Automata Language Equivalence vs. Simulations for Model-based Mutant Equivalence: An Empirical Evaluation

Xavier Devroey, Gilles Perrouin, Mike Papadakis, Axel Legay, Pierre-Yves Schobbens, Patrick Heymans

Automata Language Equivalence *vs.* Simulations for Model-based Mutant Equivalence: An Empirical Evaluation

Xavier Devroey*, Gilles Perrouin*, Mike Papadakis†, Axel Legay‡, Pierre-Yves Schobbens*, and Patrick Heymans*
*PReCISE Research Center,University of Namur, Belgium, Email: *firstname.lastname*@unamur.be
†SnT, SERVAL Team, University of Luxembourg, Email: michail.papadakis@uni.lu
‡INRIA Rennes, France, Email: axel.legay@inria.fr

***Abstract*—Mutation analysis is a popular technique to assess the effectiveness of test suites with respect to their fault-finding abilities. It relies on the mutation score, which indicates how many mutants are revealed by a test suite. Yet, there are mutants whose behaviour is equivalent to the original system, wasting analysis resources and preventing the satisfaction of the full (100%) mutation score. For finite behavioural models, the Equivalent Mutant Problem (EMP) can be addressed through language equivalence of non-deterministic finite automata, which is a well-studied, yet computationally expensive, problem in automata theory. In this paper, we report on our assessment of a state-of-the-art exact language equivalence tool to handle the EMP against 12 models of size up to 15,000 states on 4710 mutants. We introduce random and mutation-biased simulation heuristics as baselines for comparison. Results show that the exact approach is often more than ten times faster in the weak mutation scenario. For strong mutation, our biased simulations are faster for models larger than 300 states. They can be up to 1,000 times faster while limiting the error of misclassifying non-equivalent mutants as equivalent to 8% on average. We therefore conclude that the approaches can be combined for improved efficiency.**

***Keywords*-model-based mutation analysis; automata language equivalence; random simulations**

I. Introduction

Mutation analysis is a technique that injects artificial defects, called *mutations*, into the code under test, yielding *mutants*. Mutants are typically used for evaluating the effectiveness of test suites [1]–[3] and to support test generation [2], [4], [5]. The technique is quite popular in research due to the ability of mutants to simulate the behaviour of real faults [1], [6]. There is also evidence showing that tests designed to detect mutants reveal more faults than other testing criteria [2], [7].

These characteristics of mutation inspired researchers to apply the method on artefacts other than code and particularly models [2], [8]. The usual advantages of model-based testing techniques are the ability to identify defects related to missing functionality or misinterpreted specifications [9] where code-based testing fails [10], [11]. The method proved to be so powerful that it could complement existing methods. For instance, Aichernig *et al.* [12] report that model mutants lead to tests that are able to reveal implementation faults that were found neither by manual tests, nor by the actual operation, of an industrial system.

Despite its potential, mutation analysis faces a number of challenges that currently prevent wider adoption [13], [14]. One of them is the *Equivalent Mutants Problem* (EMP). It concerns the mutants whose behaviour is identical to the original artefact (code or model). Such mutants cannot be distinguished by any test, a situation that raises two issues: (i) they hamper the use of the criterion as a stopping rule by skewing the mutation score measurement (the number of detected mutants divided by the total number of mutants), and (ii) they do not bring any new value to the test generation techniques as they attempt to kill mutants that have no chance to be killed.

In this paper, we focus on the model-based formulation of the EMP, which can be expressed in terms of language equivalence. Language equivalence has been studied by the formal verification community who determined its P-SPACE complexity [15] and derived exact equivalence checking algorithms [16], [17]. While potentially helpful, such tools have, to our knowledge, never been used to tackle the EMP. This is the main contribution and novelty of this paper.

In summary, the contributions of this paper are:

- The design of two simulation algorithms relying either on random simulations (RS) or biased simulations (BS) covering infected states [18] (*i.e.*, exploiting syntactical differences between original and mutant models) to improve the chances to distinguish non-equivalent mutants;
- A configurable implementation of our simulations (available at https://projects.info.unamur.be/vibes/) that benefits from the fact that simulation can be easily distributed amongst processor cores;
- The definition of an experimental setup to apply an automata language equivalence tool (ALE) [16] to the EMP. We employed twelve models of varying origins and sizes, from nine to 15,000 states. We produced 4710 mutants using seven operators, and considered four mutation orders (one, two, five, ten), according to strong and weak mutation scenarios.
- The assessment of the ALE tool with respect to our baseline algorithms. We measured the speed and accuracy of equivalence detection. The ALE tool is particularly efficient for weak mutation by being, on average, ten times faster than simulations. However, biased simula-

tions perform well for strong mutation on models larger than 300 states: they can be 1,000 times faster. The ratio of tagging non-equivalent mutants as equivalent is 8% for biased simulations and 15% for random ones. To ease reproducibility, all our models and experimental results are available at: https://projects.info.unamur.be/vibes/mutants-equiv.html.

The remainder of the paper is organised as follows. Section II presents background information on the models used and language equivalence, while Section III details the design of our simulation heuristics and the ALE approach we used. Section IV describes our empirical assessment and provides some lessons learned. Section V covers relevant literature. Finally Section VI, wraps up the paper.

II. Background

In this section we introduce the main formalism dealt with in this paper, namely, finite transition systems, and the relevance of language equivalence for equivalent mutant detection.

A. Transition Systems & Finite Automata

Our research in model-based testing considers transition systems as a powerful abstract formalism to model system behaviour. Our definition is adapted from Baier and Katoen's book [19], where atomic propositions have been omitted (we do not consider state internals):

Definition 1 (Transition System (TS)): A TS is a tuple $(S, Act, trans, i)$ where S is a set of states, Act is a set of actions, $trans \subseteq S \times Act \times S$ is a non-deterministic transition relation (with $(s_1, \alpha, s_2) \in trans$ sometimes noted $s_1 \xrightarrow{\alpha} s_2$), and $i \in S$ is the initial state.

To deal with test generation activities, where finite behaviours are sought, we first require that sets S and Act are finite. To mimic weak and strong mutation scenarios (see Section III-A), we will stop our executions in specific states. These additional requirements make our execution semantics equivalent to that of usual non-deterministic finite automata (NFA), thereby enabling the comparison of our simulations to ALE tools. In the remainder of this paper, unless otherwise stated, we will always refer to TSs with such restrictions so that the term can be used interchangeably with NFAs[1].

Definition 2 (Trace): Let $ts = (S, Act, trans, i)$ be a TS, let $t = (\alpha_1, \ldots, \alpha_n)$ where $\alpha_1, \ldots, \alpha_n \in Act$ be a finite sequence of actions. The trace t is valid iff:

$$ts \overset{(\alpha_1,\ldots,\alpha_n)}{\Longrightarrow} s \text{ with } s \subset S,$$

where $ts \overset{(\alpha_1,\ldots,\alpha_n)}{\Longrightarrow}$ is equivalent to $\exists s \in S : i \overset{(\alpha_1,\ldots,\alpha_n)}{\Longrightarrow} s$, meaning that there exists a non-empty sequence of transitions labelled $(\alpha_1, \ldots, \alpha_n)$ from i to a state s of the TS.

[1]Our MBT framework, ViBES, uses TSs as its underlying formalism so we stick to the term "TS" for consistency.

B. Equivalent Mutant Problem

In this paper, we focus on the model-based instance of the Equivalent Mutant Problem (EMP). The equivalent mutant problem is a well-known issue in mutation analysis [13], [14]. It stems from the fact that two program variants may exhibit the same behaviour and therefore cannot be distinguished by test cases. This is particularly problematic with respect to both generation and assessment of test suites, since 100% of killed mutants is impossible to reach in case of equivalence (also the EMP leads to wasting resources spent on assessing "useless" mutants). Mutant equivalence can take two forms [13]: (a) equivalence between mutants and the original system; (b) equivalence between two mutants (not with the original system). Mutants of case (a) are called "equivalent" while mutants of case (b) are called "duplicate". In the context of this paper, we focus on mutants that are behaviourally equivalent to the original system, *i.e.*, mutants of case (a).

C. Automata Language Equivalence & EMP

In our context, the EMP corresponds to a classic problem in automata theory: *Automata Language Equivalence* (ALE). The accepted language of an automaton is formed by all the sequences of actions (words) that can be accepted *i.e.*, starting in the initial state and ending in a final state. Therefore, if a mutant m accepts the same language as the original o (language-equivalent), then there is no trace t that can distinguish the mutant from the original: $\forall t, t \in \mathcal{L}(o) \Leftrightarrow t \in \mathcal{L}(m)$.

There are various forms of relations defined between two automata that we can compute to determine whether they are language-equivalent. Among them, we can cite bisimulations or trace equivalence [19]. In the last years, the verification community came up with dedicated algorithms such as bisimulations up to congruence [16] or antichains [17] to address language equivalence. In model-based mutation testing, Aichernig *et al.* investigated language inclusion (but not equivalence) using refinement checking [20] in order to generate mutant-killing test cases.

