

HaltSim: Uma Ferramenta de Avaliação para Critérios de Parada em Simulações Estocásticas

Artur Ligieri Nunes¹, Joao P. A. Dantas² e Luiz Alberto Vieira Dias¹

¹Instituto Tecnológico de Aeronáutica, São José dos Campos, São Paulo, 12288-900, Brasil

²Instituto de Estudos Avançados, São José dos Campos, São Paulo, 12228-001, Brasil

Resumo—Determinar quando interromper uma campanha de simulação estocástica exige equilibrar precisão estatística e custo computacional. Este artigo apresenta o *HaltSim*, uma ferramenta modular para avaliar critérios de parada em ambiente controlado antes de sua integração com simuladores mais complexos. A demonstração usa um cenário aeroespacial sintético de rota econômica e compara, em 1000 campanhas por combinação, *Latin Hypercube Sampling* e *Monte Carlo* sob orçamento fixo, dois critérios de intervalo de confiança de largura fixa e uma heurística de estabilização relativa. Contra uma referência empírica de alta amostragem, os critérios de intervalo de confiança reduziram cerca de 17–19% das amostras em relação ao orçamento fixo de 300 amostras, mantendo a meta de precisão no cenário avaliado. A heurística também reduziu o custo amostral, mas não atingiu essa meta. Os resultados indicam que o *HaltSim* permite comparar o desempenho de critérios de parada antes de integrá-los a simuladores mais complexos.

Palavras-Chave—simulação estocástica, critérios de parada, arquitetura de simulação

I. INTRODUÇÃO

Simulações estocásticas incorporam elementos aleatórios por meio de distribuições de probabilidade, permitindo representar sistemas cujo comportamento depende de incertezas inerentes. Por isso, execuções com a mesma configuração nominal podem produzir respostas distintas quando números aleatórios diferentes são usados para amostrar essas distribuições [1]. A análise de uma campanha de simulação requer planejar as execuções e avaliar estatisticamente os dados de saída, de modo que a quantidade de amostras seja suficiente para estimar as métricas de interesse com incerteza documentada [2].

Nesse contexto, regras de parada são usadas para determinar o comprimento de uma campanha de simulação a partir da informação acumulada, frequentemente dentro de procedimentos de intervalo de confiança. Essa decisão é crítica: interromper a campanha cedo demais pode prejudicar precisão e cobertura, enquanto prolongá-la além do necessário aumenta custo computacional sem ganho proporcional de informação [3]. Na análise de saída de simulações estocásticas, a definição do número de repetições ou do comprimento da campanha é parte central do desenho experimental, pois afeta simultaneamente a precisão estatística e o custo computacional [2], [4]. Regras sequenciais, incluindo procedimentos de intervalo de confiança de largura fixa, procuram substituir um orçamento definido a priori por uma decisão baseada na informação acumulada; contudo, a parada dependente dos dados pode alterar a cobertura efetiva dos intervalos, especialmente em

amostras finitas [3], [5], [6]. Por isso, trabalhos posteriores passaram a avaliar não apenas a largura final do intervalo ou o tamanho amostral esperado, mas também perfis de cobertura, distribuição do tempo de parada e critérios integrados de comparação entre procedimentos [7], [8]. Essa literatura sugere que a escolha de uma regra de parada depende do problema, da resposta simulada e do critério de avaliação adotado, motivando estudos empíricos controlados antes da integração em ambientes de simulação mais complexos.

Em ferramentas de simulação complexas, avaliar um critério de parada diretamente no ambiente final pode tornar difícil separar o desempenho estatístico da regra dos custos e incertezas associados à integração do sistema. A preparação de cenários, a configuração de campanhas em lote, o tratamento de sementes, a coleta de métricas e a interpretação de saídas operacionais introduzem fontes adicionais de variabilidade e esforço técnico. Como consequência, uma falha de precisão, cobertura ou estabilidade pode decorrer tanto do próprio critério quanto de escolhas de modelagem, instrumentação, configuração experimental ou processamento das saídas. Por isso, uma etapa intermediária de experimentação controlada permite isolar o critério de parada, documentar seu comportamento e reduzir riscos antes da integração em ambientes de simulação de maior custo e complexidade. No domínio aeroespacial, esse problema é relevante para ecossistemas como o Ambiente de Simulação Aeroespacial (ASA), apresentado como ambiente modular e distribuído [9], e a AsaPy, biblioteca em *Python* para análise de dados de simulações aeroespaciais [10]. Em aplicações como cenários *Beyond Visual Range* (BVR) e combate aéreo simulado, esse ecossistema também se relaciona à avaliação de estratégias baseada em simulação [11].

