7
Avaliação

Nosso objetivo neste capítulo é avaliar as contribuições de estratégias sensíveis à história na detecção de anomalias de modularidade em dois sistemas selecionados para integrar este estudo. Para isso, consideramos três perguntas de pesquisa relacionadas à eficácia[1] dos resultados de detecção nesses sistemas.

A primeira pergunta (Q_1) motiva uma análise isolada dos resultados de estratégias sensíveis à história. Já as duas últimas (Q_2 e Q_3) motivam a realização de análises relativas dessas estratégias, comparando seus resultados com os de outras abordagens de detecção. Essas abordagens compreendem estratégias convencionais (Lanza e Marinescu 2006) e detecção baseada em recursos de visualização de código (Carneiro et al. 2010a).

As duas abordagens mencionadas para comparação foram escolhidas pois são frequentemente citadas como importantes recursos de detecção de anomalias. As questões Q_1, Q_2 e Q_3 mencionadas e que direcionam a avaliação de estratégias sensíveis à história nos sistemas selecionados estão definidas a seguir.

Q_1: Estratégias sensíveis à história apresentam-se como um recurso eficaz para detecção de anomalias de modularidade?

Q_2: Estratégias sensíveis à história tendem a se apresentar mais eficazes que estratégias convencionais?

Q_3: Estratégias sensíveis à história provêem informações capazes de contribuir com a eficácia de recursos visuais de detecção?

Nesta avaliação considerou-se amostras de versões de dois sistemas que resultaram em um total de 16 versões compondo o estudo. Nessas 16 versões os módulos foram individualmente e manualmente analisados. Análises sobre resultados de estratégias convencionais, bem como um estudo exploratório sobre detecções baseadas em recursos visuais também foram realizadas.

Destacamos que nesta avaliação não tivemos a pretensão de generalizar os resultados de eficácia obtidos, pois, para isso, um estudo mais controlado e com

[1]No decorrer do capítulo, são apresentadas as medidas estatísticas utilizadas para avaliação de eficácia das estratégias.

um maior número de amostras seria necessário. A finalidade principal desta avaliação foi a de realizar uma investigação sobre as potenciais contribuições de estratégias sensíveis à história apenas no subconjunto de amostras selecionado.

A apresentação do capítulo está organizada em duas partes. A Seção 7.1 apresenta uma descrição do estudo. Nessa seção estão descritos os procedimentos realizados e as características desta etapa do trabalho. A Seção 7.2 expõe os resultados e os discute considerando as questões Q_1, Q_2 e Q_3 anteriormente apresentadas. Todos os artefatos e dados apresentados nesta avaliação também estão disponíveis no sítio[2] oficial desta pesquisa.

7.1 Procedimentos

Esta seção descreve as características e atividades que integram este processo de avaliação. Na Seção 7.1.1 é apresentada a metodologia do estudo através da qual é possível ter uma visão geral das atividades do processo de avaliação. Os critérios para seleção dos sistemas bem como algumas de suas características são apresentadas na Seção 7.1.4. A Seção 7.1.3 apresenta o processo de obtenção das anomalias reais nos sistemas. Justificativas sobre a seleção das anomalias participantes do estudo são destacadas na Seção 7.1.2. Na Seção 7.1.5, definimos as medidas estatísticas para avaliação de eficácia das estratégias de detecção.

7.1.1 Metodologia do Estudo

Como passo inicial desta avaliação, foi realizada a detecção manual de anomalias de código ao longo das versões dos sistemas selecionados para esse estudo (Seção 7.1.4). Tais instâncias de anomalias seriam utilizadas para integrar o que apresentamos na Seção 7.1.3 como oráculos ou listas de referência. Em seguida, estratégias sensíveis à história definidas neste trabalho e estratégias convencionais definidas por Lanza e Marinescu (2006) foram aplicadas nos sistemas. Todas essas estratégias foram apresentadas no Capítulo 5 desta dissertação.

De forma geral, a aplicação das estratégias de detecção convencionais e sensíveis à história tiveram como objetivo fornecer subsídios para as discussões de Q_1 e Q_2. Já para responder Q_3 foi realizado um estudo exploratório (Carneiro et al. 2010a) sobre detecção de anomalias de código utilizando a ferramenta de visualização SourceMiner (Carneiro et al. 2009, Carneiro et al. 2010b). Esse estudo exploratório compreendeu apenas as amos-

[2]http://www.inf.puc-rio.br/ lsilva

tras de versões referente a um dos sistemas. Isso porque tornarímaos o estudo exploratório muito cansativo para os participantes caso solicitássemos que eles fizessem a inspeção ao longo das amostras de versões de ambos os sistemas. É sabido que a disponibilidade dos participantes acaba sendo também uma fator crítico na execução de estudos como o realizado.

Para as diferentes abordagens de detecção consideradas, as medidas de precisão e revocação definidas na Seção 7.1.5 foram calculadas e, então, disponibilizadas para análise. O cálculo dessas medidas tinha como objetivo possibilitar análises quantitativas e comparativas sobre a eficácia de cada uma das abordagens de detecção utilizadas.

Para aplicação das estratégias SHs foi utilizada a ferramenta proposta no Capítulo 6 desta dissertação, a Hist-Inspect. Já para a aplicação das estratégias convencionais de Marinescu (2006) foram utilizados dois mecanismos, de forma que um pudesse validar o resultado do outro. Esses mecanismos foram: (1) utilização da ferramenta proposta Hist-Inspect e (2) aplicação manual das estratégias a partir da análise dos resultados das métricas gerados pela ferramenta Together (Together 2009). Cada mecanismo confirmou devidamente o resultado obtido pelo outro.

Os detalhes sobre a utilização da SourceMiner para detecção de anomalias de código são apresentados em (Carneiro et al. 2010a). Em resumo, cinco participantes foram treinados para utilizar a ferramenta de visualização SorceMiner (Carneiro et al. 2009, Carneiro et al. 2010b). Além disso explicações e referências sobre as anomalias investigadas também foram fornecidas a eles. Após os devidos treinamentos e aquisição dos conhecimentos que lhes foram fornecidos, os participantes foram solicitados a inspecionar as versões de código do sistema Mobile Media (Seção 7.1.4) utilizando os recursos de visualização de código disponibilizados pela SourceMiner. Apenas com tais recursos eles identificariam os módulos com alguma anomalia de modularidade específica. Tal identificação foi realizada sem que os participantes pudessem investigar manualmente o código da aplicação selecionada.