Although tackling the language equivalence and inclusion problems from different angles and heuristics, all these techniques may face exponential blow-up since both language inclusion and equivalence were demonstrated to be P-SPACE complete [15]. While worst-case complexity can seem discouraging, various heuristics have been proposed to limit the effects of this complexity in practice. One of the goals of this paper is to determine the applicability of an exact language equivalence algorithm to address the EMP [16]. The algorithm selected due to its availability, reported performance over the state of the art and ability to handle non-determinism that mutations may incur. In the next section, we also present two baseline algorithms that run generated traces to distinguish original and mutants' behaviours.

III. Mutant Equivalence Analysis

A. Strong and Weak Mutation

Elizabeth Jöbstl [21] discussed the conditions, identified by DeMillo and Offutt [22], that must be fulfilled to kill a mutant:

(i) "the *necessity condition* says that the state of the mutated program after some execution of the mutated statement must be incorrect with respect to the original program. This implies that the mutated statement must be reached. This is necessary, but not sufficient"; (ii) "the *sufficiency condition* says that the final state of the mutant must differ from the final state of the original program, i.e., the necessary incorrect intermediate state must propagate to an incorrect final state." Satisfying the necessity condition alone is referred to as *weak mutation* [23], while satisfying both is *strong mutation*.

At the model level, our simulations detect an incorrect state if a trace that is valid with respect to the original TS is invalid on the mutant TS, and vice-versa. Indeed, when executed, a trace induces one or more *runs* (alternating sequences of states and actions), depending on the presence of non-determinism. If such a run does not contain all the actions of the trace (*i.e*, the run is *incomplete*), it is because of the presence of an incorrect state preventing the subsequent actions to be fired. If all runs are complete, the original and the mutant are assumed equivalent for this trace. Necessity and sufficiency conditions affect the final states of these runs. For weak mutation, these states can map to any state of the TS. For strong mutation, we need to account for the fact that TSs have no final states. A very frequent example is the modelling of user sessions in which, after a legitimate sequence of actions, the system returns to its initial state to welcome a new user. This occurs in two thirds of the systems we analyse in Section IV-A1. This is why we model strong mutation by generating traces whose runs start and end in the same initial state.

The ALE approach uses automata that have explicit initial and final states. For weak mutation, we generate automata in which all states are final, and for strong mutation the initial state is the only final state.

B. Automata Language Equivalence (ALE)

The ALE approach we selected for comparison is developed by Bonchi and Pous [16]. It can be thought of an extension to non-deterministic TSs of the Hopcroft-Karp algorithm. In particular, they introduce a new bisimulation relation called *up to congruence* that requires to explore less states than the original algorithm. This approach also avoids to build the complete deterministic finite TS and performs determinisation on-the-fly. This makes such an approach particularly relevant: (i) non-determinism may be introduced locally by mutations (our original models are deterministic), thereby limiting determinisation scope, and (ii) between 0% and 15.5% of our mutants are non-deterministic (see Section IV-A1).

C. Random and Biased Simulation

Our randomized approach to equivalence analysis is straightforward: we generate random traces from the original model and run them on the mutant model and reciprocally. If a trace fails to execute on one of the models, it serves as a counterexample and disproves equivalence. If all runs succeed, then the mutant is considered *probably equivalent* and testers have to decide if they want to perform more simulations or switch to an exact method. Algorithm 1 presents our generic simulation approach: N traces are selected (resp.) from the original model (line 1) and the mutant model (line 7), and executed (resp.) on the mutant model (line 3) and the original model (line 9). In case of non deterministic behaviour, all the possible paths are considered for the execution of the trace. If one execution fails, the algorithm stops and returns a *positive* trace such as $(o \stackrel{t}{\Longrightarrow}) \wedge \neg(m \stackrel{t}{\Longrightarrow})$ (line 4) or a *negative* trace such as $\neg(o \stackrel{t}{\Longrightarrow}) \wedge (m \stackrel{t}{\Longrightarrow})$ (line 10) .

Algorithm 1 Generic simulation

Require: $o : TS$ {the original system}
$m : TS$ {the mutant to compare to o}
N {total number of traces to generate}
k {trace length}
Ensure: returns a positive or negative trace differentiating m from o or a special value ($none$) if m is equivalent to o.

1: $traceset \leftarrow select(o, \frac{N}{2}, k)$
2: **for all** $t \in tracetset$ **do**
3: **if** $\neg(m \stackrel{t}{\Longrightarrow})$ **then**
4: **return** $pos(t)$ {if the mutant TS fails to execute t, returns a positive trace t}
5: **end if**
6: **end for**
7: $traceset \leftarrow select(m, \frac{N}{2}, k)$
8: **for all** $t \in traceset$ **do**
9: **if** $\neg(o \stackrel{t}{\Longrightarrow})$ **then**
10: **return** $neg(t)$ {if the original TS fails to execute t, returns a negative trace t}
11: **end if**
12: **end for**
13: **return** $none$

This generic simulation algorithm is instantiated through two strategies for trace generation (lines 1 and 7): *Random Simulation* (RS) and *Biased Simulation* (BS). The parameter N is computed using the Chernoff-Hoeffding bound as explained hereafter.

1) Random Simulation (RS): Random simulation (RS) assumes a uniform distribution of traces over the model, that is, such traces are selected randomly ($select$ call on lines 1 and 7 in Algorithm 1) by accumulating the actions α_i triggered by a random walk of a given length $\leq k$ in the TS. For weak mutation (WM RS), the only constraint is to start the random walk from the initial state i. Strong mutation (SM RS) requires a random walk starting from and ending in i: after few tries, this method (*i.e.*, using a random walk until the initial state i is reached) showed very poor results on our largest models (we set a timeout of one hour for one equivalence detection) and is therefore not further discussed in this paper.

2) Biased Simulation (BS): The biased simulation (BS) approach exploits the basic characteristics of mutation testing: mutations are localised and they create (most of the time) behavioural differences. It assumes that those differences are detected by a trace t which, when executed on the original TS

o or on its mutant m, goes through one of the states affected by the mutation. For instance, the transition missing (TMI in Table II) operator produces a mutant by removing a transition $a \xrightarrow{\alpha_i} b$ from the original TS. The BS approach generates traces in o and m, such that their executions $m \stackrel{t}{\Longrightarrow}$ or $o \stackrel{t}{\Longrightarrow}$ cover a or b. Such states, called *infected states*, have been shown to help identifying equivalent mutants at the code level [24], [25] and to speed up mutation analysis at the model level [26]. This motivates us to adopt this strategy in our biased simulation.

In practice, the set of infected states S_{infect} is computed by checking syntactic differences between the original and mutant TSs. It will include: (i) connected states (i.e, states accessible from the initial state) from one model which are not present in the other, and (ii) states with differences in their input/output transitions: in number of transitions or in action names, considering any pair of states $< s_o, s_m >$ where s_o is a state in the original TS, s_m a state in the mutant TS, such that their names are identical. An alternative is to instrument the mutant generator to keep track of the list of infected states while generating the mutants. Our goal is to be able to apply this strategy without any information on how the mutants are generated (*e.g.*, generated by other frameworks than ours) and to fairly compare with an exact approach that makes no assumption on the locality of differences. Once the set of infected states S_{infect} is obtained (by any means), the second step is to generate traces that cover such infected states.

For weak mutation (WM BS), a trace t is selected ($select$ call on lines 1 and 7 in algorithm 1) by concatenating the actions of (i) the shortest walk from the initial state i to a randomly chosen state $a \in S_{infect}$ and (ii) a random walk starting from a. To proceed, the first step during trace generation is to compute the shortest distance (*i.e.*, the number of transitions) between each state of the original TS o (or its mutant m resp.) and the initial state i of o (or m resp.) using a standard breadth-first search [27]. For strong mutation (SM BS), instead of a random walk starting from a, the algorithm will consider the actions of a path starting from a and returning to i using the computed shortest distance: the distance from a to i will (not strictly) decrease each time a transition is taken in the path.

3) Estimating the Number of Required Runs: An important parameter for simulation is the number of runs N. Under the hypothesis that traces are uniformly distributed we can bound the equivalence probability and estimating the number of runs needed achieve these bounds. Herault *et al.* [28] suggested to use the Chernoff-Hoeffding bound to estimate the number N of required runs to limit the equivalence probability depending on the approximation parameter $\epsilon > 0$ and a confidence parameter $\delta < 1$. If $N \geq \frac{4\log(2/\delta)}{\epsilon^2}$ then we have: $Pr\left[equiv(m,o)\right] = Pr\left[\left|\frac{A}{N} - p\right| \leq \epsilon\right] \geq 1 - \delta$. Where A is the number of successful runs that is either $m \stackrel{t}{\Longrightarrow}$ or $o \stackrel{t}{\Longrightarrow}$ for a given trace t. In practice, we compute A/N only when the algorithm has exhausted all the runs and set $N = \frac{8\log(2/\delta)}{\epsilon^2}$ for the number of runs as we have to account for two-way simulation: the number of runs is thus doubled.

