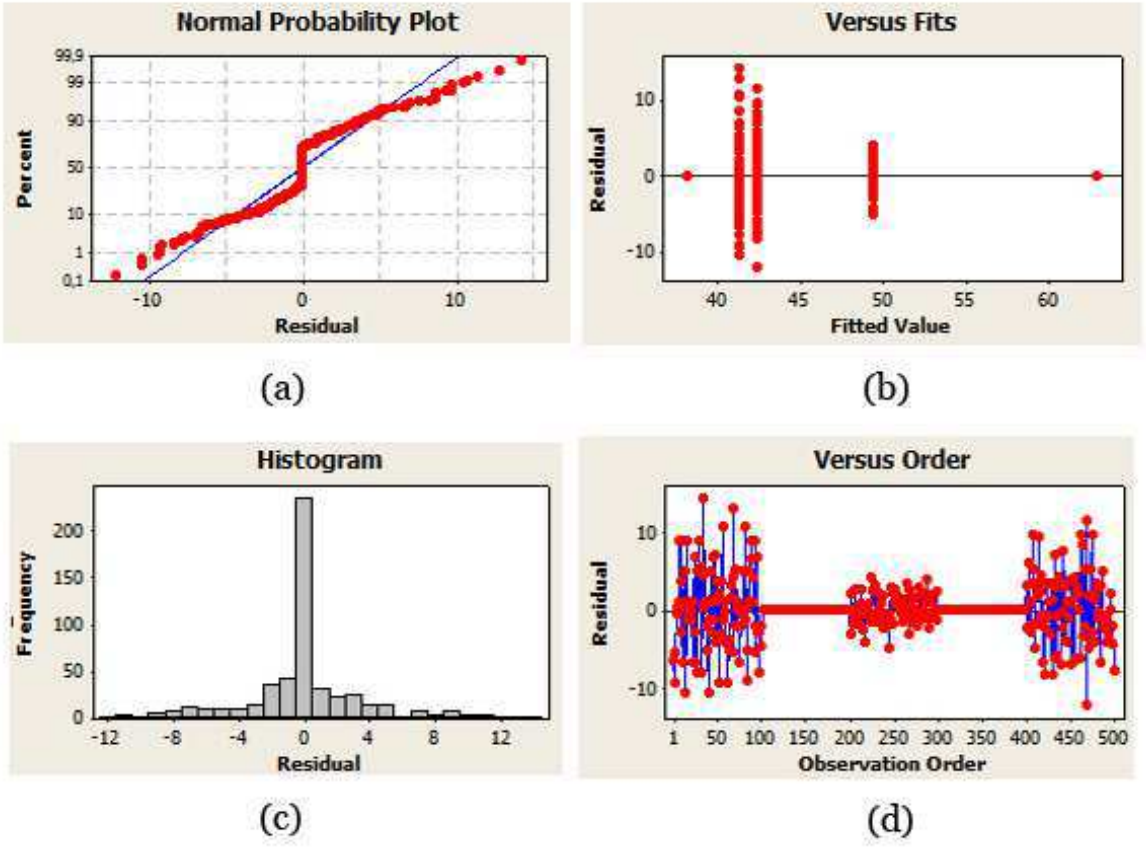

Figura C.5: Investigação da adequação dos dados residuais às premissas de ANOVA no tocante à densidade de faltas.

premissas de ANOVA são violadas pelos dados.

O próximo elemento analisado é a independência dos resíduos com relação aos níveis do fator. Observando o histograma da Figura C.5 (c), verificamos que alguns valores são mais frequentes que os demais, indicando que os resíduos não são independentes dos níveis. Por fim, observamos os resíduos pelas ordem em que foram observados, e novamente, verificamos que alguns trechos apresentam variância nula, o que indica a presença de tendências neste gráfico. Diante disto não podemos afirmar que os resíduos são homoscedásticos e independentes dos níveis do fator.

Após verificar os gráficos da Figura C.5 observamos que as premissas de ANOVA não são respeitadas pelos dados de densidade de faltas das técnicas. Portanto, não é adequada a utilização de ANOVA para o teste de hipóteses. Diante disto, decidimos utilizar o teste de

Kruskal-Wallis para obter os resultados das hipóteses estruturadas nos projetos experimentais de cada variável dependente.

Apêndice D

Análise de Desempenho das Técnicas

Neste apêndice é apresentada a análise realizada com as técnicas, para obter um desempenho comparativo das técnicas de re-teste seletivo analisadas neste estudo experimental. Esta análise é fundamentada nos resultados descritos no Capítulo 7, assim como é utilizada para complementar a análise realizada no referido capítulo. Cada seção deste apêndice descreve a análise para uma variável dependente analisada.