Este artigo propõe uma etapa intermediária: uma ferramenta de avaliação modular para avaliar critérios de parada em ambiente controlado. Embora a demonstração utilize cenários aeroespaciais sintéticos, a contribuição de engenharia de *software* do *HaltSim* é estruturar um mecanismo de geração de cenários por composição de modelos. Nessa organização, cada cenário explicita suas variáveis de entrada, os modelos que o compõem, como consumo de combustível, vento estocástico e ruído residual neste artigo, e as métricas produzidas. A infraestrutura comum permanece em outra camada da arquitetura, permitindo combinar cenários com diferentes métodos de amostragem, selecionar ou adicionar critérios de parada e estender o módulo de análise com novos fatores, visualizações e relatórios. Assim, o recorte aeroespacial é a aplicação inicial da ferramenta, não o limite conceitual de sua arquitetura.

A contribuição principal deste trabalho é apresentar o *HaltSim* como uma ferramenta de avaliação para investigar

A. L. Nunes, nunesaln@ita.br; J. P. A. Dantas, jpdantas@ita.br; L. A. V. Dias, vdias@ita.br.

critérios de parada em campanhas de simulação estocástica antes de sua integração em ambientes de simulação mais complexos. A partir dessa contribuição, o artigo descreve: (i) uma arquitetura modular que separa cenário, amostragem, execução, critérios de parada e análise; (ii) uma demonstração em cenário aeroespacial sintético de rota econômica; (iii) uma comparação empírica entre orçamento fixo, critérios de intervalo de confiança de largura fixa e heurística de estabilização relativa; e (iv) o comportamento dessa heurística frente à meta externa de precisão. Com isso, a questão experimental é verificar se critérios de parada baseados na incerteza acumulada podem reduzir o número de amostras sem comprometer, nesta ferramenta de avaliação, a precisão definida por semi-largura relativa e a cobertura contra a referência empírica. O objetivo não é otimizar combustível com realismo aeronáutico, mas comparar custo amostral, erro contra referência e efeito do método de amostragem sobre os critérios de parada.

O restante do artigo está organizado da seguinte forma: a Seção II apresenta o referencial teórico; a Seção III descreve a arquitetura da ferramenta; a Seção IV detalha a metodologia experimental; a Seção V discute os resultados empíricos; e a Seção VI apresenta as conclusões.

II. REFERENCIAL TEÓRICO

Esta seção resume os conceitos necessários para a avaliação proposta, incluindo campanhas de simulação estocástica, critérios de parada baseados em intervalos de confiança, amostragem por LHS e Monte Carlo e a heurística de estabilização relativa usada como linha de base operacional.

A. Simulações, Campanhas e Critérios de Parada

De forma geral, uma execução estocástica de um cenário de simulação é uma tomada de amostra: o modelo do cenário é avaliado com uma configuração das variáveis de entrada e com uma semente pseudoaleatória (*seed*) que gera uma sequência específica de valores aleatórios. Essa execução retorna uma amostra de saída do cenário. Como simulações estocásticas podem produzir respostas distintas para uma mesma configuração nominal, a análise normalmente depende de múltiplas execuções e de estimativas estatísticas sobre as saídas acumuladas [1], [2]. Neste artigo, uma campanha é uma rodada composta por várias execuções acumuladas, e o critério de parada decide quando essa sequência de execuções deve ser interrompida. No cenário de rota econômica usado como demonstração, cada execução recebe altitude e velocidade como variáveis de entrada; vento e ruído residual são gerados a partir da semente da amostra; e as saídas principais são consumo de combustível e tempo de voo. Assim, diferentemente de uma grade fixa, a campanha amostral é tratada como uma sequência acumulada de pontos no espaço altitude-velocidade, em que o número de amostras influencia simultaneamente o custo computacional e a qualidade estatística da estimativa [2].