It has to be observed that regarding biased simulations, the distribution of traces will not be uniform as the infected states "force" traces to explore only given portions of the model, *viz.* where the mutations are. Although this inequality may not hold in this case, we alleviate this threat by not trying to interpret the δ and ϵ values for biased simulations: they are for us a convenient means to compute N. Furthermore, keeping the same number of runs for random and biased simulations allows comparing their execution times and recalls.

TABLE I
MODELS CHARACTERISTICS

Model	States	Trans.	Act.	Avg. deg.	BFS height	Back lvl tr.
S. V. Mach.	9	13	14	1.44	5	3
C. P. Term.	16	17	15	1.55	7	4
Minepump	25	41	23	4.64	15	9
Claroline	106	2,055	106	19.37	1	105
Elsa-RR	384	1,214	384	3.16	194	174
Elsa-RRN	615	1,771	615	2.88	369	289
AGE-RR	772	6,639	772	8.60	328	408
AGE-RRN	1,101	10,960	1,101	9.96	426	662
Random 1	10,000	13,652	120	1.37	7,924	3,303
Random 2	15,000	20,488	300	1.37	11,865	4,899
Random 3	15,000	20,488	210	1.37	11,865	4,899
Random 4	15,000	20,488	150	1.37	11,865	4,899

TABLE II
TRANSITION SYSTEM MUTATION OPERATORS

SMI	State Missing operator removes a state (other than the initial state) and all its incoming/outgoing transitions.
WIS	Wrong Initial State operator changes the initial state.
AEX	Action Exchange operator replaces the action linked to a given transition by another action.
AMI	Action Missing operator removes an action from a transition.
TMI	Transition Missing operator removes a transition.
TAD	Transition Add operator adds a transition between two states.
TDE	Transition Destination Exchange operator modifies the destination of a transition.

IV. EMPIRICAL ASSESSMENT

This section presents our empirical assessment of the ALE, RS, and BS approaches. We define the following research questions:

RQ1 *What is the impact of weak and strong mutation on BS/RS vs. ALE performance?*

RQ2 *How many non-equivalent mutants are effectively detected by the RS and BS approaches?*

RQ3 *What are the worst case execution times for the ALE and BS/RS approaches?*

A. Protocol

To answer these RQs, we consider several models of different kinds of systems and apply the following procedure to each of them: (i) we generate a set of mutants from the model using the operators presented in Table II for orders 1, 2, 5, and 10; (ii) for each order, we sample 100 non-equivalent mutants (using the ALE algorithm to guarantee non-equivalence) to form the mutant set M; (iii) for each mutant in M, we measure the execution time and result of: 3 executions of weak mutation

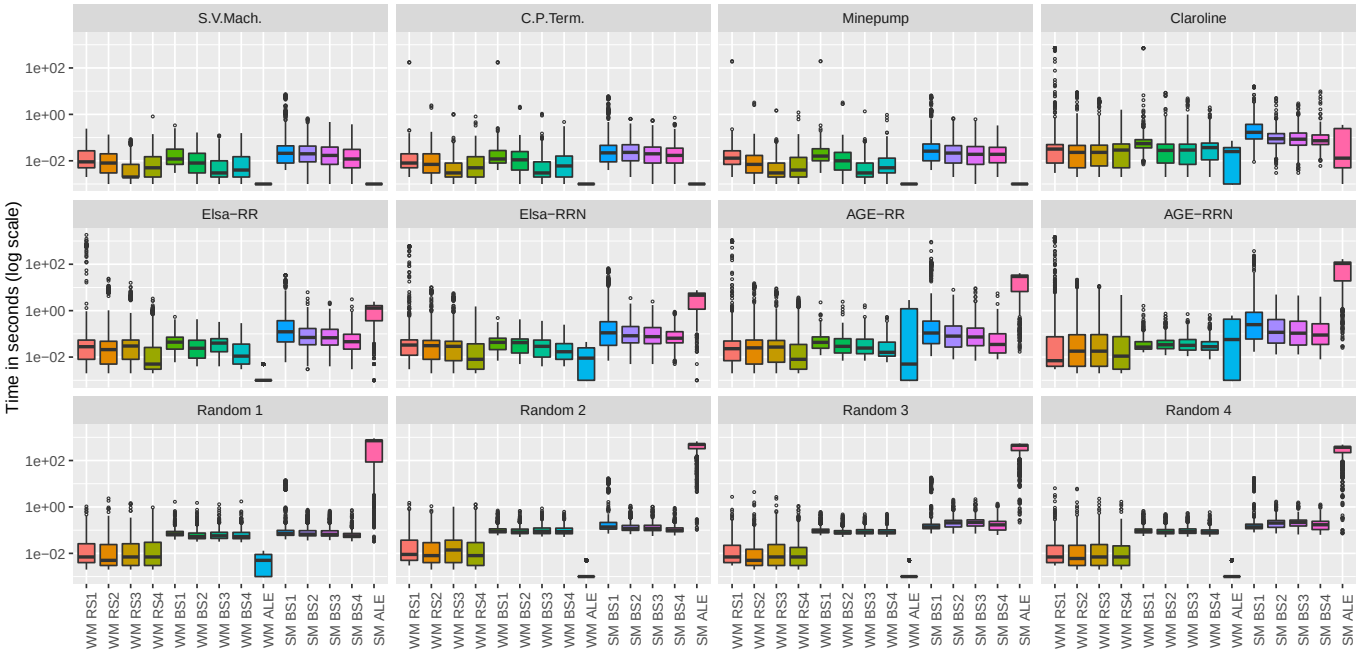

Fig. 1. Execution time of the equivalent mutant detection

random and biased search (WM RS/BS), and 3 executions of strong mutation-biased search (SM BS) algorithms[2] with 4 different values of δ and ϵ; and the executions of the ALE algorithm. In the following we detail the different steps of the procedure. The assessment has been performed on a Debian 3.16.7 x86_64 GNU/Linux running on a 16 cores, 2.2 GHz, 16Gb RAM virtual machine.

1) Models: We carry out the assessment on 12 different models coming from different sources and with varying size detailed in Table I. The different characteristics considered are: the number of states (*St.*); the number of transitions (*Tr.*); the number of actions (*Act.*); the number of incoming plus outgoing transitions per state (*Avg. deg.*); the maximal number of states between the initial state and any other state when traversing the TS in breadth-first search (*BFS h.*); the number of transitions whose source state has a higher *BFS h.* value than its destination state (*Back lvl tr.*). The models are: the soda vending machine model (*S.V.Mach.*), a small example describing the behaviour of a machine selling soda and tea [29]; the card payment terminal (*C.P.Term.*), also a small example describing the behaviour of a terminal used in a store to pay by card; the mine pump (*Minepump*), a well-known specification exemplar that models the behaviour of a pump keeping a mine safe from flooding by pumping water from a sink while avoiding methane explosions [29]; the Claroline website (*Claroline*), representing the navigational usages of a real online course management platform. The latter was reverse-engineered from an Apache log using a 2-gram inference method from Sprenkle *et al.* [30]; WordPress models (*AGE-RR*, *AGE-RRN*, *Elsa-RR*, and *Elsa-RRN*) that represent the navigational usages of two different real WordPress instances. They were also reverse-engineered using the same 2-gram inference method [30]. the *AGE-RR* and *Elsa-RR*, we considered only request type (*e.g.*, POST, GET, HEAD) and the requested resource (*e.g.*, "/index.php") in the sequences used. For the *AGE-RRN* and *Elsa-RRN* models, we considered request type, requested resource, and parameter names (*e.g.*, "?page=") in the sequences used as input of the 2-gram inference method. The random models (*Random 1* to *Random 4*) were generated according to the following procedure: (i) generate a set of random oriented graphs and compute the different measures from Table I (except number of actions); (ii) select those graphs that are likely to represent a real system according to Pelánek [31], *i.e.*, those having a small average degree, a large BFS height and a small number of back level edges (in this order); (iii) apply a random labelling multiple times and compute the occurrence probability, *i.e.*, the probability of the labels to obtain a set of randomly generated TSs; (iv) select the TS that has the following properties[3]: the probability of the most frequently occurring label in the TS is less than, or equal to, 6%, and the cumulated probability of the 5 most frequently occurring labels is less than or equal to 20% [32]. We end up with 4 random models as recorded in Table I.

2) Mutant Generation and Sampling: First-order mutants are generated using the operators presented in Table II. Each operator is applied (arbitrarily) 10 times on the *S.V.Mach.*, *C.P.Term.*, and *Minepump* models. Due to the small size of the models, applying the same mutation operator more than 10 times is not relevant. Operators are also applied (arbitrarily) 500 times on the other models. In the same way, *N*-order

[2]As explained in section III-C1, SM RS is not considered for the assessment due to the poor results during our initial attempts.