D.1 Desempenho das Técnicas – Inclusão

Após verificar (através do teste de Kruskal-Wallis) que as técnicas são diferentes entre si, procuramos então comparar o desempenho de inclusão de cada técnica, procurando identificar a técnica com melhor inclusão. De acordo com os dados observados, as técnicas T_1, T_2 e T_4 possuíram variância nula e médias 50; $92,857$ e $78,571$, respectivamente. Dados sobre inclusão de nenhuma outra técnica observada apresentaram comportamento similar (i.e. variância nula). Para auxiliar na definição de hipóteses sobre a inclusão das técnicas, um intervalo de confiança dos dados foi construído, como ilustrado na Figura D.1.

Observando os intervalos de confiança, verificamos que T_2 apresentou o melhor resultado de inclusão, seguido por T_4. Portanto, como não há sobreposição, entre os intervalos de confiança de T_2 e T_4, e a média de $T_2 > T_4$, podemos afimar que T_2 possui um desempenho melhor que T_4.

O próximo Intervalo de confiança com melhor desempenho é o de $T_{5-75\%}$. A disposição deste Intervalo de Confiança, quando comparado com o de T_4, indica que $T_4 > T_{5-75\%}$, uma

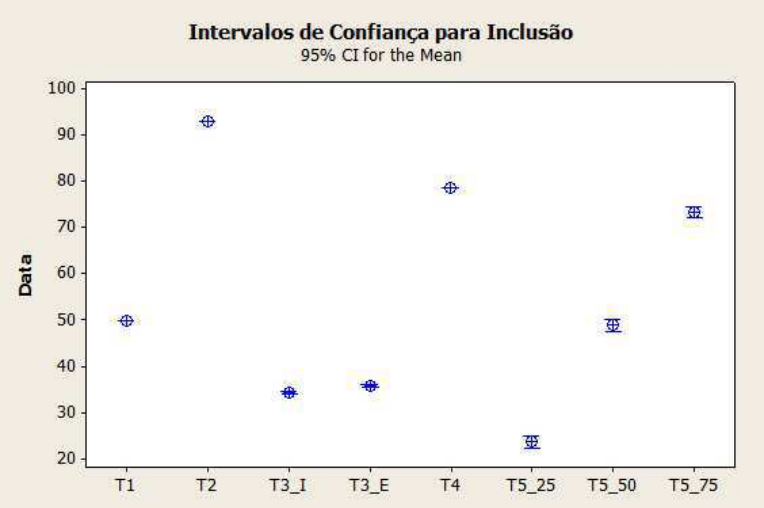

Figura D.1: Intervalos de confiança para a variável dependente inclusão.

vez que o intervalo de confiança de $T_5 - 75\%$ é $[72.097, 74.616]$ e não contém 75.571, a média de T_4.

De acordo com os Intervalos de Confiança, é necessário investigar o desempenho de T_1 e $T_{5-50\%}$. Observando os intervalos de confiança destas duas técnicas, verificamos que a média de T_1 (50) encontra-se no intervalo de confiança de $T_{5-50\%}$ ($[47.512, 50.272]$). Logo, não é possível afirmar que T_1 e $T_{5-50\%}$ são diferentes.

Para as demais técnicas, uma vez que não há uma clara separação dos intervalos de confiança, é preciso realizar o teste estatístico Mann-Whitney. As técnicas foram ordenadas a partir da disposição dos intervalos de confiança na Figura D.1, e então, foi estabelecida a seguinte ordem entre as técnicas.

$$T_{5-50\%} > T_{3e} > T_{3i} > T_{5-25\%}$$

A partir desta ordenação foram estabelecidas as seguintes hipóteses nulas e alternativas:

$H0_A : T_{5-50\%} = T_{3e}$ $\quad$ $H1_A : T_{5-50\%} > T_{3e}$

$H0_B : T_{3e} = T_{3i}$ $\quad$ $H1_B : T_{3e} > T_{3i}$

$H0_C : T_{3i} = T_{5-25\%}$ $\quad$ $H1_C : T_{3i} > T_{5-25\%}$

Estas hipóteses foram verificadas e os resultados de cada verificação segue abaixo.