Os dados do estudo exploratório foram transportados para este capítulo conforme o seguinte critério. Foram marcados como anômalos todos os módulos detectados por pelo menos um dos cinco partipantes. Ou seja cada detecção era devidamente considerada no estudo exploratório sem ignorar as indicações de nenhum dos participantes treinados. Dessa forma, um verdadeiro positivo na apresentação dos dados deste experimento significa que pelo menos um dos cinco participantes conseguiu identificar uma anomalia real no sistema através dos recursos visuais. De forma equivalente, um falso positivo indica que a utilização do recurso visual induziu a pelo menos um dos participantes

a detectar uma anomalia que não era real no sistema.

Após a obtenção dos resultados de cada uma das abordagens, uma análise detalhada dos dados foi realizada. Tal análise forneceria os parâmetros para se verificar: (i) se as estratégias sensíveis à história poderiam ser eficazes na detecção de anomalias de código, (ii) se estratégias sensíveis à história poderiam resultar em menos falsos positivos e negativos que estratégias convencionais e ainda (iii) se estratégias sensíveis à história poderiam ser utilizadas para minimizar possíveis deficiências de detecções baseadas em recursos visuais.

7.1.2 Anomalias Selecionadas

As anomalias selecionadas para integrar esta avaliação foram *God Class* (GC), *Shotgun Surgery* (SS) e *Divergent Change* (DC), todas definidas no Capítulo 5 desta dissertação. *God Class* foi escolhida pois, segundo pesquisas (Ducasse et al. 2004), classes com tal problema costumam ser bastante instáveis e propensas a erros. Desta forma, alguns pesquisadores suspeitam que análises sensíveis à história possam trazer importantes contribuições na detecção dessas anomalias. Além disso, alguns estudiosos (Lanza e Marinescu 2006) da área já haviam definido estratégias convencionais para a detecção de classes com tal problema. Isso contribui para a realização de comparações dos resultados de estratégias sensíveis à história com os de estratégias de destaque na literatura.

Shotgun Surgery e *Divergent Change* foram escolhidas pois diversas pesquisas sobre mecanismos de avaliação de modularidade contemplam tais anomalias. Esses estudos consideram tanto recursos visuais (Carneiro et al. 2009) quanto estratégias formadas por métricas não convencionais, por exemplo sensíveis a interesses (Figueiredo et al. 2009). Tal fato viabiliza a realização de pesquisas que associem a abordagem sensível à história com essas outras abordagens de detecção. Além disso, no caso da anomalia *Divergent Change*, não existe na literatura nenhuma estratégia convencional (Lanza e Marinescu 2006) para apoiar tal detecção. Tal fato nos possibilita verificar se estratégias sensíveis à história podem contribuir na detecção dessas classes.

7.1.3 Elaboração dos Oráculos ou Listas de Referência

Para obtenção das anomalias reais dos sistemas selecionados (Seção 7.1.4) foi solicitada a contribuição de dois especialistas desses sistemas. Com a apresentação das definições das anomalias, solicitamos que eles: (i) apresentassem

uma lista dos módulos que eles julgavam infectados por cada anomalia considerada, e (ii) destacassem a confiabilidade de cada detecção. Esse critério seria utilizado caso fosse necessária alguma reunião de consenso entre as detecções.

A lista elaborada pelos especialistas dos sistemas teria como objetivo servir de referência para avaliar os erros e acertos dos mecanismos de detecção considerados. Entretanto, é sabido que, devido a subjetividade comumente encontrada nas definições das anomalias, a conclusão sobre a existência ou não do problema pode, frequentemente, gerar controvérsias. Dessa forma, como cada sistema foi avaliado por mais de um especialista, de forma não surpreendente, os módulos detectados por um não foram exatamente os considerados pelo outro.

Para contornar tal situação, uma reunião foi realizada entre os mesmos para que cada um explicasse seu ponto de vista e então se chegasse a consenso sobre os módulos com anomalias. O resultado das discussões geradas foi considerado como a opinião de um terceiro especialista, destacando uma opinião que era comum aos dois especialistas. Considerarou-se como anomalias reais de cada um dos sistemas apenas as listagens resultantes da reunião de consenso entre esses.

7.1.4 Sistemas Selecionados e Critérios de Seleção

A elaboração de oráculos ou de listas de referência, como as definidas na Seção 7.1.3, é altamente dependende da disponibilização de desenvolvedores ou de projetistas que conheçam o código das aplicações alvo. Os oráculos realizados para esta avaliação possibilitam que os resultados sejam comparados a anomalias reais e não a supostas anomalias detectadas por abordagens já existentes.

É possível notar que a maioria das pesquisas (Olbrich et al., 2009, Shatnawi e Li 2006) que avaliam a detecção de um elevado número de sistemas e de versões se rende a tomar as detecções de estratégias convencionais da literatura como fontes verdadeiras de referência. Isso é feito pois análises manuais módulo a módulo, versão a versão de uma grande número de sistemas são frequentemente muito dispendiosas.

Entretanto, estudos que visam contribuir com detecções de anomalias e que tomam como verdadeiras as detecções realizadas via estratégias da literatura possibilitam muito fortemente a propagação de erros. Isso por causa das frequentes ocorrências de falsos positivos e negativos. Para evitar tal risco, optamos por utilizar um número reduzido de sistemas e de versões mas ter a vantagem de ter como referência anomalias detectadas manualmente por

especialistas. Sabemos que análises manuais também são passíveis de falhas. Entretanto, ainda assim, são consideradas um parâmetro mais confiável que detecções automáticas não validadas, principalmente quando os inspetores possuem suficiente conhecimento das aplicações avaliadas.

Com base nos critérios discutidos acima, selecionamos dois sistemas de pequeno e médio porte: Mobile Media (MM) (Figueiredo et al., 2008) e Health Watcher (HW) (Greenwood et al., 2007). O primeiro, trata-se da implementação de uma linha de produtos (Saleh e Gomma 2005) para manipulação de fotos, músicas e vídeos em aparelhos celulares. O segundo é um sistema de informação de tecnologia Web cuja principal funcionalidade é o registro de queixas relacionadas à qualidade de serviços de estabelecimentos públicos. A seleção desses sistemas foi uma decisão estratégica baseada principalmente nas semelhanças e diferenças existentes entre eles e que foram consideradas igualmente importantes para esta avaliação. Tais semelhanças e diferenças são descritas a seguir.

Semelhanças importantes:

1. Para ambos os sistemas seria possível acessar projetistas que poderiam identificar manualmente e/ou confirmar as anomalias reais das aplicações. Isso tornaria possível a avaliação confiável dos resultados em relação a opinião desses especialistas.