[3]These properties are likely to represent real systems [31]

mutants (with N equal to 2, 5, or 10 in our case) are generated by applying the same operators 10 or 500 times (depending on the model) on $(N-1)$-order mutants. After the generation, we perform a random sampling of 100 mutants (when available) for orders 1, 2, 5, and 10, giving us a set M with 370 mutants for the *S.V.Mach.*, *C.P.Term.*, and *Minepump* models, and 400 mutants for the other models. To ease mutant generation, we use our compact representation [33].

3) Non-determinism: We checked all the 4710 mutants and found that only 3.54% of them are non-deterministic. Nevertheless, there is a great disparity amongst models as the non-determinism rate varies from 0% for *Elsa-RRN* to 15.5% for *Claroline*. Higher-order mutation greatly influenced non-determinism rates: the sole order 10 is responsible for 53% of all non-deterministic mutants. In terms of mutation operators, TAD accounts for a large majority of non-deterministic first-order mutants (78%) and AEX for the remaining 22%. At higher orders, these two operators are largely involved. They are absent only in the *Minepump* model where TDE and AMI appear for two non-deterministic mutants.

4) Algorithm Execution: To run the language equivalence algorithms (for WM and SM), we use the HKC library [34], an OCaml implementation of the ALE algorithm [16] compiled using OCamlbuild. This tool handles non-deterministic TSs using different strategies: the automata may be processed either forward of backwards, and the exploration strategy may be breadth-first or depth-first. For each mutant, we execute the HKC library using each of the 4 possible configurations. The input models and their mutants have been transformed from our XML format to the Timbuk input format supported by HKC.

The random and biased simulation algorithms are implemented in Java using multi-threading to parallelize trace selection and execution as described in Algorithm 1 (lines 1, 3, 7, and 9). In our experiments, we set up the algorithm with 4 threads and run 4 instances in parallel on our virtual machine with 16 cores. We run the simulation algorithms with 4 different values of δ and ϵ determining the number of traces selected and executed (N in Algorithm 1):

- RS1/BS1: ($\delta = 1e-10$, $\epsilon = 0.01$, $N = 1,897,519$);
- RS2/BS2: ($\delta = 1e-10$, $\epsilon = 0.1$, $N = 18,975$);
- RS3/BS3: ($\delta = 1e-5$, $\epsilon = 0.1$, $N = 9,764$);
- RS4/BS4: ($\delta = 1e-1$, $\epsilon = 0.1$, $N = 2,396$).

For all the simulation configurations and all models, we fixed the trace length k to 3,000, which was our compromise between performance and non-equivalence detection: setting k to BFS height led to crashes in some cases. In order to answer RQ3, we also run each algorithm (RS1/BS1 to RS4/BS4, plus the 4 possible ALE configurations) with the model itself as the "mutant". Those (unrealistic) equivalent detection runs between the model and itself are only used to approximate the worst computation time of the different algorithms.

B. Results and Discussion

1) Random/biased simulations and ALE - Answering RQ1: Figure 1 presents the execution time per mutant of the studied

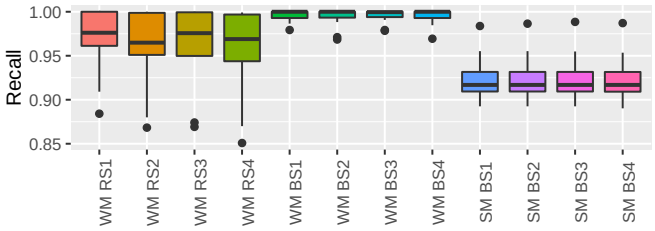

Fig. 2. Recall

algorithms, which is detailed in the Appendix. Regarding weak mutation scenarios, the ALE approach is the fastest in all cases in eleven of our models. On the *AGE-RN* model, biased simulations are faster for the largest numbers of runs. However, the results are at the limit of non-significance (see Table III), so that the only clearly significant result is for *BS1* on this model. For *AGE-RNN*, execution times for biased simulations are non-significant. Random simulations are also faster than ALE on *AGE-RRN* but only certain settings are significant. We thus conclude that the ALE approach is more interesting in terms of execution time. When we compare the two forms of simulations, for the smallest models, biased simulations are either on par for the smallest models or slightly better. Additional computations such as the breath-first search used for biased simulation do not cause significant overhead. For the largest random models, random simulations are faster. In these cases, the overhead of computing infected states and paths that cover these states is greater and random simulation is faster. However, lower standard deviations for biased simulation execution times over random ones make the BS approach easier to use.

Regarding strong mutation, several observations can be made. First, random simulations provide very high execution times compared to biased simulations or the ALE algorithm (the analysis of one model is stopped after one hour). This may be due to the difficulty to reach the initial state again when performing random walks in the TSs. Second, biased simulations are faster than ALE executions for models larger than 300 states. On the largest models, biased simulations can be up to 1,000 times faster. We thus conclude that these are the most interesting situations in which to use BS, for mutation analysis. On smaller models, the ALE algorithm's performance is quite impressive and therefore should be privileged.

2) Non-equivalent mutant detection - Answering RQ2: To answer RQ2, we compute the non-equivalent mutant classification recall of the BS/RS algorithms (in Figure 2), *i.e.*, the percentage of non-equivalent mutants detected by the BS/RS amongst the selected mutants. By construction, the ALE algorithm has a recall of 100%, it is therefore not shown here. It is also noted that the precision is 100% since all the non-equivalent mutants detected are indeed killable, by construction of our mutant set.

All our simulations obtain a recall higher than 85%, with a clear advantage for biased simulations which never achieve worse than 95% for the weak mutation scenario. As for time, deviation in the recall is smaller for biased simulations thus

making the approach more predictable in addition of being more reliable. We also observe that the random simulations are more sensitive to the number of runs: we need more of them to discover discrepancies by luck. This effect cannot be observed for biased simulations. A possible explanation is that the number of runs required to cover infected states with traces is lower than the number we provided.

For strong mutation, the BS approach's recall decreases to around 92% ($\overline{recall} = 92\%$, $\sigma = 3\%$): amongst the 5113 non-equivalent mutant non-detections (over a total of 64529 non-equivalent mutant evaluations), 1905 (37%) were TAD mutants, 1755 (34%) were WIS mutants, 545 (11%) were TDE mutants, and 459 (9%) were 2nd-order TAD mutants (*i.e.*, TAD-TAD mutants); the rest of non-equivalent mutants not detected is distributed amongst different operators with less than 2% for each. This decrease may be due to the difficulty to find a path to the initial state: for strong mutation, the BS trace selection algorithm will consider traces starting from, and ending in, the initial state. This means that mutations creating (TAD) or modifying (TDE) a back-level transition will not be detected using SM BS. Concerning WIS mutants, we believe that, as the WIS operator only changes the initial state of the TS, the set of infected states (S_{infect}) is empty, which is equivalent in our implementation of SM BS to considering all the states infected.

3) Worst case scenario (execution time) - Answering RQ3: Figure 3 presents a compact view of the worst execution time of the different algorithms (RQ3). We grouped the different results by the kind of model: embedded system, web-application, or randomly generated model. As expected, the RS/BS execution time is directly correlated to the δ and ϵ values: a lower number of traces selected and executed (N) takes less time. Overall, the time of the ALE executions grows with the size of the model, reaching 5660 seconds (more than one and a half hour) for the worst WM ALE execution time on the Random 2 model.

C. Threats to Validity

1) Internal Validity: We performed our experiment on 12 models: 3 academic examples (*S. V. Mach.*, *C. P. Term.*, *Minepump*), 5 larger real-world models (*Claroline*, *Elsa-RR*, *Elsa-RRN*, *AGE-RR*, and *AGE-RRN*) and 4 randomly generated models (*Random 1-4*). These models come from different sources and represent two different kinds of systems: embedded systems designed by an engineer and web-based applications where the model has been reverse-engineered from a running instance using a 2-gram inference method [30]. The random models were built from a set of generated TSs in order to match the real system state-space measures, as described by Pelánek [31], [32].

2) Construct Validity: The RS/BS δ and ϵ values have been arbitrarily chosen. The first values (RS1/BS1: $\delta = 1e{-}10$, $\epsilon = 0.01$) are the same as in Hérault *et al.* [28]. As the number of traces selected and executed N equals to $\frac{8\log(2/\delta)}{\epsilon^2}$, we chose to run the algorithm with 3 higher parameters values in order to reduce N. We cannot guarantee that our parameter values are relevant for any model. They will rather depend on the model size, the desired approximation (ϵ) and confidence (δ), and the time budget allowed for the equivalence analysis.

To the best of our knowledge, the HKC library [34] was the only publicly available tool able to perform ALE checking on non-deterministic TSs. We cannot guarantee that there are no other other tools providing the same features with lower execution time. To avoid bias in the random selections in the RS/BS algorithms, we execute each configuration of the different algorithms 3 times.