D.1.1 Verificação de $H0_A$ e $H1_B$

O desempenho de inclusão entre $T_{5-50\%}$ e T_{3e} é verificado pelas hipóteses $H0_A$ e $H1_A$. Lembrando que a partir do teste de Anderson-Darling, verificamos que os dados de $T_{5-50\%}$ e T_{3e} (Apêndice A) não seguem uma distribuição normal, utilizamos o teste não-paramétrico de Mann-Whintey da ferramenta Minitab para o teste de hipóteses. Os dados deste teste, com um nível de significância (α) de $0,05$ são apresentados a seguir:

```
Mann-Whitney Test and CI: T5_50; T3_E

         N  Median
T5_50  100  50,000
T3_E   100  35,714

Point estimate for ETA1-ETA2 is 14,286 95,0 Percent CI for ETA1-ETA2 is (12,500;14,286)
W = 14699,0
Test of ETA1 = ETA2 vs ETA1 > ETA2 is significant at 0,00001
The test is significant at 0,00001 (adjusted for ties)
```

Como podemos ver no resultado do teste de Mann-Whitney, o p (p = 0,00001) é menor que o nível de significância ($\alpha = 0,05$). Portanto podemos rejeitar a hipótese nula, com 95% de nível de confiança em favor da hipótese alternativa. Ou seja, para um nível de confiança de 95%, podemos afirmar que $T_{5-50\%}$ possui melhor inclusão que T_{3e}.

D.1.2 Verificação de $H0_B$ e $H1_B$

A partir das hipóteses $H0_B$ e $H1_B$, podemos investigar qual técnica dentre T_{3e} e T_{3i} possui um melhor desempenho de inclusão. Verificando que tanto T_{3e} e T_{3i} apresentam uma variância e que não possuem uma distribuição normal (Apêndice A), aplicamos o Teste de Mann-Whitney para o teste de hipóteses. Os resultados do teste, para um nível de significância (α) de $0,05$ são apresentados a seguir:

```
Mann-Whitney Test and CI: T3_E; T3_I
```

```
        N  Median
T3_E  100  35,714
T3_I  100  33,929

Point estimate for ETA1-ETA2 is 1,786
95,0 Percent CI for ETA1-ETA2 is (1,786;1,786)
W = 12364,5
Test of ETA1 = ETA2 vs ETA1 > ETA2 is significant at 0,00001
The test is significant at 0,00001 (adjusted for ties)
```

A partir dos resultados obtidos no teste de Mann-Whitney, para a comparação entre T_{3e} e T_{3i}, observamos que o p (p = 0,00001) é menor que o nível de significância estabelecido. Portanto, podemos rejeitar, com 95% de nível de confiança a hipótese nula, em favor da hipótese alternativa. Ou seja, T_{3e} possui um desempenho de inclusão melhor que T_{3i}.

D.1.3 Verificação de $H0_C$ e $H1_C$

O desempenho de inclusão entre T_{3i} e $T_{5-25\%}$ é verificado pelas hipóteses $H0_C$ e $H1_C$. Partimos então para o teste de não-paramétrico de Mann-Whitney, uma vez que as amostras de T_{3i} e $T_{5-25\%}$ não seguem uma distribuição normal (Apêndice A). Os resultados do teste, para um nível de significância (α) de $0,05$ são apresentados a seguir:

```
Mann-Whitney Test and CI: T3_I; T5_25

         N  Median
T3_I   100  33,929
T5_25  100  25,000

Point estimate for ETA1-ETA2 is 10,714
95,0 Percent CI for ETA1-ETA2 is (8,930;12,501)
W = 14651,0
Test of ETA1 = ETA2 vs ETA1 > ETA2 is significant at 0,00001
The test is significant at 0,00001 (adjusted for ties)
```

O p obtido como resultado do teste ($p = 0,00001$) é menor que o nível de sigificância ($\alpha = 0,05$) estabelecido para o teste. Diante disto, rejeitamos a hipótese nula, em favor da

hipótese alternativa. Portanto, podemos afirmar que, para um nível de confiança de 95%, a técnica T_{3i} possui um desempenho de inclusão melhor que a técnica $T_{5-25\%}$.

D.1.4 Conclusões sobre o Desempenho de Inclusão

Como foi possível observar nos testes realizados com o desempenho de inclusão entre os pares de técnicas, podemos concluir que as técnicas estão ordenadas quanto ao desempenho de inclusão, da seguinte forma:

$$T_2 > T_4 > T_{5-75\%} > T_1 = T_{5-50\%} > T_{3e} > T_{3i} > T_{5-25\%}$$

Onde, a técnica com a melhor inclusão foi T_2, enquanto que a técnica que apresentou o menor desempenho em inclusão, dentre as analisadas, foi $T_{5-25\%}$.

D.2 Desempenho das Técnicas – Precisão

A partir dos resultados observados no teste de Kruskal-Wallis para a precisão das técnicas (Seção 7.3), verificamos a necessidade de analisar o desempenho comparativo da precisão das técnicas. Diante disto, os dados foram investigados, e foram construídos Intervalos de Confiança, que permitisse a construção de hipóteses baseadas no desempenho das técnicas. Os Intervalos de Confiança construídos podem ser observados na Figura D.2.