Um critério de parada define a regra usada para interromper essa sequência. O orçamento fixo é reproduzível e simples de auditar, mas não responde à variabilidade observada: ele pode executar amostras desnecessárias quando a estimativa já está estável ou parar antes de precisão suficiente quando a resposta é mais variável. Critérios adaptativos procuram resolver esse problema monitorando evidências acumuladas

durante a campanha e encerrando a execução quando uma condição de precisão ou estabilidade é atendida [3], [12].

Essa adaptação, entretanto, precisa ser avaliada empiricamente. Uma regra que reduz amostras pode alterar a cobertura efetiva do intervalo, aumentar o erro contra uma referência ou produzir decisões sensíveis ao primeiro cruzamento de uma fronteira [5], [6]. Procedimentos de múltiplos cruzamentos mostram que continuar a amostragem após o primeiro atendimento da fronteira pode melhorar o comportamento de cobertura sem exigir sobreamostragem excessiva, embora essa relação dependa do problema e do procedimento adotado [12], [13]. A ferramenta de avaliação proposta neste artigo foi construída para observar esse comportamento em campanhas repetidas antes de qualquer integração com sistemas de simulação mais complexos.

B. Intervalos de Confiança de Largura Fixa

Uma abordagem inicial consiste em monitorar o intervalo de confiança da média e relacionar a continuação da campanha a uma precisão prescrita. Procedimentos sequenciais de *fixed-width confidence interval* (FWCI) formalizam essa decisão ao estimar o tamanho amostral necessário para atingir uma semi-largura alvo [12], [14]. Variações recentes tratam metas absolutas, múltiplos estágios e avaliações por simulação em diferentes contextos [15]–[17]. Seja $\bar{x}_n$ a média amostral do consumo após n amostras, s_n o desvio padrão amostral e h_n a semi-largura do intervalo de confiança. Usa-se um intervalo bilateral de 95% com distribuição t de Student:

$$\bar{x}_n \pm h_n, \quad h_n = t_{1-\alpha/2, n-1} \frac{s_n}{\sqrt{n}}, \quad (1)$$

em que $\alpha = 0,05$ representa o nível de significância associado ao intervalo de 95%. Como a métrica de consumo é positiva no cenário, o critério de largura fixa relativa verifica se:

$$\frac{h_n}{\bar{x}_n} \leq \epsilon_r. \quad (2)$$

Na configuração experimental, adota-se $\epsilon_r = 0,02$ como tolerância relativa de precisão, $n_{\min} = 100$ como número mínimo de amostras antes da primeira checagem e $n_{\max} = 300$ como limite máximo de amostras por campanha. Regras sequenciais interrompidas no primeiro cruzamento de uma fronteira de precisão podem apresentar cobertura efetiva diferente da cobertura nominal [5]. Além disso, a validade e a largura de intervalos em análise de saída de simulações dependem do método de construção do intervalo e da resposta simulada [18]. Por isso, o primeiro cruzamento é tratado aqui como linha de base estatística, não como garantia geral de validade.

C. FWCI Confirmado e Heurística de Estabilização Relativa

A literatura sobre cruzamentos múltiplos sugere continuar a amostragem após cruzamentos iniciais da fronteira de precisão, reduzindo a dependência de uma única decisão local [12], [13]. A variante proposta aqui, chamada FWCI confirmado, adapta essa cautela para execução em lotes. Para o lote ou ponto de verificação j , define-se I_j como indicador de atendimento da precisão e c_j como contador de confirmações consecutivas:

$$I_j = \mathbb{1} \left\{ n_j \geq n_{\min} \text{ e } \frac{h_{n_j}}{\bar{x}_{n_j}} \leq \epsilon_r \right\}, \quad (3)$$

$$c_j = \begin{cases} c_{j-1} + 1, & \text{se } I_j = 1, \\ 0, & \text{se } I_j = 0. \end{cases} \quad (4)$$

A campanha para quando $c_j \geq k_{\text{confirm}}$ ou quando atinge $n_{\max}$. O parâmetro k_{confirm} representa o número de confirmações consecutivas exigidas; usa-se $k_{\text{confirm}} = 2$, valor adotado como salvaguarda operacional preliminar, não como escolha teoricamente ótima.