3) External Validity: We cannot guarantee that our results are generalizable to all behavioural models. However, we recall the diversity of the model sources (hand-crafted, reverse-engineered, and randomly generated to match real system state-space) as well as the diversity of the considered systems. Variations in performance of the algorithms also suggest mitigation of this threat.

4) Conclusion Validity: To confirm our observations on the recall of the RS/BS algorithms, we test the null hypothesis between the outputs of our algorithm (the mutant is equivalent/non-equivalent) and a random equivalent/non-equivalent assignment using a Wilcoxon rank sum test. The p-value lower than 2.2e-16[4] discredits the null hypothesis showing that the equivalent/non-equivalent detection recall is significant.

To confirm the statistical difference between the execution times of the RS/BS and ALE algorithms, we test the null hypothesis between RS/BS execution time and ALE execution time for weak and strong mutation for each of our input models using a Wilcoxon rank sum test. For weak mutation, the results of this statistical test are shown in Table III: for every model except *AGE-RR*/*AGE-RRN* models, the p-value is lower than 2.2e-16, discrediting the null hypothesis and showing a significant difference in the execution times. The execution times of *AGE-RR*/*AGE-RRN* model are only significant for RS1 to RS3, BS1, and BS3 (for *AGE-RR*); and RS2 to RS4 (for *AGE-RRN*). For strong mutation, all the p-values were lower than 2.2e-16, showing a significant difference in execution time between the BS algorithm and the ALE algorithm in a strong mutation scenario.

5) Verifiability: The input models, as well as the tools and scripts used to perform the empirical assessment, are available online at https://projects.info.unamur.be/vibes/mutants-equiv.html. The input models are encoded using an XML format and are processed by our Java tools (part of VIBeS [35]) for the RS/BS algorithms. The ALE execution is done using HKC [34]. Both VIBeS and HKC are released under open source licences (MIT license for VIBeS, GNU LGPL for HKC), allowing one to inspect, reuse, or adapt the code. VIBeS's source code is available online in the Git repository at https://forge.info.unamur.be/scm/git/vibes, and the different Maven artefacts were deployed on the Maven central repository. As our assessment involves randomization, the complete

[4]Value $2.2e-16$ corresponds to the smallest possible p-value computable with R.

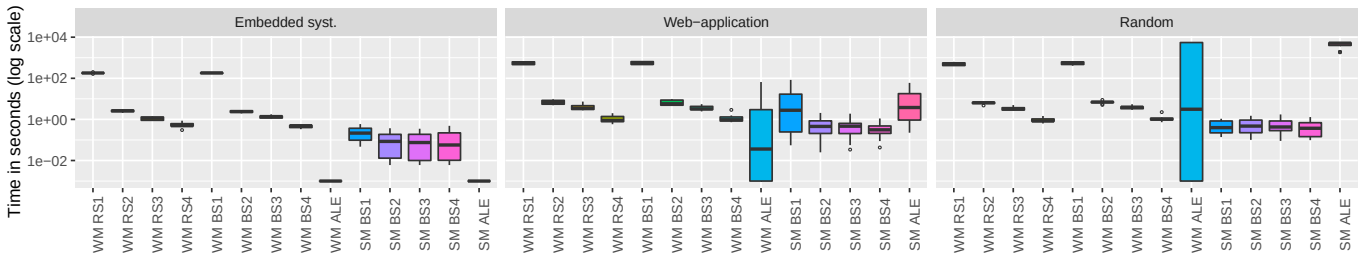

Fig. 3. Worst execution time of the equivalent mutant detection using the model itself as mutant

TABLE III
P-VALUES OF THE WILCOXON RANK SUM TEST BETWEEN THE WM RS/BS EXECUTION TIMES AND THE WM ALE EXECUTION TIMES.

Model	WM RS1	WM RS2	WM RS3	WM RS4	WM BS1	WM BS2	WM BS3	WM BS4
S.V.Mach.	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
C.P.Term.	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
Minepump	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
Claroline	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
Elsa-RR	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
Elsa-RRN	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
AGE-RR	$2.866e-03$	$9.676e-03$	$2.021e-02$	$\mathbf{3.249e-01}$	$9.107e-03$	$4.744e-02$	$\mathbf{6.405e-02}$	$\mathbf{1.382e-01}$
AGE-RRN	$\mathbf{8.143e-02}$	$8.379e-04$	$6.981e-04$	$2.162e-02$	$\mathbf{5.991e-01}$	$\mathbf{7.076e-01}$	$\mathbf{5.674e-01}$	$\mathbf{5.168e-01}$
Random 1	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
Random 2	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
Random 3	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
Random 4	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$

results are also downloadable as well as the script files used to perform the analysis described in section IV-B. Finally, one may (re-)run the complete assessment using the provided Makefile.

D. Lessons Learned

From our experiment we draw the following lessons. (i) Regarding weak mutation and independently of the size or nature of the models, the ALE approach provides faster and exact answers. This indicates that state-of-the-art language equivalence algorithms can be used successfully for such a task. (ii) Regarding strong mutation, biased random simulations are of interest for the web and the random models, and gains increase with the size (from one to three orders of magnitude). Recalls of 90% and above allow to use such simulations as reasonably reliable fast filters to discard non-equivalent mutants, leaving to ALE algorithms "difficult" cases so as to accelerate the analysis of large mutants bases. (iii) Biased simulations are more predictable in terms of execution time and recall. Additionally, drastically increasing the number of runs does not affect their performance as opposed to random simulations. (iv) The configuration of the ALE algorithm (forward/backward processing, or breadth-first or depth-first exploration) has very little influence on the total execution time (regarding equivalent mutant detection). This may be explained by the fact that mutations occur randomly and therefore do not privilege any graph traversal strategy.

V. RELATED WORK

The usage of simulation heuristics for testing purposes is presented in Section V-A). Approaches related to the equivalent mutant problem and model-based mutation are then discussed in Sections V-B and V-C, respectively.

A. Simulation

Our random simulation heuristic, which yields a probabilistic interpretation of the problems under analysis by making several repeated samples, is akin to Monte-Carlo simulation. Monte-Carlo methods were found to be quite efficient for searching and reasoning on large data spaces. In software verification, Monte Carlo simulations have been used to devise statistical model-checking techniques [28], [36] that alleviate state explosion. In software testing, Langdon *et al.* [37] used them, together with genetic programming, in order to identify subsuming higher-order mutants. Poulding and Feldt [38] used a variant of the method, called Nested Monte-Carlo Search, to generate random data structures to be used for testing. Along the same lines, Nested Monte-Carlo Search was used, by Poulding and Feldt [39] to heuristically perform model checking of Java programs. All these methods are related to ours since they use Monte-Carlo. However, none of them aims at modelling mutants or tackling the equivalent mutant problem.

Walkinshaw and Bogdanov [40] advocate that using random selection (like Lo and Khoo [41]) in order to compare automata languages may be biased due to the impossibility to obtain a representative sample of the language. In their work, they use a model-based testing approach (the W-method [42]) to compare two automata from the accepted language perspective, and a *diff* algorithm to compare them with respect to their transition structures (which is a more elaborate version of our heuristic used to compute the set of infected states S_{infect}). In contrast, we look for difference instead of similarity, which motivates the choice of easier-to-compute random heuristics as baselines to compare with an ALE approach.

B. Equivalent mutants

Previous work demonstrated that equivalent mutants skew the mutation score measurements and thus hinder the effectiveness of the method [43]. Unfortunately, it has been proven that judging whether a code mutant is equivalent to the original code is an undecidable problem [44]. This means that there is no solution to the general case of this problem. Luckily, since mutations are small syntactic changes, heuristics can identify several classes of them [13]. Two types of such heuristics exist in the literature: those that operate in a *static* manner and those that are *dynamic*.

Static techniques include the use of compiler optimizations [45], constraint solving [24], program slicing [46], data-flow patterns [47], and formal verification [25]. All these techniques are effective at detecting certain types of equivalent mutants, *i.e.*, trivial equivalencies [13], but unfortunately, they are not applicable to model mutants.

Dynamic techniques measure the differences between the test executions of the original and mutant programs and identify likely non-equivalent mutants. Schuler and Zeller [48] and Papadakis *et al.* [49] measure the impact on coverage, while Kintis *et al.* [50] measure the impact on other mutants (second-order mutants). Our technique shares the same notion of equivalence because we check the model trace in order to judge it. However, we do not consider executable code as we only deal with model mutants. We also sample execution in order to increase the efficiency of the process. It is to be noted that we have a different notion of equivalence since we deal with behavioural models. Therefore, differences in traces imply different behaviours, which is not the case for executable code.

Non-determinism complicates equivalence detection both at the code [51] and model levels [52]. Patel and Hierons [51] associate predictions from pairs of inputs and outputs of the mutant program and check whether these predictions can be discarded by the original program, hence showing non-equivalence. This is not applicable to our case since our models do not have outputs. Aichernig and Jöbstl [52] also encode the semantics of the action models in terms of constraints and use refinement to check conformance in the context of non-determinism. In our case, RS/BS manage non determinism in the TSs by considering all the possible runs.