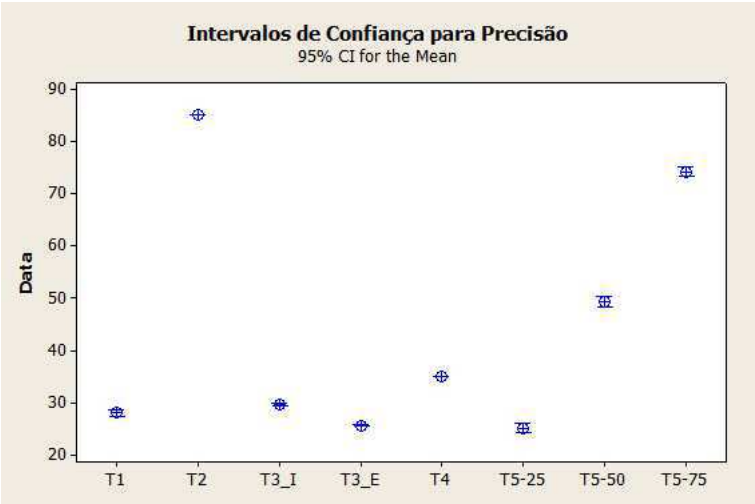

Figura D.2: Intervalos de confiança para a variável dependente precisão.

De acordo com os dados observados, as técnicas T_2 e T_4 possuíram variância nula e médias $85,135$ e $35,135$, respectivamente. Dados sobre precisão de nenhuma outra técnica observada apresentaram comportamento similar. De acordo com a disposição destes intervalos de confiança, é possível afirmar que $T_2 > T_{5-75\%} > T_{5-50\%} > T_4 > T_{3i}$, uma vez que não há sobreposição entre estes intervalos de confiança. No entanto, para as demais técnicas, é preciso realizar o teste Mann-Whitney.

Para as demais técnicas, e a partir da disposição dos intervalos de confiança da Figura D.2, as seguintes Hipóteses foram construídas:

$H0_D : T_{3i} = T_1$ $\qquad$ $H1_D : T_{3i} > T_1$

$H0_E : T_1 = T_{3e}$ $\qquad$ $H1_E : T_1 > T_{3e}$

$H0_F : T_{3e} = T_{5-25\%}$ $\qquad$ $H1_F : T_{3e} > T_{5-25\%}$

D.2.1 Verificação de $H0_D$ e $H1_D$

As hipóteses $H0_D$ e $H1_D$, verificam o desempenho de precisão entre T_{3i} e T_1. Utilizando o teste de Mann-Whintey da ferramenta Minitab, para um nível de confiança de 95% (i.e. $\alpha = 0,05$) temos os seguintes resultados:

```
Mann-Whitney Test and CI: T3_I; T1

        N  Median
T3_I  100  29,730
T1    100  27,778

Point estimate for ETA1-ETA2 is 1,952
95,0 Percent CI for ETA1-ETA2 is (0,601;1,952)
W = 12082,0
Test of ETA1 = ETA2 vs ETA1 > ETA2 is significant at 0,00001
The test is significant at 0,00001 (adjusted for ties)
```

Observando os resultados dos testes, verificamos que o p ($p = 0,00001$) foi menor que o nível de significância ($\alpha = 0,05$), portanto, rejeitamos a hipótese nula, em favor da hipótese alternativa. Ou seja, podemos afirmar com 95% de confiança, que T_{3i} possui uma melhor precisão que T_1.

D.2.2 Verificação de $H0_E$ e $H1_E$

O desempenho de precisão entre T_1 e T_{3e} é verificado pelas hipóteses $H0_E$ e $H1_E$. Para realizar este teste de hipótese, submetemos os dados dessas técnicas ao teste de Mann-Whitney da ferramenta Minitab. Para um nível de confiança de 95%, obtivemos os seguintes resultados:

```
Mann-Whitney Test and CI: T1; T3_E

        N  Median
T1    100  27,778
T3_E  100  25,676


Point estimate for ETA1-ETA2 is 2,102
95,0 Percent CI for ETA1-ETA2 is (2,102;3,453)
W = 12660,0
Test of ETA1 = ETA2 vs ETA1 > ETA2 is significant at 0,00001
The test is significant at 0,00001 (adjusted for ties)
```

O p obtido como resultado do teste ($p = 0,00001$) é menor que o nível de significância estabelecido para o teste ($\alpha = 0,05$), indicando que a hipótese nula pode ser rejeitada em favor da hipótese alternativa. Diante disto, podemos afirmar com um nível de confiança de 95%, que T_1 possui melhor desempenho de precisão que T_{3e}.