2. Ambos os sistemas foram desenvolvidos com a preocupação de atender requisitos de modularidade de código. Tal característica torna a detecção de anomalias nesses sistemas longe de ser trivial. Esta particularidade foi considerada interessante pois poderia contribuir para explorarmos o apóio a detecções não óbvias e que frequentemente passariam despercebidas por equipes de desenvolvimento ou por estratégias convencionais.

Diferenças importantes:

1. Tais aplicações pertencem a diferentes domínios, além de utilizarem diferentes tecnologias de implementação. Tal fato contribui para que a possível obtenção de resultados comuns entre as aplicações não seja induzida pela utilização de sistemas de elevada similaridade, de mesmo domínio ou mesma tecnologia.

2. Esses sistemas apresentam cenários de evolução bastante diferenciados. No MM, cada versão se diferencia da anterior pela adição de novas funcionalidades. Enquanto no HW, as diferenças relativas às evoluções são geralmente resultantes de sucessivas reestruturações.

Tabela 7.1: Principais características das aplicações do estudo

	Mobile Media	**Health Watcher**
Tipo de Aplicação	Linha de Produto	Sistema Web
Ling. de Programação	Java	Java
Número de Versões Selecionadas	6	10
LOC[1] em v_1 : LOC em v_n	760 : 2670	5295 : 7593
Crescimento Médio de LOC	238,75	229,8
NOC[2] em v_1 : NOC em v_n	16 : 51	88 : 135
Crescimento Médio de NOC	5,6	4,7
Média de GC[3]	11/6 = 1,83	32/10 = 3,2

[1]LOC: Número de linhas de código, desconsiderando brancos e comentários; [2]NOC: Número de Classes; [3]GC: *God Classes*

A Tabela 7.1 apresenta, de forma resumida, algumas características dessas aplicações. Nesse estudo, foram consideradas 6 versões do Mobile Media (v2-v7) e 10 versões do Health Watcher (v1-v10). Além disso, ambos são escritos exclusivamente em linguagem Java, como mostra a tabela mencionada. Também podemos observar que no MM exige-se um menor esforço na detecção manual dos módulos em virtude do seu menor número de linhas. Como podemos notar, enquanto na primeira versão avaliada do MM existem 760 linhas, no HW existem 5295 linhas, o que representa uma diferença de quase 7 vezes mais linhas de código no segundo sistema.

N o que diz respeito ao número de classes, também podemos notar um número bem maior de classes no segundo sistema, tanto na primeira versão avaliada quanto na última. A última linha da tabela confirma a informação sobre a preocupação com modularidade de ambos os sistemas. É possível perceber que ambos apresentam uma média de anomalias consideravelmente baixa. No caso da anomalia de tipo *God Class* tem-se uma média aproximada de 2 anomalias por versão no MM e 3 no Health Watcher.

7.1.5 Avaliação de Eficácia das Estratégias

A avaliação da eficácia no estudo foi suportada pela análise das medidas de revocação e precisão (Rijsbergen 2001). Além disso, outras variáveis a serem disponibilizadas para análises dos dados são: número de acertos ou verdadeiros positivos, número de falsos positivos e número de falsos negativos.

Um acerto ou verdadeiro positivo (VP) ocorre quando a estratégia avaliada identifica uma entidade que também está presente na lista de referência. Um falso positivo (FP) ocorre quando a entidade detectada não se encontra na lista de detecção previamente elaborada. Um falso negativo (FN) ocorre

quando uma entidade está na lista de referência mas não conseguiu ser detectada pela estratégia avaliada.

É a partir das variáveis mencionadas que as medidas de revocação e precisão são calculadas. A revocação representa o percentual de acertos em relação as entidades presentes na lista de referência. Enquanto que a precisão representa o percentual de acertos em relação as entidades detectadas por uma dada abordagem. Os cálculos são realizados de acordo com as seguintes fórmulas:

$$\text{Revocação} = \frac{\text{VP}}{\text{VP} + \text{FN}} \qquad \text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Considerando Refatorações *Rename*

Cada estratégia sensível à história foi aplicada nos sistemas selecionados em dois casos. No primeiro, chamado caso1, as estratégias foram aplicadas em cada versão, sem considerar o mapeamento de histórico das classes que sofreram refatorações de tipo "*rename*". Ou seja, uma classe renomeada passava a ser considerada uma nova classe, perdendo parte da sua história. No segundo, chamado caso2, as mesmas estratégias foram reaplicadas assumindo que as entidades renomeadas ainda permeneciam com seus nomes originais.

Refatorações desse tipo foram observadas apenas na versão 7 do MM e na versão 2 no HW. Nosso interesse em testar o segundo caso era o de avaliar os efeitos da abordagem sensível à história caso ela fosse estendida de forma a conseguir detectar a ocorrência de refatorações desse tipo. Para isso, as classes renomeadas foram acessadas e refatoradas para que recebessem seus nomes originais e, assim, fosse o possível mapear o histórico de cada uma nas detecções sensíveis à história.

7.1.6 Outras Considerações

Apesar de terem sido selecionadas três anomalias para esta avaliação, algumas restrições podem ser destacadas. A anomalia *Shotgun Surgery* (SS) não foi considerada no estudo com o sistema Health Watcher. Tal decisão foi tomada pois não foi disponibilizada pelos especialistas desse sistema uma lista de referência (Seção 7.1.3) para tal anomalia. A inexistência dessa lista impediria a avaliação de eficácia dos resultados de detecção referentes a essa anomalia.

Além disso, o estudo exploratório que realizamos sobre o uso dos recursos visuais não contemplou resultados referentes à anomalia SS. Na época do

estudo exploratório sobre a utilização dos recursos visuais, houve o interesse em pesquisar sobre a detecção da anomalia *God Class*, *Shotgun Surgery* e de uma outra anomalia relacionada a métodos. Já na avaliação das estratégias sensíveis à história consideramos, à priori, apenas anomalias de granularidade de classe.

7.2 Resultados e Discussões

A apresentação dos dados para avaliação das estratégias sensíveis à história nos sistemas selecionados está organizada em três etapas. Tal organização visa facilitar a discussão das três perguntas de pesquisa apresentadas no início deste capítulo. Dessa forma, na primeira etapa de discussões realizamos uma análise isolada dos resultados de estratégias sensíveis à história. Em seguida, comparamos tais resultados com os de estratégias convencionais. Por fim, realizamos uma análise relativa em relação aos resultados obtidos através do uso de recursos de visualização de código. Tais etapas são apresentadas individualmente para cada um dos sistema que compreendem tal avaliação.