Perhaps the closest work is that of Papadakis and Malevris [53] who sample execution paths according to their length (select the k-shortest paths), symbolically execute them and judge mutant equivalence based on the selected paths. The main differences with our approach are that we additionally sample paths that cover infected states and we operate on behavioural models instead of actual code representation.

C. Model-based mutation

Specification mutation testing aims at identifying defects on the implementations under test by altering the models of the system and requiring the design of tests that identify these differences [9]. The main point about this technique is that it complements code-based testing by targeting problems related to missing functionality [10], [11].

Given the plethora of the existing models and languages, many model-based mutation techniques have been developed. Woodward [54], Fabbri *et al.* [55] and Hierons and Merayo [56] suggested a set of mutant operators for algebraic specifications, finite state machines and Statecharts, and probabilistic finite state machines, respectively. Similarly, Henard *et al.* [57], Arcaini *et al.* [58] and Papadakis *et al.* [8] mutated feature models and combinatorial interaction models.

Regarding behavioural models, like the ones we used here, Aichernig *et al.* [20], [59] developed a mutation-based test generation technique for state machines. Belli and Beyazit [60] compare mutation-testing strategies when applied on event-based and state-based models, and found that both had similar effectiveness. In follow-up studies, Belli *et al.* [61] and Aichernig *et al.* [12] evaluated their model-based mutation testing approaches on industrial systems and found that they were complementary, in terms of fault detection, to code-based testing.

Generally, the EMP is seldom the single focus of the above approaches as it is in the present study.

VI. Conclusion

In this paper, we investigated the relevance of an exact language equivalence approach to tackle the equivalent mutant problem at the model level. To do so, we offered two baseline algorithms based on random simulation, and compared them to language equivalence under weak and strong mutation scenarios. Our experiments demonstrated the efficiency of the exact approach for the weak mutation scenario. For strong mutation, our biased simulations – that pre-process the models to detect states that are infected by mutations – are efficient (up to 1,000 times faster) on models that contain more than 300 states, limiting detection errors to 8%. These results suggest using simulations first to quickly discard many non-equivalent mutants, and then employing exact approaches only on a small amount of "probably" equivalent mutants to speed up equivalence analysis.

There is room for improvement. First, we will extend our experiments to other forms of equivalence and tools. We would also like to switch from the pure equivalence analysis to test generation concerns by analysing counter-examples. Our long-term goal is to draw attention on the applications of language equivalence for mutation testing and develop further EMP-dedicated solutions.

Appendix

This appendix presents the results of the different weak and strong mutations ALEs/BSs/RSs algorithms. For each algorithm, a table gives the recall, the average execution time ($\overline{\mathbf{time}}$), and the standard deviation (σ).

S.V.Mach.

			Weak Mutation			Strong Mutation		
	δ	ϵ	Recall	**time**	σ	Recall	**time**	σ
ALE			100%	<0.01	<0.01	100%	<0.01	<0.01
BS	1e-10	0.01	98%	0.02	0.03	91%	0.26	1.00
	1e-10	0.10	97%	0.02	0.02	91%	0.04	0.06
	1e-05	0.10	97%	<0.01	0.02	91%	0.03	0.05
	0.10	0.10	98%	0.01	0.02	91%	0.02	0.04
RS	1e-10	0.01	97%	0.02	0.03	N/A	N/A	N/A
	1e-10	0.10	96%	0.01	0.02	N/A	N/A	N/A
	1e-05	0.10	97%	<0.01	0.01	N/A	N/A	N/A
	0.10	0.10	97%	0.01	0.03	N/A	N/A	N/A

C.P.Term.

			Weak Mutation			Strong Mutation		
	δ	ϵ	Recall	**time**	σ	Recall	**time**	σ
ALE			100%	<0.01	<0.01	100%	<0.01	<0.01
BS	1e-10	0.01	97%	0.49	9.05	91%	0.21	0.76
	1e-10	0.10	96%	0.02	0.10	91%	0.04	0.05
	1e-05	0.10	97%	0.01	0.05	91%	0.03	0.05
	0.10	0.10	96%	0.01	0.03	91%	0.03	0.04
RS	1e-10	0.01	97%	0.49	9.04	N/A	N/A	N/A
	1e-10	0.10	96%	0.02	0.11	N/A	N/A	N/A
	1e-05	0.10	97%	<0.01	0.05	N/A	N/A	N/A
	0.10	0.10	96%	0.01	0.04	N/A	N/A	N/A

Minepump

			Weak Mutation			Strong Mutation		
	δ	ϵ	Recall	**time**	σ	Recall	**time**	σ
ALE			100%	<0.01	<0.01	100%	<0.01	<0.01
BS	1e-10	0.01	98%	0.40	8.54	92%	0.21	0.80
	1e-10	0.10	98%	0.02	0.15	92%	0.04	0.06
	1e-05	0.10	99%	<0.01	0.04	92%	0.03	0.05
	0.10	0.10	98%	0.01	0.04	92%	0.03	0.04
RS	1e-10	0.01	98%	0.39	8.43	N/A	N/A	N/A
	1e-10	0.10	98%	0.02	0.15	N/A	N/A	N/A
	1e-05	0.10	98%	<0.01	0.06	N/A	N/A	N/A
	0.10	0.10	98%	0.01	0.05	N/A	N/A	N/A

Claroline

			Weak Mutation			Strong Mutation		
	δ	ϵ	Recall	**time**	σ	Recall	**time**	σ
ALE			100%	0.02	0.02	100%	0.10	0.12
BS	1e-10	0.01	99%	3.62	49.96	98%	0.59	2.00
	1e-10	0.10	99%	0.09	0.57	98%	0.17	0.42
	1e-05	0.10	99%	0.07	0.32	98%	0.17	0.28
	0.10	0.10	99%	0.05	0.12	98%	0.18	0.71
RS	1e-10	0.01	96%	29.99	139.34	N/A	N/A	N/A
	1e-10	0.10	95%	0.39	1.52	N/A	N/A	N/A
	1e-05	0.10	94%	0.23	0.80	N/A	N/A	N/A
	0.10	0.10	94%	0.10	0.25	N/A	N/A	N/A

Elsa-RR

			Weak Mutation			Strong Mutation		
	δ	ϵ	Recall	**time**	σ	Recall	**time**	σ
ALE			100%	<0.01	<0.01	100%	1.05	0.67
BS	1e-10	0.01	99%	0.06	0.05	95%	0.96	3.86
	1e-10	0.10	100%	0.04	0.04	95%	0.15	0.27
	1e-05	0.10	99%	0.05	0.04	95%	0.13	0.19
	0.10	0.10	100%	0.02	0.03	95%	0.09	0.16
RS	1e-10	0.01	88%	73.03	209.50	N/A	N/A	N/A
	1e-10	0.10	86%	0.92	2.56	N/A	N/A	N/A
	1e-05	0.10	86%	0.51	1.38	N/A	N/A	N/A
	0.10	0.10	87%	0.13	0.33	N/A	N/A	N/A

Elsa-RRN

			Weak Mutation			Strong Mutation		
	δ	ϵ	Recall	**time**	σ	Recall	**time**	σ
ALE			100%	0.01	0.01	100%	3.64	2.29
BS	1e-10	0.01	100%	0.05	0.05	90%	2.93	10.34
	1e-10	0.10	100%	0.04	0.04	90%	0.18	0.25
	1e-05	0.10	99%	0.04	0.04	90%	0.16	0.21
	0.10	0.10	100%	0.03	0.03	90%	0.10	0.11
RS	1e-10	0.01	97%	19.24	100.73	N/A	N/A	N/A
	1e-10	0.10	95%	0.37	1.42	N/A	N/A	N/A
	1e-05	0.10	95%	0.22	0.75	N/A	N/A	N/A
	0.10	0.10	94%	0.08	0.21	N/A	N/A	N/A

AGE-RR

			Weak Mutation			Strong Mutation		
	δ	ϵ	Recall	**time**	σ	Recall	**time**	σ
ALE			100%	0.64	0.94	100%	21.18	13.70
BS	1e-10	0.01	100%	0.06	0.08	90%	9.38	42.87
	1e-10	0.10	100%	0.05	0.10	90%	0.24	0.45
	1e-05	0.10	100%	0.04	0.08	90%	0.18	0.47
	0.10	0.10	100%	0.03	0.04	89%	0.09	0.25
RS	1e-10	0.01	96%	38.68	188.18	N/A	N/A	N/A
	1e-10	0.10	94%	0.68	2.50	N/A	N/A	N/A
	1e-05	0.10	95%	0.35	1.27	N/A	N/A	N/A
	0.10	0.10	94%	0.14	0.47	N/A	N/A	N/A