D.2.3 Verificação de $H0_F$ e $H1_F$

Para investigar o desempenho de precisão entre T_{3e} e $T_{5-25\%}$, realizamos um teste de hipóteses com $H0_F$ e $H1_F$. Aplicando o teste de Mann-Whitney da ferramenta Minitab, sob um nível de confiança de 95%, obtivemos os seguintes resultados.

```
Mann-Whitney Test and CI: T3_E; T5-25

         N  Median
T3_E   100  25,676
T5-25  100  24,324
```

```
Point estimate for ETA1-ETA2 is -0,000
95,0 Percent CI for ETA1-ETA2 is (-0,000;1,351)
W = 10356,0
Test of ETA1 = ETA2 vs ETA1 > ETA2 is significant at 0,2277
The test is significant at 0,2229 (adjusted for ties)
```

A partir dos resultados do teste, verificamos que o p resultante ($p = 0,2229$) é maior que o nível de significância estabelecido ($\alpha = 0,05$). Portanto, não podemos rejeitar a hipótese nula em favor da hipótese alternativa. Ou seja, para um nível de confiança de 95%, não podemos rejeitar a hipótese de que a precisão de T_{3e} é melhor que a precisão de $T_{5-25\%}$.

D.2.4 Conclusões sobre o Desempenho de Precisão

A partir dos resultados obtidos com os testes visuais dos Intervalos de Confiança, assim como, os resultados dos testes de hipóteses (utilizando o teste de Mann-Whitney), obtivemos a seguinte ordenação (decrescente) de precisão entre as técnicas analisadas:

$$T_2 > T_{5-75\%} > T_{5-50\%} > T_4 > T_{3i} > T_1 > T_{3e} = T_{5-25\%}.$$

Onde, a técnica com a melhor precisão foi T_2, enquanto que as técnicas que apresentaram o menor desempenho em precisão, dentre as analisadas, foram T_{3e} e $T_{5-25\%}$.

D.3 Desempenho das Técnicas – Eficiência

De acordo com os dados sobre eficiência (em nanossegundos) observados, nenhuma técnica possuiu variância nula. Para auxiliar na definição de hipóteses sobre as técnicas, intervalos de confiança foram construídos, como ilustrado na figura D.3.

Na Figura D.3 (a), é possível observar que T_1 é a menos eficiente dentre as técnicas, pois demora mais tempo que todas as outras. É possível afirmar também que $T1 < T_{2i} < T_{2e}$, de acordo com a Figura D.3 (b).

Para decidir se T_{3e} ou se T_{3i} é mais eficiente, um teste Mann-Whitney unilateral foi realizado. Os resultados do teste são mostrados a seguir:

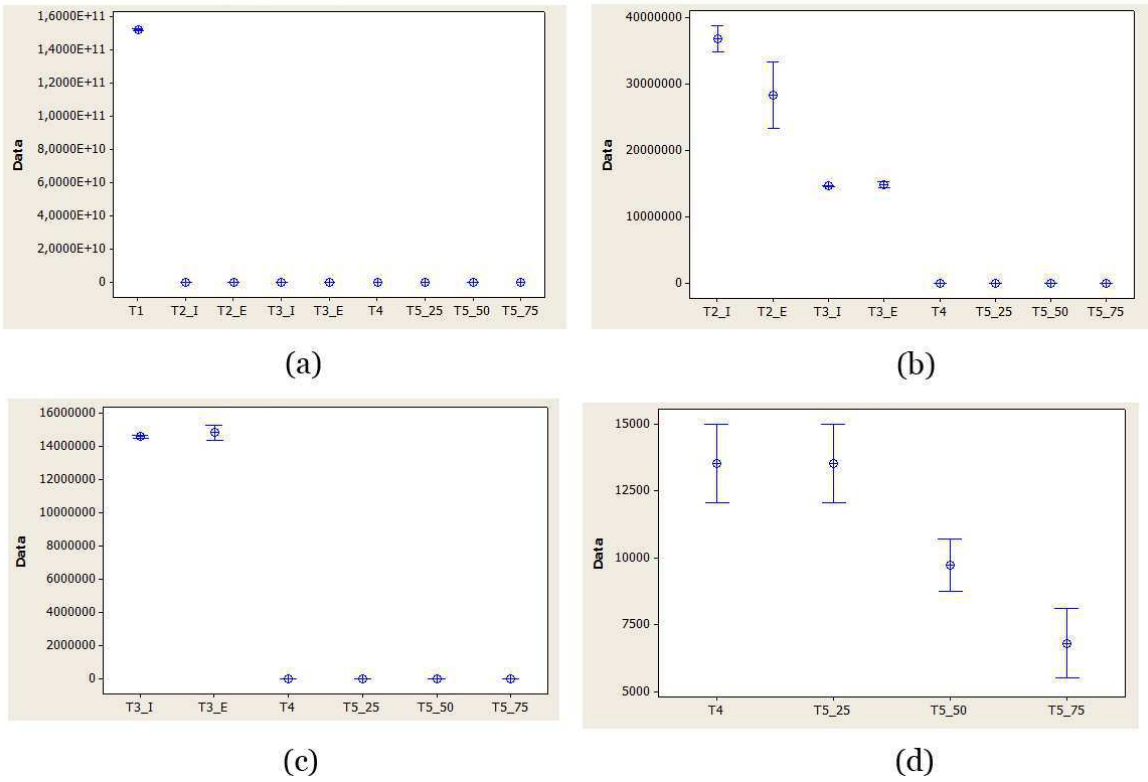

Figura D.3: Intervalos de confiança para a variável dependente eficiência.

```
        N    Median
T3_E  100  14430249
T3_I  100  14510029


Point estimate for ETA1-ETA2 is -103980
95,0 Percent CI for ETA1-ETA2 is (-183762;-25637)
W = 8985,5
Test of ETA1 = ETA2 vs ETA1 < ETA2 is significant at 0,0047
The test is significant at 0,0047 (adjusted for ties)
```

De acordo com o teste realizado, é possível concluir que T_{3e} é menos eficiente que T_{3i}, pois $p = 0.0047 < 0.05 = \alpha$. Um teste similar será aplicado entre as técnicas T_4 e $T_{5-25\%}$, resultado em:

```
          N  Median
T5_25   100   10541
T4      100   10541


Point estimate for ETA1-ETA2 is -0
95,0 Percent CI for ETA1-ETA2 is (-1;1)
W = 10050,0
Test of ETA1 = ETA2 vs ETA1 < ETA2 is significant at 0,5000
The test is significant at 0,5000 (adjusted for ties)
```

Uma vez que o p é maior que o α, não há evidências de que T_4 seja mais eficiente que $T_{5-25\%}$. Um teste bilaterial será realizado para verificar se há diferenças entre estas técnicas:

```
          N  Median
T4      100   10541
T5_25   100   10541


Point estimate for ETA1-ETA2 is -0
95,0 Percent CI for ETA1-ETA2 is (-1;1)
W = 10050,0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 1,0000
The test is significant at 1,0000 (adjusted for ties)
```

Como p também é maior que α, não há evidências para rejeição da hipótese nula. Portanto, considera-se que $T_4 = T_{5-25\%}$. Para finalizar as conclusões sobre as técnicas a respeito da eficiência, será verificado se há diferenças entre as técnicas $T_{5-50\%}$ e $T_{5-75\%}$:

```
          N  Median
T5_50   100  7667,0
T5_75   100  4792,0


Point estimate for ETA1-ETA2 is 2875,0
95,0 Percent CI for ETA1-ETA2 is (2873,9;2875,0)
W = 13606,0
```

```
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0,00001
The test is significant at 0,00001 (adjusted for ties)
```

Como p foi menor que α, então há diferenças entre as técnicas. Um teste unilateral será realizado para verificar se $T_{5-50\%} > T_{5-75\%}$:

```
         N  Median
T5_50  100  7667,0
T5_75  100  4792,0


Point estimate for ETA1-ETA2 is 2875,0
95,0 Percent CI for ETA1-ETA2 is (2873,9;2875,0)
W = 13606,0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0,00001
The test is significant at 0,00001 (adjusted for ties)
```

De acordo com o teste realizado, $p < \alpha$ e, portanto, rejeita-se a hipótese nula que as técnicas são iguais e aceita-se a hipótese de que $T_{5-50\%} > T_{5-75\%}$.

Por fim, para a variável dependente eficiência, pode-se afirmar que as técnicas se comportam da seguinte maneira: $T1 < T_{2i} < T_{2e} < T_{3e} < T_{3i} < T_4 = T_{5-25\%} < T_{5-50\%} < T_{5-75\%}$.

D.4 Desempenho das Técnicas – Potencial de Redução

Utilizando o teste de Kurskal-Wallis, verificamos que as técnicas possuem comportamentos distintos, com relação ao seu respectivo pontencial de redução. A partir destes resultados, decidimos realizar uma investigação para obter um desempenho comparativo entre o potencial de redução das técnicas. Considerando que a maioria das técnicas ($T_2, T_4, T_{5-25\%}, T_{5-50\%}$ e $T_{5-75\%}$) apresentaram variância nula, não seria adequada a realização de testes de hipóteses. Diante disto, observamos os intervalos de confiança do potencial de redução de cada técnica, para estabelecer o desempenho comparativo entre as técnicas.