Como o estudo exploratório sobre o uso de recursos de visualiação de código compreendeu apenas o sistema Mobile Media, destacamos que Q3 que considera tal abordagem de detecção será discutida apenas no contexto deste sistema. Além disso, resultados sobre a anomalia *Shotgun Surgery* não foram considerados para o caso do sistema Health Watcher. Isso porque Essa anomalia foi desconsiderada nesse sistema devido à inexistência de um oráculo para essa anomalia. Em cada um das etapas, reportamos as medidas definidas na Seção 7.1.5 em cada uma das versões dos sistemas. Além disso, também consideramos nas discussões de comparação entre as abordagens a apresentação das médias de precisão e revocação de cada mecanismo de detecção considerado.

Destacamos ainda que em todas as tabelas à seguir as medidas de precisão e revocação foram apresentadas em sua forma percentual. Quando houve a impossibilidade de efetuar o cálculo de uma dessas medidas devido a divisões por zero, um traço foi utilizado para simbolizar a situação. A Seção 7.2.1 apresenta os resultados de avaliação com o sistema Mobile Media nas etapas mencionadas. A Seção 7.2.2 apresenta os resultados da avaliação com o sistema Health Watcher. A Seção x, apresenta uma breve discussão considerando a média dos resultados entre os dois sistemas.

7.2.1 Primeiro Sistema: Mobile Media

Q1: Eficácia de Estratégias Sensíveis à História

A Tabela 7.2 apresenta os dados de avaliação referentes à detecção de todas as anomalias consideradas, em todas as versões do Mobile Media. Esses dados nos fornecem insumos para analisar a eficácia de estratégias sensíveis à história nesse sistema.

Podemos observar que tanto na detecção de GCs quanto na de DCs foi possível detectar 100% das anomalias em pelo menos duas das versões analisadas. Isso ocorre em v5 e v6 no caso da detecção de GCs e de v2 à v4 no caso de DCs. Tal valor máximo de revocação também é possível ser observado em v7-caso2 dentre as detecções de SS. Consideramos ainda que, de forma bastante surpreendente, mesmo nessas versões em que o número de acertos de GCs e DCs foi máximo (revocação = 100%), não houve a detecção de nenhum falso positivo. Tal fato também repercutiu em precisão de 100% nessas versões.

God Class (GC)						
	v2	**v3**	**v4**	**v5**	**v6**	**v7**[2]
Acertos	1	1	1	1	2	0 (1)
Falsos Positivos	0	0	0	0	0	0 (0)
Falsos Negativos	1	1	1	0	0	2 (1)
Revocação	50%	50%	50%	100%	100%	0% (50%)
Precisão	100%	100%	100%	100%	100%	- (100%)
Divergent Change (DC)						
	v2	**v3**	**v4**	**v5**	**v6**	**v7**[2]
Acertos	2	2	2	2	2	2 (2)
Falsos Positivos	0	0	0	0	1	2 (1)
Falsos Negativos	0	0	0	1	1	2 (2)
Revocação	100%	100%	100%	67%	67%	50% (50%)
Precisão	100%	100%	100%	100%	67%	50% (67%)
Shotgun Surgery (SS)						
	v2[1]	**v3**[1]	**v4**[1]	**v5**	**v6**	**v7**[2]
Acertos	n/a	n/a	n/a	2	2	3 (4)
Falsos Positivos	0	0	0	1	1	3 (4)
Falsos Negativos	0	0	0	1	1	1 (0)
Revocação	n/a	n/a	n/a	67%	67%	75%(100%)
Precisão	n/a	n/a	n/a	67%	67%	50% (50%)

[1]n/a (não aplicável): não existem anomalias do tipo SS a serem detectadas nessas versões. [2]Em v7, os resultados entre parênteses são referentes à aplicação da estratégia SH no caso2 (Seção x).

Tabela 7.2: Detecção sensível à história no Mobile Media

O que é bastante comum quando se obtêm valores de revocação máximos é que como efeito colateral costuma-se obter muitos falsos positivos e isso não

foi o caso. Se considerarmos a análise de todas as versões de v2 à v6 e a versão v7-caso2 e considerarmos os três tipos de anomalias, a menor precisão e revocação foi de 50%. Esses resultados são considerados bastante satisfatórios.

Além disso, dentre as detecções dos três tipos de anomalias, foi possível a obtenção de pelo menos 1 acerto (VP) em todas as versões, exceto em v7-caso1 das detecções de GCs. Em v7, por ser a última versão e por apresentar o maior histórico dentre as versões, era esperado que fosse obtido um dos melhores percentuais de acerto. Entretanto, surpreendentemente, nessa versão não foi possível a localização de nenhuma GC. Existe uma justificativa razoável para tal falha em v7. Tal justificativa toma como base a análise comparativa dos resultados de detecção de GC em v7 nos casos 1 e 2, esse último entre parênteses. Como explicado na Seção 7.1.5, o caso 2 considera a aplicação da estratégia como se a avaliação fosse capaz de identificar refatorações do tipo "*rename*".

A justificativa é que as classes detectadas corretamente em versões anteriores sofreram renomeação em v7. Ao considerarmos o caso2, pudemos observar que uma dentre as duas GCs existentes em v7 passou a ser detectada. Tal fato possibilitou a obtenção de 100% de precisão na detecção de GCs em todas as versões. A estratégia SH para GC só não conseguiu detectar a outra anomalia existente pois essa apresentou uma redução de complexidade de aproximadamente 100 linhas de código. Isso fez com que a estratégia SH tivesse considerado a eliminação da anomalia detectada na versão anterior. Esse fato foi o único que impossibilitou a obtenção de uma revocação de 100% no caso2, valor que já tinha sido obtido para a precisão.

Elevados valores de precisão e revocação também puderam ser observadas nas detecções de DC e SS. Outras observações importantes de serem feitas é que no caso de DC ainda nem existe dentre as estratégias de Lanza e Marinescu (2006) uma estratégia convencional para detectar tal anomalia. Contudo, a estratégia sensível à história avaliada para DC apresentou-se bastante eficiente. Como relatado anteriormente, foram obtidos valores de 100% de precisão e revocação na metade das versões avaliadas (v2-v4).