AGE-RRN

			Weak Mutation			Strong Mutation		
	δ	ϵ	Recall	**time**	σ	Recall	**time**	σ
ALE			100%	0.21	0.22	100%	75.29	51.92
BS	1e-10	0.01	100%	0.04	0.07	95%	7.10	32.32
	1e-10	0.10	100%	0.05	0.05	95%	0.32	0.46
	1e-05	0.10	100%	0.04	0.04	95%	0.27	0.43
	0.10	0.10	100%	0.04	0.04	95%	0.21	0.31
RS	1e-10	0.01	90%	117.21	362.41	N/A	N/A	N/A
	1e-10	0.10	88%	1.98	4.78	N/A	N/A	N/A
	1e-05	0.10	87%	1.12	2.63	N/A	N/A	N/A
	0.10	0.10	85%	0.41	0.87	N/A	N/A	N/A

Random 1

			Weak Mutation			Strong Mutation		
	δ	ϵ	Recall	**time**	σ	Recall	**time**	σ
ALE			100%	<0.01	<0.01	100%	448.61	339.19
BS	1e-10	0.01	100%	0.08	0.07	92%	0.78	2.50
	1e-10	0.10	100%	0.07	0.07	92%	0.09	0.08
	1e-05	0.10	100%	0.07	0.07	92%	0.09	0.06
	0.10	0.10	99%	0.07	0.07	92%	0.07	0.05
RS	1e-10	0.01	100%	0.03	0.07	N/A	N/A	N/A
	1e-10	0.10	100%	0.03	0.11	N/A	N/A	N/A
	1e-05	0.10	100%	0.03	0.09	N/A	N/A	N/A
	0.10	0.10	99%	0.03	0.08	N/A	N/A	N/A

Random 2

			Weak Mutation			Strong Mutation		
	δ	ϵ	Recall	**time**	σ	Recall	**time**	σ
ALE			100%	<0.01	<0.01	100%	412.37	168.90
BS	1e-10	0.01	100%	0.11	0.06	89%	1.22	3.35
	1e-10	0.10	100%	0.10	0.06	89%	0.14	0.09
	1e-05	0.10	100%	0.11	0.07	89%	0.14	0.08
	0.10	0.10	100%	0.11	0.06	89%	0.12	0.07
RS	1e-10	0.01	100%	0.04	0.10	N/A	N/A	N/A
	1e-10	0.10	100%	0.03	0.07	N/A	N/A	N/A
	1e-05	0.10	100%	0.03	0.07	N/A	N/A	N/A
	0.10	0.10	99%	0.03	0.09	N/A	N/A	N/A

Random 3

			Weak Mutation			Strong Mutation		
	δ	ϵ	Recall	**time**	σ	Recall	**time**	σ
ALE			100%	<0.01	<0.01	100%	367.99	154.80
BS	1e-10	0.01	100%	0.11	0.06	91%	1.04	3.20
	1e-10	0.10	100%	0.09	0.04	91%	0.23	0.15
	1e-05	0.10	100%	0.09	0.04	91%	0.23	0.14
	0.10	0.10	100%	0.09	0.05	91%	0.19	0.12
RS	1e-10	0.01	100%	0.03	0.10	N/A	N/A	N/A
	1e-10	0.10	100%	0.03	0.16	N/A	N/A	N/A
	1e-05	0.10	99%	0.03	0.12	N/A	N/A	N/A
	0.10	0.10	99%	0.02	0.07	N/A	N/A	N/A

Random 4

			Weak Mutation			Strong Mutation		
	δ	ϵ	Recall	**time**	σ	Recall	**time**	σ
ALE			100%	<0.01	<0.01	100%	306.37	127.02
BS	1e-10	0.01	100%	0.11	0.06	91%	1.09	3.23
	1e-10	0.10	100%	0.10	0.05	91%	0.22	0.14
	1e-05	0.10	100%	0.10	0.05	91%	0.23	0.12
	0.10	0.10	100%	0.09	0.04	91%	0.19	0.11
RS	1e-10	0.01	100%	0.04	0.25	N/A	N/A	N/A
	1e-10	0.10	99%	0.03	0.25	N/A	N/A	N/A
	1e-05	0.10	100%	0.03	0.10	N/A	N/A	N/A
	0.10	0.10	99%	0.02	0.09	N/A	N/A	N/A

REFERENCES

[1] J. H. Andrews, L. C. Briand, Y. Labiche, and A. S. Namin, "Using Mutation Analysis for Assessing and Comparing Testing Coverage Criteria," *IEEE Transactions on Software Engineering*, vol. 32, no. 8, pp. 608–624, 2006.

[2] J. Offutt, "A mutation carol: Past, present and future," *Information and Software Technology*, vol. 53, no. 10, pp. 1098–1107, Oct. 2011.

[3] M. Gligoric, A. Groce, C. Zhang, R. Sharma, M. A. Alipour, and D. Marinov, "Comparing non-adequate test suites using coverage criteria," in *International Symposium on Software Testing and Analysis, ISSTA*. Lugano, Switzerland,: ACM, July 15-20 2013, pp. 302–313.

[4] M. Papadakis and N. Malevris, "Automatic mutation test case generation via dynamic symbolic execution," in *International Symposium on Software Reliability Engineering, ISSRE*. IEEE, 2010, pp. 121–130.

[5] G. Fraser and A. Arcuri, "Achieving scalable mutation-based generation of whole test suites," *Empirical Software Engineering*, pp. 1–30, 2014.

[6] R. Just, D. Jalali, L. Inozemtseva, M. D. Ernst, R. Holmes, and G. Fraser, "Are Mutants a Valid Substitute for Real Faults in Software Testing?" in *International Symposium on the Foundations of Software Engineering, FSE*. ACM, 2014, pp. 654–665.

[7] R. Baker and I. Habli, "An empirical evaluation of mutation testing for improving the test quality of safety-critical software," *IEEE Transactions on Software Engineering*, vol. 39, no. 6, pp. 787–805, 2013.

[8] M. Papadakis, C. Henard, and Y. Le Traon, "Sampling program inputs with mutation analysis: Going beyond combinatorial interaction testing," in *International Conference on Software Testing, Verification and Validation, ICST*. IEEE, 2014, pp. 1–10.

[9] T. A. Budd and A. S. Gopal, "Program testing by specification mutation," *Computer Languages*, vol. 10, no. 1, pp. 63–73, Jan. 1985.

[10] W. E. Howden, "Reliability of the path analysis testing strategy." *IEEE Transactions on Software Engineering*, vol. 2, no. 3, pp. 208–215, 1976.

[11] J. M. Voas and G. McGraw, *Software Fault Injection: Inoculating Programs Against Errors*. John Wiley & Sons, Inc., 1997.

[12] B. K. Aichernig, J. Auer, E. Jöbstl, R. Korosec, W. Krenn, R. Schlick, and B. V. Schmidt, "Model-based mutation testing of an industrial measurement device," in *Tests and Proofs*, ser. LNCS, vol. 8570. Springer, 2014, pp. 1–19.

[13] M. Papadakis, Y. Jia, M. Harman, and Y. Le Traon, "Trivial compiler equivalence: A large scale empirical study of a simple fast and effective equivalent mutant detection technique," in *International Conference on Software Engineering, ICSE*. IEEE, 2015, pp. 936–946.

[14] Y. Jia and M. Harman, "An Analysis and Survey of the Development of Mutation Testing," *IEEE Transactions on Software Engineering*, vol. 37, no. 5, pp. 649–678, Sep. 2011.

[15] O. Kupferman and M. Y. Vardi, "Verification of fair transition systems," in *Computer Aided Verification*. Springer, 1996, pp. 372–382.

[16] F. Bonchi and D. Pous, "Checking NFA equivalence with bisimulations up to congruence," in *Symposium on Principles of Programming Languages, POPL*. Rome, Italy: ACM, 2013, pp. 457–468. [Online]. Available: http://doi.acm.org/10.1145/2429069.2429124

[17] L. Doyen and J. Raskin, "Antichain algorithms for finite automata," in *Tools and Algorithms for the Construction and Analysis of Systems, TACAS*, ser. Lecture Notes in Computer Science, vol. 6015. Springer, 2010, pp. 2–22. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12002-2_2

[18] R. Just, M. D. Ernst, and G. Fraser, "Efficient mutation analysis by propagating and partitioning infected execution states," in *ISSTA*. ACM, 2014, pp. 315–326.

[19] C. Baier and J. Katoen, *Principles of model checking*. MIT Press, 2008.

[20] B. K. Aichernig, E. Jöbstl, and S. Tiran, "Model-based mutation testing via symbolic refinement checking," *Science of Computer Programming*, vol. 97, pp. 383–404, Jan. 2015.

[21] E. Jöbstl, "Model-based mutation testing with constraint and smt solvers," Ph.D. dissertation, Graz University of Technology, 2014.

[22] R. A. DeMillo and A. J. Offutt, "Experimental results from an automatic test case generator," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 2, no. 2, pp. 109–127, 1993.

[23] W. E. Howden, "Weak mutation testing and completeness of test sets," *IEEE Transactions on Software Engineering*, vol. SE-8, no. 4, pp. 371–379, July 1982.