Os intervalos de confiança das técnicas foram calculados, com um nível de confiança de 95%, a partir dos dados obtidos durante a execução do experimento. O gráfico da Figura D.4 (a) apresenta os intervalos de confiança de cada técnica (eixo-x) obtidos a partir do seu respectivo potencial de redução (em porcentagem) apresentados no eixo-y. Uma vez que a escala do potencial de redução das técnicas dificulta a vizualizção dos intervalos de confiança, o gráfico da Figura D.4 (a) foi separado nos gráficos das Figuras D.4 (b), (c) e (d).

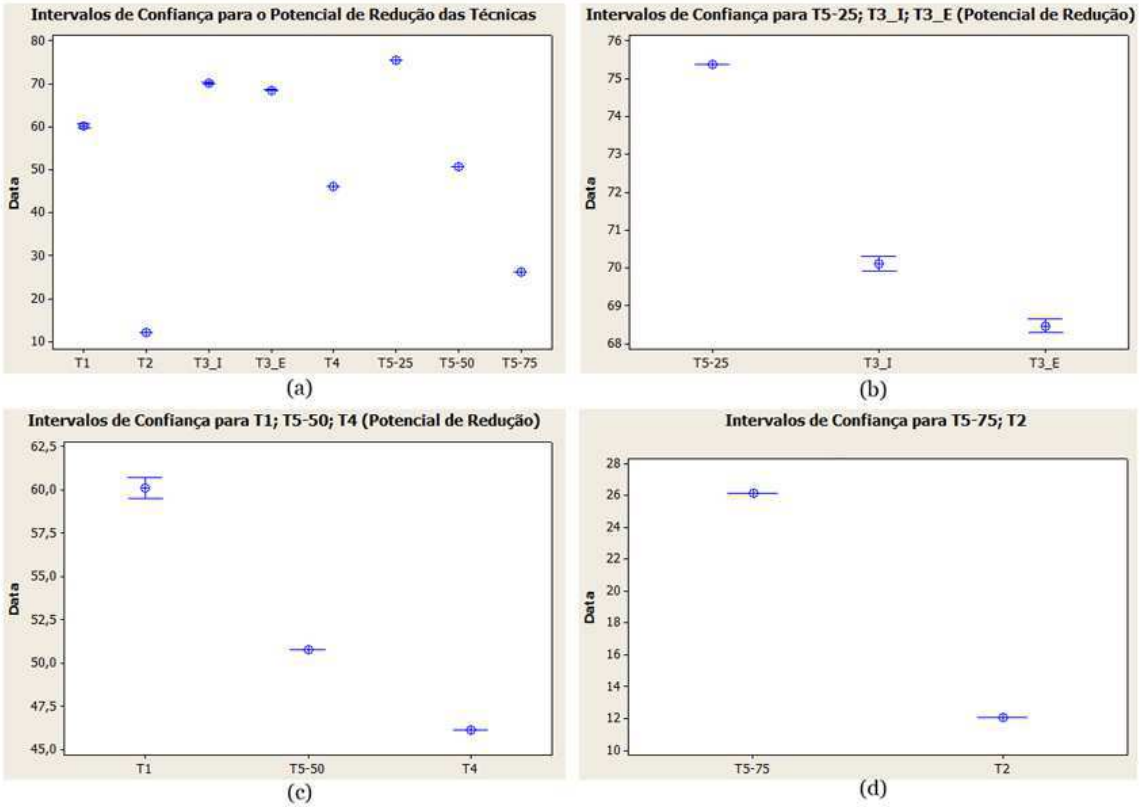

Figura D.4: Intervalos de confiança para o potencial de redução. (a) Todas as técnicas. (b) $T_{5-25\%}, T_{3i}$ e T_{3e}. (c) T_1, $T_{5-50\%}$ e T_4. (d) $T_{5-75\%}$ e T_2.

Uma vez que os intervalos de confiança não se sobrepõem, a ordem observada nos intervalos de confiança caracteriza o próprio desempenho das técnicas. Ou seja, as técnicas apresentam a seguinte ordenação (descrescente) de desempenho no potencial de redução:

$$T_{5-25\%} > T_{3i} > T_{3e} > T_1 > T_{5-50\%} > T_4 > T_{5-75\%} > T_2$$

Portanto, as técnicas com melhores desempenhos de potencial de redução são $T_{5-25\%}$ r T_{3i}, enquanto que T_2 apresentou o menor desempenho.

D.5 Desempenho das Técnicas – Densidade de Faltas

Com o objetivo de caracterizar o desempenho comparativo das técnicas de re-teste seletivo referente à densidade de faltas, intervalos de confiança foram construídos a partir dos dados obtidos na execução do experimento. Dentre os dados observados a densidade de faltas de T_2 e T_4 apresentaram variância nula, com média $38,23$ e $62,85$, respectivamente. Como estas técnicas não apresentaram variância, a média encontrada pode ser comparada com os intervalos de confiança das demais técnicas para observar o desempenho de densidade de faltas. Os intervalos de confiança dos dados foram construídos e são apresentados na Figura D.5.