No que diz respeito à detecção de SS, segundo inspeção manual nem haveria a necessidade de avaliar a estratégia correspondente nas versões iniciais do MM. Isso porque de v2 à v4 não existiam anomalias a serem detectadas. Dessa forma, não seria possível contabilizar o número de entidades detectadas corretamente nem de efetuar os cálculos de precisão e revocação. Apesar disso, a estratégia foi aplicada e, de fato, ela não considerou a detecção de nenhuma anomalia desse tipo entre v2 e v4, passando a identificá-la apenas a partir da versão em que se era esperado iniciar a aplicação das estratégias. Nesses casos,

de v3 à v7, também foi observado que, considerando a possibilidade de mapear o histórico de entidades renomeadas (caso2, entre parênteses), foi possível a detecção de 100% das anomalias desse tipo. No caso em que o mapeamento de renomeações não foi considerado também foi possível obter uma revocação satisfatória de 75%.

Se nos basearmos apenas nos resultados obtidos na avaliação com o Mobile Media podemos dizer que são apresentadas evidências iniciais que suportam a seguinte resposta para Q1: *"Estratégias sensíveis à história podem contribuir com detecções eficazes de anomalias de modularidade"*.

RQ2: Estratégias Sensíveis à História vs. Estratégias Convencionais

God Class. A Tabela 7.3 reúne os resultados das estratégias convencional e sensível à história para GC de a forma a permitir uma comparação entre esses resultados.

Detecção Convencional de GC						
	v2	**v3**	**v4**	**v5**	**v6**	**v7**
Acertos	0	1	1	0	0	0
Falsos Positivos	0	0	0	0	0	0
Falsos Negativos	2	1	1	1	2	2
Revocação	0%	50%	50%	0%	0%	0%
Precisão	-	100%	100%	-	-	-

Detecção Sensível à História de GC						
	v2	**v3**	**v4**	**v5**	**v6**	**v7**[1]
Acertos	1	1	1	1	2	0 (1)
Falsos Positivos	0	0	0	0	0	0 (0)
Falsos Negativos	1	1	1	0	0	2 (1)
Revocação	50%	50%	50%	100%	100%	0% (50%)
Precisão	100%	100%	100%	100%	100%	- (100%)

[1]Em v7, os resultados entre parênteses são referentes à aplicação da estratégia SH no caso2 (Seção x).

Tabela 7.3: Detecção convencional vs. sensível à história de GC no Mobile Media

Como podemos observar a utilização de estratégia sensível à história possibilitou melhorias dos resultados de v2, v5, v6 e v7, quando analisada a v7-caso2. Além disso, dentre as versões que não houve melhoria dos valores, também não houve impacto negativo nos resultados. Ou seja, não houve diminuição dos valores de precisão e revocação nessas versões, v3 e v4, em que não houve melhoria dos resultados. Os resultados permaneceram exatamente os mesmos dos obtidos pelas estratégias covencionais. As versões 2, 5, 6 e 7 que não apresentavam acertos provenientes da aplicação da estratégia convencional,

passaram a ter pelo menos um acerto em cada versão, propiciando precisão de 100% nessas versões. A seguir apresentamos os resultados médios de precisao e revocação de ambas as abordagens.

- Detecção Convencional de GC:
  média de revocação = 16.6%; média de precisão = 33%
- Detecção Sensível à História de GC:
  média de revocação = 66%; média de precisão = 100%

Se considerarmos a média de precisão e de revocação para avaliarmos comparativamente ambas as abordagens, podemos observar o quanto as estratégias SHs se apresentaram melhores que as convencionais. Tanto a revocação média quanto a precisão média da abordagem sensível à história foram superiores às medidas médias da abordagem convencional. Na detecção de GCs, a utilização de estratégias sensíveis à história não somente se apresentou eficaz como também possibilitou acertos não obtidos pela abordagem convencional.

Com base nos dados apresentados, podemos realizar a seguinte afirmação para RQ2 em relação à detecção de *God Classes*: *"Estratégias sensíveis à história podem ser mais eficazes que estratégias convencionais"*. De fato, as suspeitas em relação a RQ2 era que a abordagem SH por ter a sua disposição um maior número de informações possibilitaria resultados mais eficazes que os de estratégias convencionais. A observação desse resultado no MM não garante resultados similares em outros sistemas ou outras anomalias. Entretanto, sabemos que é possível que estratégias SHs contribuam com melhores resultados de detecção.

Shotgun Surgery. A Tabela 7.4 reúne os resultados das estratégias convencional e sensível à história para SS. Como podemos observar, a utilização de estratégia sensível à história possibilitou melhorias nas três versões em que existia esse tipo de anomalia (v5, v6 e v7).

Uma observação importante no caso dessa anomalia é que as estratégias convencional e sensível à história se diferem apenas pelo acréscimo de uma única métrica sensível à história. Ou seja, propositalmente, consideramos as mesmas métricas e valores limites utilizados pela estratégia convencional e adicionamos uma única informação sensível à história, como pode ser observado no Capítulo 5. Isso nos possibilita resultados em que qualquer diferença entre as abordagens possa ser associada às influências meramente da informação sensível à história.

Comparando os resultados dessas estratégias, observamos que a informação sensível à história foi suficiente para apontar anomalias que não eram apontadas pela abordagem convencional em v5, v6 e v7. Ou seja, uma única

Detecção Convencional de SS						
	v2[1]	**v3**[1]	**v4**[1]	**v5**	**v6**	**v7**[2]
Acertos	n/a	n/a	n/a	1	1	1
Falsos Positivos	0	0	0	1	1	3
Falsos Negativos	0	0	0	2	2	2
Revocação	n/a	n/a	n/a	33%	33%	50%
Precisão	n/a	n/a	n/a	50%	50%	40%

Detecção Sensível à História de SS						
	v2[1]	**v3**[1]	**v4**[1]	**v5**	**v6**	**v7**[2]
Acertos	n/a	n/a	n/a	2	2	3 (4)
Falsos Positivos	0	0	0	1	1	3 (4)
Falsos Negativos	0	0	0	1	1	1 (0)
Revocação	n/a	n/a	n/a	67%	67%	75%(100%)
Precisão	n/a	n/a	n/a	67%	67%	50% (50%)

[1]n/a (não aplicável): não existem anomalias do tipo SS a serem detectadas nessas versões. [2]Em v7, os resultados entre parênteses são referentes à aplicação da estratégia SH no caso2 (Seção x).