[24] A. J. Offutt and J. Pan, "Automatically detecting equivalent mutants and infeasible paths," *Software Testing, Verification and Reliability*, vol. 7, no. 3, pp. 165–192, 1997.

[25] S. Bardin, M. Delahaye, R. David, N. Kosmatov, M. Papadakis, Y. Le Traon, and J. Marion, "Sound and quasi-complete detection of infeasible test requirements," in *International Conference on Software Testing, Verification and Validation, ICST*. Graz, Austria: IEEE, 2015, pp. 1–10. [Online]. Available: http://dx.doi.org/10.1109/ICST.2015.7102607

[26] W. Krenn and R. Schlick, "Mutation-driven test case generation using short-lived concurrent mutants - first results," *CoRR*, vol. abs/1601.06974, 2016.

[27] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press Cambridge, 2001, vol. 6.

[28] T. Hérault, R. Lassaigne, F. Magniette, and S. Peyronnet, "Approximate probabilistic model checking," in *Verification, Model Checking, and Abstract Interpretation, 5th International Conference, VMCAI*, ser. Lecture Notes in Computer Science, vol. 2937. Venice, Italy: Springer, 2004, pp. 73–84.

[29] A. Classen, "Modelling with FTS: a Collection of Illustrative Examples," PReCISE Research Center, University of Namur, Namur, Belgium, Tech. Rep. P-CS-TR SPLMC-00000001, 2010. [Online]. Available: https://projects.info.unamur.be/fts/publications/

[30] S. E. Sprenkle, L. L. Pollock, and L. M. Simko, "Configuring effective navigation models and abstract test cases for web applications by analysing user behaviour," *Software Testing, Verification and Reliability*, vol. 23, no. 6, pp. 439–464, 2013.

[31] R. Pelánek, "Typical Structural Properties of State Spaces," in *International SPIN Workshop*, ser. LNCS, vol. 2989. Springer, 2004, pp. 5–22.

[32] ——, "Properties of state spaces and their applications," *International Journal on Software Tools for Technology Transfer*, vol. 10, no. 5, pp. 443–454, 2008.

[33] X. Devroey, G. Perrouin, M. Papadakis, P.-Y. Schobbens, and P. Heymans, "Featured Model-based Mutation Analysis," in *International Conference on Software Engineering, ICSE*. Austin, TX, USA: ACM, 2016.

[34] F. Bonchi and D. Pous, "HKC Library v. 1.0," https://perso.ens-lyon.fr/damien.pous/hknt/, 2013.

[35] X. Devroey and G. Perrouin, "Variability Intensive system Behavioural teSting (VIBeS) v. 1.1.4," https://projects.info.unamur.be/vibes/, Namur, Belgium, 2015.

[36] H. L. S. Younes and R. G. Simmons, "Probabilistic verification of discrete event systems using acceptance sampling," in *International Conference on Computer Aided Verification, CAV*, ser. Lecture Notes in Computer Science, vol. 2404. London, UK, UK: Springer-Verlag, 2002, pp. 223–235. [Online]. Available: http://dl.acm.org/citation.cfm?id=647771.760735

[37] W. B. Langdon, M. Harman, and Y. Jia, "Efficient multi-objective higher order mutation testing with genetic programming," *Journal of Systems and Software*, vol. 83, no. 12, pp. 2416–2430, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.jss.2010.07.027

[38] S. M. Poulding and R. Feldt, "Generating structured test data with specific properties using nested monte-carlo search," in *Genetic and Evolutionary Computation Conference, GECCO*. Vancouver, BC, Canada: ACM, 2014, pp. 1279–1286. [Online]. Available: http://doi.acm.org/10.1145/2576768.2598339

[39] ——, "Heuristic model checking using a monte-carlo tree search algorithm," in *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO*. Madrid, Spain: ACM, 2015, pp. 1359–1366. [Online]. Available: http://doi.acm.org/10.1145/2739480.2754767

[40] N. Walkinshaw and K. Bogdanov, "Automated Comparison of State-Based Software Models in Terms of Their Language and Structure," *ACM Transactions on Software Engineering and Methodology*, vol. 22, no. 2, pp. 1–37, mar 2013.

[41] D. Lo and S. c. Khoo, "Quark: Empirical assessment of automaton-based specification miners," in *13th Working Conference on Reverse Engineering*, Oct 2006, pp. 51–60.

[42] A. P. Mathur, *Foundations of software testing*. Pearson Education, 2008.

[43] L. Madeyski, W. Orzeszyna, R. Torkar, and M. Jozala, "Overcoming the equivalent mutant problem: A systematic literature review and a comparative experiment of second order mutation," *IEEE Transactions on Software Engineering*, vol. 40, no. 1, pp. 23–42, 2014.

[44] T. A. Budd and D. Angluin, "Two Notions of Correctness and Their Relation to Testing," *Acta Informatica*, vol. 18, no. 1, pp. 31–45, March 1982.

[45] A. J. Offutt and W. M. Craft, "Using compiler optimization techniques to

detect equivalent mutants," *Software Testing, Verification and Reliability*, vol. 4, no. 3, pp. 131–154, 1994.

[46] R. M. Hierons, M. Harman, and S. Danicic, "Using program slicing to assist in the detection of equivalent mutants," *Software Testing, Verification and Reliability*, vol. 9, no. 4, pp. 233–262, 1999.

[47] M. Kintis and N. Malevris, "MEDIC: A static analysis framework for equivalent mutant identification," *Information & Software Technology*, vol. 68, pp. 1–17, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2015.07.009

[48] D. Schuler and A. Zeller, "Covering and uncovering equivalent mutants," *Software Testing, Verification and Reliability*, vol. 23, no. 5, pp. 353–374, 2013.

[49] M. Papadakis, M. E. Delamaro, and Y. Le Traon, "Mitigating the effects of equivalent mutants with mutant classification strategies," *Science of Computer Programming*, vol. 95, pp. 298–319, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.scico.2014.05.012

[50] M. Kintis, M. Papadakis, and N. Malevris, "Employing second-order mutation for isolating first-order equivalent mutants," *Software Testing, Verification and Reliability*, vol. 25, no. 5-7, pp. 508–535, 2015. [Online]. Available: http://dx.doi.org/10.1002/stvr.1529

[51] K. Patel and R. M. Hierons, "Resolving the equivalent mutant problem in the presence of non-determinism and conicidental correctness," in *28th IFIP International Conference on Testing Software and Systems*, 2016.

[52] B. K. Aichernig and E. Jobstl, "Towards symbolic model-based mutation testing: Pitfalls in expressing semantics as constraints," in *2012 IEEE Fifth International Conference on Software Testing, Verification and Validation*, April 2012, pp. 752–757.

[53] M. Papadakis and N. Malevris, "Mutation based test case generation via a path selection strategy," *Information & Software Technology*, vol. 54, no. 9, pp. 915–932, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2012.02.004

[54] M. R. Woodward, "Errors in algebraic specifications and an experimental mutation testing tool," *Software Engineering Journal*, vol. 8, no. 4, pp. 221–224, July 1993.

[55] S. Fabbri, J. C. Maldonado, T. Sugeta, and P. C. Masiero, "Mutation testing applied to validate specifications based on statecharts," in *International Symposium on Software Reliability Engineering, ISSRE*. IEEE, 1999, pp. 210–219.

[56] R. M. Hierons and M. G. Merayo, "Mutation testing from probabilistic and stochastic finite state machines," *Journal of Systems and Software*, vol. 82, no. 11, pp. 1804–1818, 2009.

[57] C. Henard, M. Papadakis, G. Perrouin, J. Klein, and Y. Le Traon, "Assessing Software Product Line Testing Via Model-Based Mutation: An Application to Similarity Testing," in *International Conference on Software Testing, Verification and Validation Workshops, ICSTW*. Luxembourg, Luxembourg: IEEE, 2013, pp. 188–197.

[58] P. Arcaini, A. Gargantini, and P. Vavassori, "Generating tests for detecting faults in feature models," in *International Conference on Software Testing, Verification and Validation, ICST*. Graz, Austria: IEEE, April 2015, pp. 1–10. [Online]. Available: http://dx.doi.org/10.1109/ICST.2015.7102591

[59] W. Krenn, R. Schlick, S. Tiran, B. K. Aichernig, E. Jöbstl, and H. Brandl, "Momut: : UML model-based mutation testing for UML," in *International Conference on Software Testing, Verification and Validation, ICST*. Graz, Austria: IEEE, April 13-17 2015, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/ICST.2015.7102627

[60] F. Belli and M. Beyazit, "Event-Based Mutation Testing vs. State-Based Mutation Testing - An Experimental Comparison," in *International Conference on Computers, Software & Applications, COMPSAC*. IEEE, Jul. 2011, pp. 650–655.

[61] F. Belli, C. J. Budnik, A. Hollmann, T. Tuglular, and W. E. Wong, "Model-based mutation testing - approach and case studies," *Science of Computer Programming*, vol. 120, pp. 25–48, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.scico.2016.01.003