Também se avalia uma heurística de estabilização relativa, prática inspirada na função `basic_stop` da `AsaPy`, baseada em mudanças relativas entre pontos de verificação. Nessa regra, μ_{atual} e μ_{anterior} representam as médias acumuladas nos pontos de verificação atual e anterior, enquanto σ_{atual} e σ_{anterior} representam os desvios padrão correspondentes:

$$\Delta_\mu = \frac{\mu_{\text{atual}} - \mu_{\text{anterior}}}{\mu_{\text{anterior}}}, \quad \Delta_\sigma = \frac{\sigma_{\text{atual}} - \sigma_{\text{anterior}}}{\sigma_{\text{anterior}}}. \quad (5)$$

As quantidades Δ_μ e Δ_σ medem, respectivamente, a mudança relativa da média e do desvio padrão entre pontos de verificação. A parada ocorre quando $|\Delta_\mu| < \tau$ e $|\Delta_\sigma| < \tau$, em que τ é o limiar de estabilização relativa. No *HaltSim*, usa-se $\tau = 0,02$ sobre a métrica de consumo, sem integração real com `AsaPy`. Esse valor tem a mesma escala percentual de ϵ_r , mas significado diferente: ϵ_r mede semi-largura relativa do intervalo de confiança, enquanto τ mede estabilização relativa entre pontos de verificação. Assim, a heurística não estima a incerteza da média nem controla a largura do intervalo; ela apenas verifica estabilidade local de estatísticas acumuladas. Heurísticas desse tipo são úteis como mecanismos operacionais simples, mas devem ser interpretadas com cautela em análise de saída de simulações [19].

III. ARQUITETURA

A arquitetura do *HaltSim*, ilustrada na Fig. 1, organiza a ferramenta em módulos separados para controle experimental, modelos de cenário, execução adaptativa e análise dos resultados. Essa separação permite distinguir o que pertence à definição do cenário do que pertence à execução da campanha.

Na configuração do cenário, são definidos as variáveis de entrada, as faixas admissíveis, as sementes e os modelos que compõem a resposta simulada. No cenário de rota econômica, essa camada inclui as faixas de altitude e velocidade, o tratamento de vento e ruído residual e as métricas produzidas pela execução.

Na configuração da execução, são definidos o método amostral, o tamanho de lote, o orçamento máximo, os parâmetros dos critérios de parada e os artefatos de saída. A partir dessa configuração, a ferramenta gera um plano amostral, executa amostras em lotes, atualiza estatísticas acumuladas e delega a decisão de continuidade ao critério de parada selecionado. Ao final da campanha, os resultados são comparados a uma referência empírica e registrados em tabelas rastreáveis.

Na implementação, o *HaltSim* foi organizado como um pacote Python modular. O núcleo de modelos reúne funções de atmosfera, aerodinâmica, consumo de rota e componentes estocásticos, enquanto a camada de cenário compõe esses

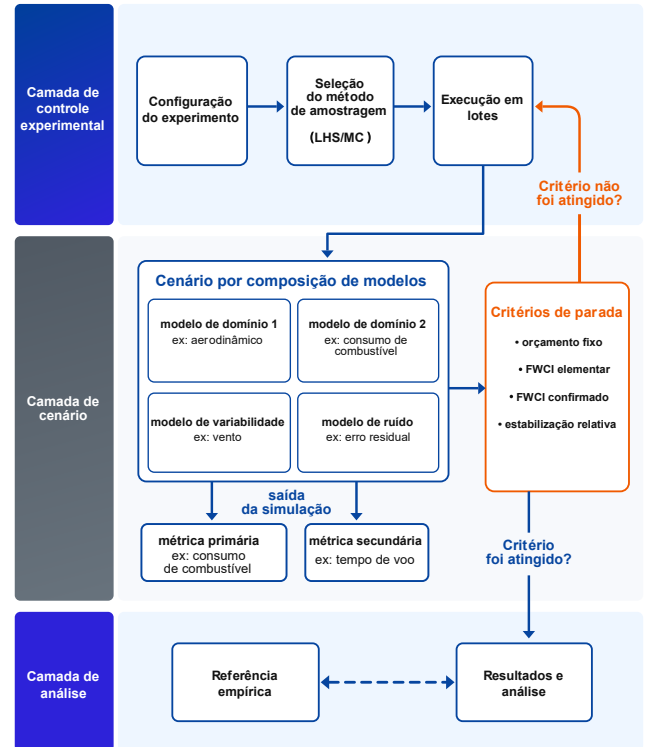


Fig. 1. Arquitetura modular do *HaltSim*, com camadas de controle experimental, cenário e análise.

modelos em uma execução unitária a partir das entradas e da semente da amostra.