Tabela 7.4: Detecção convencional vs. sensível à história de SS no Mobile Media

informação sensível à história possibilitou melhorias em 100% das versões que continham a anomalia. Se considerarmos as médias de precisão e revocação, também podemos observar os efeitos da utilização de informações sensíveis à história em detecções de *Shotgun Surgery*. As médias de precisão e revocação na detecção dessa anomalia são:

- Detecção Convencional de SS:
  média de revocação = 38.6%; média de precisão = 46.6%
- Detecção Sensível à História de SS:
  média de revocação = 78%; média de precisão = 61.3%

A abordadem sensível à história apresentou valores médios superiores tanto na revocação quanto na precisão. Com base nos dados apresentados, assim como ocorreu na detecção de GCs, podemos realizar a seguinte afirmação em relação à detecção de *Shotgun Surgery* no MM: *"Estratégias sensíveis à história podem ser mais eficazes que estratégias convencionais"*.

RQ3: Estratégias Sensíveis à História vs. Múltiplas Perspectivas Visuais

God Class. A Tabela 7.5 reúne os resultados de detecções de GCs baseadas em múltiplas perspectivas visuais e em estratégias sensíveis à história. Podemos observar que a utilização da ferramenta SourceMiner possibilitou a detecção de 100% das anomalias em todas as versões avaliadas pelos participantes (v2 à v7). Contudo, um efeito colateral desse número de acertos foi o

elevado número de falsos positivos. Em v6 e v7 os valores de precisão chegam a ser inferiores a 20%. Além disso, em nenhuma das versões analisadas obteve-se valores de precisão superiores a 40%. Já a aboragem sensível à história não possibilitou a totalidade de acertos como é verificado pela utilização das perspectivas visuais, contudo, o número de falsos positivos obtidos pela estratégia SH foi igual a zero em todas as versões.

Detecção por Múltiplas Perspectivas Visuais de GC						
	v2[1]	**v3**	**v4**	**v5**	**v6**	**v7**[1]
Acertos	n/a	2	2	1	2	2
Falsos Positivos	n/a	3	3	5	8	9
Falsos Negativos	n/a	0	0	0	0	0
Revocação	n/a	100%	100%	100%	100%	100%
Precisão	n/a	40%	40%	17%	20%	18%
Detecção Sensível à História de GC						
	v2	**v3**	**v4**	**v5**	**v6**	**v7**[1]
Acertos	1	1	1	1	2	0 (1)
Falsos Positivos	0	0	0	0	0	0 (0)
Falsos Negativos	1	1	1	0	0	2 (1)
Revocação	50%	50%	50%	100%	100%	0% (50%)
Precisão	100%	100%	100%	100%	100%	- (100%)

[1]n/a (não aplicável): essa versão não foi considerada no estudo exploratório utilizando a abordagem de detecção visual.

Tabela 7.5: Detecção baseada em múltiplas perspectivas visuais vs. sensível à história de GC no Mobile Media

Notamos que, enquanto a abordagem visual apresenta melhores valores de revocação, a sensível à história apresenta melhores valores de precisão. Isso nos faz acreditar que talvez as duas abordagens pudessem ser utilizadas de forma complementar. As informações sensíveis à história poderiam ser utilizadas para eliminação dos falsos positivos obtidos pela abordagem visual. Enquanto a abordagem visual contribuiria nas detecções não realizadas pela abordagem sensível à história. As informações sobre as médias de revocação e precisão para cada abordagem, destacam que, diferentemente do foi observado na análise anterior de RQ2, nenhuma das abordagens se apresentou superior em revocação e precisão ao mesmo tempo. Os valores médios de precisão e revocação são:

- Detecção por Múltiplas Perspectivas de GC:
  média de revocação = 100%; média de precisão = 27%
- Detecção Sensível à História de GC:
  média de revocação = 66%; média de precisão = 100%

Diante dos resultados apresentados pela Tabela 7.5, podemos realizar a seguinte afirmação para RQ3 em relação à detecção de GCs: *"Estratégias sensíveis à história provêem informações capazes de melhorar a eficácia de alguns recursos visuais de detecção?"*

Divergent Change. A Tabela 7.6 reúne os resultados de detecções de DCs utilizando-se múltiplas perspectivas visuais e estratégias sensíveis à história.

Detecção por Múltiplas Perspectivas Visuais de DC						
	v2[1]	**v3**	**v4**	**v5**	**v6**	**v7**[1]
Acertos	n/a	2	2	3	3	4
Falsos Positivos	n/a	2	2	6	9	10
Falsos Negativos	n/a	0	0	0	1	0
Revocação	n/a	100%	100%	100%	75%	100%
Precisão	n/a	50%	50%	33%	25%	29%
Detecção Sensível à História de DC						
	v2	**v3**	**v4**	**v5**	**v6**	**v7**[1]
Acertos	2	2	2	2	2	2 (2)
Falsos Positivos	0	0	0	0	1	2 (1)
Falsos Negativos	0	0	0	1	1	2 (2)
Revocação	100%	100%	100%	67%	67%	50% (50%)
Precisão	100%	100%	100%	100%	67%	50% (67%)

[1]n/a (não aplicável): essa versão não foi considerada no estudo exploratório utilizando a abordagem de detecção visual.

Tabela 7.6: Detecção baseada em múltiplas perspectivas visuais vs. sensível à história de DC no Mobile Media

A partir da Tabela 7.6, podemos notar uma situação bem semelhante a que observamos na seção anterior na avaliação de GCs. Nos ressultados referentes à detecção apoiada por recursos visuais observa-se que o número de acertos é melhor em três das versões consideradas (v5 à v7). Nas demais versões, de v2 à v4, o número de acertos foi igual em ambas as abordagens. Entretanto em todas as versões o número de falsos positivos apontados pelas estratégias sensíveis à história mais uma vez foi menor. O número de falsos positivos da abordagem sensível à história é nulo em 50% das versões (v2, v3, v4). Somando os falsos positivos da abordagem visual nas duas últimas versões temos um total de 19 falsos positivos contra 2 falsos positivos da abordagem sensível à história. Como valores médios de precisão e revocação, temos:

- Detecção por Múltiplas Perspectivas de DC:
  média de revocação = 95%; média de precisão = 37%
- Detecção Sensível à História de DC:
  média de revocação = 80.6%; média de precisão = 89%

Assim, como nas duas anomalias consideradas anteriormente, observamos que a abordagem sensível à história é mais eficaz que a abordagem visual no que diz respeito à precisão das detecções. Entretanto, a abordagem visual é mais eficaz no que diz respeito ao número de anomalias corretamente identificadas. Tais informações possibilitam também na detecção dessa anomalia a seguinte afirmação para RQ3: *"Estratégias sensíveis à história provêem informações capazes de melhorar a eficácia de alguns recursos visuais de detecção?"*. Tais resultados na detecção dessa anomalia também confirmam indicíos de que talvez seja interessante investir em estratégias híbridas em que uma minimize as possíveis deficiências da outra.