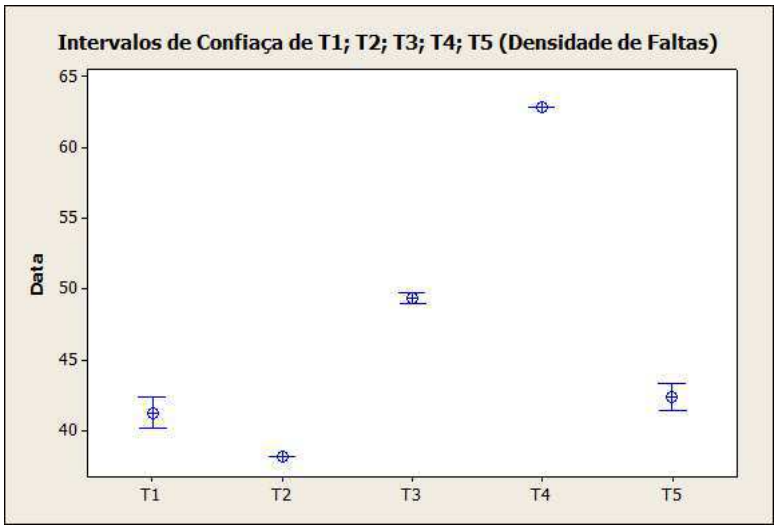

Figura D.5: Intervalos de confiança para a densidade de faltas das técnicas.

Observando os intervalos de confiança, verificamos que a maior média encontrada é T_4, seguida pelo intervalo de confiança de $T3e$. Como a média de T_4 $(62,85)$ é maior que intervalo de confiança de T_{3e} $([49,028; 49,736])$, podemos dizer que $T_4 > T_{3e}$. Os intervalos de confiança que seguem T_3 são os de T_1 e T_5. Considerando que estes intervalos se sobrepõem, é necessário realizar uma investigação mais precisa acerca do desempenho de densidade de faltas destas duas técnicas. Diante disto, é recomendável partir para um teste de hipóteses. Então, estabelecemos as hipóteses $H0_G$ (Hipótese Nula) e $H1_G$ (Hipótese Alternativa) de forma que:

$$H0_G\text{: } D(T_1) = D(T_5)$$
$$H1_G\text{: } D(T_1) \neq D(T_5)$$

Considerando que os dados de densidade de faltas de T_1 e T_5 apresentaram distribuição normal (Apêndice A), é possível aplicarmos o teste paramétrico t de Student para realizar o teste de hipóteses. Este teste foi realizado, sob um nível de confiança de 95% utilizando a ferramenta Minitab, e os resultados são apresentados a seguir:

```
Two-Sample T-Test and CI: T1; T5

Two-sample T for T1 vs T5

      N   Mean  StDev  SE Mean
T1  100  41,28   5,55     0,55
T5  100  42,38   4,63     0,46

Difference = mu (T1) - mu (T5)
Estimate for difference:  -1,097
95\% CI for difference:  (-2,522; 0,329)
T-Test of difference = 0 (vs not =): T-Value = -1,52  P-Value = 0,131  DF = 191
```

Observando o resultado do teste, verificamos que o p encontrado ($p = 0,131$) é maior que o nível de significância utilizado no teste ($\alpha = 0,05$). Portanto, com 95% de nível de confiança, não podemos rejeitar a hipótese nula, o que indica que T_1 e T_5 apresentam uma densidade de faltas semelhante. É importante lembrar que o teste-t é um teste paramétrico, e a quantidade de execuções das técnicas realizadas, contribuem significativamente para a confiança neste resultado, já que o teste-t possui um alto poder estatístico [Jain 1991, Siegel and Junior 1988].

A última técnica a ser analisada é T_2, e como podemos observar, sua média $(38,23)$ é menor que os intervalos de confiança de todas as outras técnicas (T_1, T_2, T_3 e T_4). Portanto, podemos dizer que T_2 apresentou o menor desempenho de densidade de faltas dentre as técnicas analisadas.

A partir dos resultados observados nesta análise, podemos resumir o desempenho de densidade de faltas das técnicas analisadas neste experimento da seguinte forma:

$$T_4 > T_3 > T_1 = T_5 > T_2$$

Ou seja, as técnicas com os melhores desempenhos foram T_4 e T_3, enquanto que a técnica com o menor desempenho foi T_2. A partir dos resultados do teste-t realizado (sob um nível

de confiança de 95%) com T_1 e T_5 não foi possível, rejeitar a hipótese de que T_1 e T_5 são diferentes entre si, com relação à densidade de faltas.