7.2.2 Segundo Sistema: Health Watcher

Como justificado na intrudução da Seção 7.2, não será possível discutir RQ2 (abordagem visual vs. abordagem sensível à história) no caso do sistema Health Watcher, nem avaliar Q1 (eficácia de estratégias sensíveis à história) e Q2 (estratégia convencional vs. sensível à história) para a anomalia *Shotgun Surgery*. Nesse caso, discutiremos Q1 em relação as anomalias *God Class* e *Divergent Change* e Q2 apenas para a anomalia GC, uma vez que não existe estratégia convencional para comparação dos resultados de detecção para DC.

Q1: Eficácia de Estratégias Sensíveis à História

A Tabela 7.7 apresenta os resultados da aplicação das estratégias sensível à história e convencional no sistema Health Watcher para a detecção de GC. As informações referentes à aplicação da estratégia convencional foram disponibilizadas para a discussão de Q2 a ser realizada na seção seguinte. Para a análise de Q1 consideraremos os dados da estratégia sensível à história apresentados na Tabela 7.7 bem como os dados sensíveis à história da detecção de DC apresentados na Tabela 7.8 Observamos que, enquanto no Mobile Media as estratégias sensíveis à história apresentaram-se bastante eficazes em todas as anomalias avaliadas, no Health Watcher elas não apresentaram-se tão eficazes.

Como pode ser observado na Tabela 7.7, a estratégia SH possibilitou a detecção de GCs em apenas 3 das 10 versões avaliadas. Além disso, nessas versões, v4, v7 e v9, todos os valores de revocação foram abaixo de 35%. Um ponto positivo, entretanto, é que nessas três versões a precisão foi de 100%. Ou seja, apesar de não ter possibilitado um bom número de acertos na detecção de GCs, a estratégia também não prejudica o desenvolvedor em relação a incidência de falsos positivos. Em todas as versões de v1 à v10 os falsos positivos para GC foram nulos. Apesar do número de falsos positivos

Detecção Convencional de GC										
	v1	**v2**	**v3**	**v4**	**v5**	**v6**	**v7**	**v8**	**v9**	**v10**
Acertos	0	0	0	0	0	0	0	0	0	0
Falsos Positivos	1	1	1	1	1	1	1	1	1	1
Falsos Negativos	4	4	4	4	3	3	3	3	3	3
Revocação	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Precisão	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Detecção Sensível à História de GC										
	v1	**v2**	**v3**	**v4**	**v5**	**v6**	**v7**	**v8**	**v9**	**v10**
Acertos	0	0	0	1	0	0	1	0	1	0
Falsos Positivos	0	0	0	0	0	0	0	0	0	1
Falsos Negativos	4	4	4	3	3	3	2	3	2	3
Revocação	0%	0%	0%	25%	0%	0%	33%	0%	33%	0%
Precisão	-	-	-	100%	-	-	100%	-	100%	0%

Tabela 7.7: Detecção convencional vs. sensível à história de GC no Health Watcher

ser nulo em quase todas as versões, o resultado das detecções para GC não foi considerado satisfatório, devido aos baixos valores de precisão e revocação como podemos observar na Tabela 7.7.

Na detecção de DCs, como pode ser observado na Tabela 7.8 também não foram obtidos elevados valores de precisão e revocação. A revocação máxima foi de 50%, obtida apenas em v5. Entretanto em nenhuma das versões obteve-se precisão ou revocação nulos. A revocação mínima foi de 33% enquanto a precisão mínima foi de 40%. Em resumo, a média de revocação foi de 38.9% enquanto a média de precisão foi de 58.5%. Enquanto a detecção de DC no MM reesultou em poucos falsos positivos, no HW, falsos positivos foram detectados em todas as versões. Além disso, enquanto no MM obteve-se precisão e revocação de 100% em três versões, no HW isso não foi possível em nenhuma das versões.

Como mencionamos na Seção 7.1.4, algumas diferenças existentes entre

Detecção Sensível à História de DC										
	v1	**v2**	**v3**	**v4**	**v5**	**v6**	**v7**	**v8**	**v9**	**v10**
Acertos	2	2	2	3	2	2	2	2	2	2
Falsos Positivos	1	2	1	2	2	1	1	1	2	3
Falsos Negativos	4	4	4	3	3	3	3	3	3	3
Revocação	33%	33%	33%	50%	40%	40%	40%	40%	40%	40%
Precisão	67%	50%	67%	60%	50%	67%	67%	67%	50%	40%

Tabela 7.8: Detecção sensível à história de DC no Health Watcher

o Mobile Media e Health Watcher foram consideradas importantes para essa avaliação, justamente para evitar replicações forçadas de bons ou maus resultados. Ou ainda para não haver o favorecimento de uma ou de outra abordagem. Era consideralvelmente importante para essa pesquisa avaliar a eficácia das estratégias em diferentes contextos de evolução. Evidências como as apresentadas através do estudo com o Health Watcher nos fazem suspeitar que provavelmente em sistemas com evolução pouco expressiva, isto é, em que as versões são geradas principalmente à partir de refatorações como no caso do HW, as estratégias sensíveis à história podem não apresentar bons resultados. De fato, os execelentes resultados de detecções sensíveis à história observados no primeiro sistema não foram igualmente constatados no segundo. Se nos baseássemos apenas nos resultados do HW não poderímaos afirmar que as estratégias sensíveis à história possibilitam detecções eficazes de anomalias (Q1).

RQ2: Estratégias Sensíveis à História vs. Estratégias Convencionais

God Class. Apesar da estratégia sensível à história não ter se apresentado muito eficaz na avaliação com o HW, pudemos observar que, ainda assim, tal estratégia apresentou-se mais eficaz que a estratégia convencional. Analisando a Tabela 7.7 podemos observar que enquanto a estratégia SH detectou falso positivo apenas em v10, com a aplicação da estratégia convencional falsos positivos foram detectados em todas as versões. Além disso, a estratégia convencional não conseguiu obter nenhum acerto em nenhuma das 10 versões avaliadas, enquanto a estratégia SH possibilitou acertos em v4, v7 e v9. Nesse caso, apesar do resultado não satisfatório da estratégia SH, pudemos observar que na valiação com o HW também podemos realizar a seguinte afirmação em relação a RQ2: *"Estratégias sensíveis à história podem ser mais eficazes que estratégias convencionais"*. Tal afirmação também pode ser justificada pelos valores médios de revocação e precisão obtidos na detecção de GC nesse sistema. Eles são:

- Detecção Convencional de GC:
  média de revocação = 0%; média de precisão = 0%
- Detecção Sensível à História de GC:
  média de revocação = 10%; média de precisão = 33%

7.2.3 Resumo Geral dos Resultados

A avaliação apresentada neste trabalho visou identificar as possíveis contribuições de se utilizar informações sobre a evolução das características do código na detecção de anomalias de modularidade de código. Pudemos observar que estratégias sensíveis à história apresentaram resultados superiores aos de estratégias convencionais nos dois sistemas considerados (Q2). Além disso, que em sistemas com características de evolução mais expressivas, ou seja, com um maior número de alterações realizadas entre as versões, estratégias SH tendem a se apresentar bastante eficazes (Q1). Isso foi verificado no estudo com o Mobile Media.

Já em sistemas em que as evoluções entre as versões são pouco significativas, ou seja, são poucas as mudanças realizadas entre as versões, as estratégias sensíveis à história tendem a não apresentar resultados eficazes, mas que ainda assim podem ser melhores que o de estratégias convencionais. Isso foi observado no estudo com o Health Watcher. Durante esta avaliação, observamos ainda que ao compararmos as detecções baseadas em recursos visuais com as de estratégias sensíveis à história, nenhuma se apresentou totalmente superior a outra.

A abordagem visual obteve maiores números de acertos, contudo, como efeito colateral desses acertos apresentou sempre um grande número de falsos positivos. Em contrapartida, a abordagem sensível à história, em algumas versões não conseguia alcançar a totalidade de acertos obtidos pela abordagem visual, mas apresentava valores nulos de falsos positivos ou sempre menores que os da abordagem de detecção visual. As comparações entre os resultados da abordagem visual com os da sensível à história, nos chamou a atenção para as possibilidades de se considerar abordagens híbridas de detecção em que uma minimize as deficiências da outra.

Em resumo, todos os resultados apresentados nesta avaliação fornecem indícios de que a utilização de informações sensíveis à história no contexto de detecção de anomalias merece ser explorado. As métricas e a ferramenta proposta neste trabalho fornecem alguns recursos iniciais que permitem a exploração de diversas configurações de estratégias para a continuidade de pesquisas nesse contexto.

7.2.4 Ameaças à Validade dos Resultados

Apesar das análises importantes realizadas considerando Q1, Q2 e Q3, algumas limitações podem ser identificadas nesse processo de avaliação.

Número de Sistemas e de Versões Avaliadas: Este estudo considerou a avaliação da abordagem em dois sistemas, Mobile Media e Health Watcher. Esse número pode ser considerado baixo para um processo de avaliação que se propõe a analisar exatamente as contribuições de detecção automática de anomalias de código. Entretanto, foram consideradas 6 versões do Mobile Media e 10 versões do Health Watcher, totalizando um total de 16 versões analisadas. Essas 16 versões poderiam ser vistas em avaliações de outras abordagens como 16 aplicações independentes, pois as detecções foram realizadas em cada uma de forma individual. Entretanto, como se trata de uma abordagem sensível à história seria interessante se tivéssemos considerado sistemas com um maior número de versões. Isso traria uma maior número de informações em relação à possíveis degradações do código ao longo das versões e maiores informações sobre as informações histórias em sistemas que pasaram por um grande número de mudanças.

Tipo de Sistema Avaliado: o Mobile Media foi desenvolvido sobre o paradigma de Linhas de Produto com implementação baseada em compilação condicional. Tal fato pode levar a crer que os resultados de métricas podem ter sido diretamente influenciados pelo tipo desse sistema. Entretanto, seria apenas uma questão relacionada aos valores limites considerados em métricas de tamanho, complexidade ou outras. Acreditamos que em um sistema desse mesmo domínio e que não estivesse desenvolvido como uma linha de produto, as únicas alterações possivelmente necessárias seriam as de valores limires. Entretanto, não constatamos tal informação em outros tipos de sistemas com características evolutivas semelhantes aos do Mobile Media.

Métricas Utilizadas: algumas métricas convencionais utilizadas na avaliação como LOC e WMC são constantemente criticadas por pesquisadores. Na verdade, resultados obtidos por métricas de tamanho podem ser difíceis de interpretar pois altos valores podem ser resultantes da formatação do código. Entretanto, a influência na qualidade das estratégias relacionadas a um maior ou menor número de linhas de código só interfere no ajuste de valores limites na utilização das estratégias propostas. Como esses valores não foram prédefinidos, dificilmente o uso dessas métricas poderia impactar como fator principal na obtenção dos resultados. Além disso, essas métricas não foram utilizadas de forma isolada, o que mimiza as possibilidades de interpretações indevidas relacionadas a elas.

Cálculo das Métricas Convencionais: as métricas convencionais utilizadas como base da maioria das estratégias sensíveis à historia foram calculadas pela ferramenta Together. Assumimos a corretude dos resultados apresentados por tal ferramenta. Entretanto, a geração de valores de métricas convencio-

nais por uma outra ferramenta talvez pudesse trazer valores diferenciadas que possivelmente trariam diferenças nos resultados das estratégias. Mais uma vez, consideramos que esse também seria um caso que possivelmente seria resolvidos com ajustes de valores limites.

Valores Limites Utilizados na Estratégias Convencionais: Na avaliação das estratégias convencionais, os valores limites não foram explorados empiricamente. Ao invés disso, utilizamos os valores limites sugeridos pelos proponentes das estratégias convencionais. Como utilizamos sistemas diferentes dos considerados por esses autores na época em que eles avaliaram suas estratégias, é possível que os limites tenham interferido na eficácia das detecções convencionais. Por considerar tal fato, entramos em contato com tais autores, por e-mail, para que pudéssemos replicar nossa avaliação utilizando os mesmos sistemas utilizados por eles quando foi proposta a abordagem convencional. Como os sistemas utilizados não eram gratuitos, esses não puderam nos ser disponibilizados para replicarmos tal avaliação.

Opinião dos especialistas sobre as anomalias reais: apesar dos oráculos terem sido resultantes de inspeções manuais, ainda assim tal processo é passível de erros. Não se pode garantir que todas as anomalias existentes nos sistemas foram identificadas. Além disso as definições textuais dessas anomalias podem gerar diferentes interpretações. Outros oráculos com módulos diferentes dos considerados como anômalos poderiam influenciar diretamente nos resultados. O recurso que utilizamos para tentar minimizar tal problema foi trabalhar com a opinião de dois especialistas. Só consideramos como anômalos os módulos apontados pelos dois especialistas simultaneamente.