A camada de amostragem transforma a configuração experimental em um plano reproduzível, com método amostral, entradas de cada execução e sementes associadas. A implementação usa `NumPy` para geração pseudoaleatória e operações numéricas, `SciPy` para o plano LHS e `pandas` para estruturas tabulares de campanha, permitindo reexecutar e auditar cada ponto.

A camada experimental executa o plano em lotes, acumula saídas e registra o progresso. Sobre esse histórico, a camada estatística calcula média, intervalo de confiança, semi-largura relativa e mudanças entre pontos de verificação; a camada de critérios de parada consome essas evidências por uma interface comum e decide se a campanha continua, para por sucesso ou encerra por limite máximo.

Por fim, a camada de comparação organiza os resultados finais por método amostral e critério de parada, avaliando custo amostral, erro em relação à referência empírica, precisão e cobertura.

Essa separação entre modelos, cenário, amostragem, execução, estatística, decisão e comparação é a principal contribuição arquitetural da ferramenta. O *HaltSim* funciona como um ambiente intermediário de experimentação: critérios de parada podem ser implementados, comparados e documentados em um cenário controlado antes de serem incorporados a ambientes de simulação mais complexos, nos quais integração, custo computacional, configuração operacional e interpretação das métricas tornam a avaliação mais difícil. Assim, a ferramenta reduz o risco de integrar uma regra ainda pouco caracterizada a um simulador de maior custo.

Do ponto de vista de engenharia de *software*, essa modularidade permite que novos cenários, modelos de domínio, métodos amostrais, critérios de parada ou relatórios sejam

incorporados sem reescrever o motor de execução em lotes. O cenário de rota econômica apresentado na metodologia é, portanto, uma instanciação dessa arquitetura, e não o limite conceitual do *HaltSim*.

IV. METODOLOGIA

Esta seção descreve como a arquitetura apresentada foi instanciada no estudo experimental. Primeiro, apresenta-se o cenário sintético de rota econômica e suas variáveis de entrada, fontes de variabilidade e métricas de saída. Em seguida, detalham-se a configuração amostral, os critérios de parada avaliados, a referência empírica usada na comparação e as métricas adotadas para analisar custo amostral, precisão, erro e cobertura.

A. Ferramenta de Avaliação e Cenário

O cenário avaliado, ilustrado na Fig. 2, representa uma aeronave em rota fixa de 150 NM e instancia a camada de cenário descrita na arquitetura. Para cada amostra, altitude e velocidade são definidas pelo plano amostral, nas faixas de 10000–35000 ft e 350–550 kt, enquanto a semente controla vento efetivo, modelado por $\mathcal{N}(0, 20 \text{ kt})$, e ruído residual lognormal multiplicativo com $\sigma = 0,03$. As saídas observadas são o consumo de combustível, usado como métrica primária, e o tempo de voo, usado como métrica auxiliar. A execução unitária distingue entradas controladas, variáveis estocásticas associadas à semente e métricas produzidas pelo cenário.

Embora simplificado, o modelo de consumo se inspira em relações clássicas de desempenho em cruzeiro, nas quais alcance e consumo dependem de velocidade, eficiência aerodinâmica e parâmetros propulsivos [20], [21]. Essas referências são usadas como orientação qualitativa para estruturar uma resposta positiva e sensível às variáveis de entrada, sem pretensão de reproduzir uma aeronave específica ou validar o cenário como simulador de alta fidelidade.

B. Amostragem, Sementes e Replicação

O desenho amostral principal é o *Latin Hypercube Sampling* (LHS), uma alternativa estratificada à amostragem aleatória simples proposta para melhorar a eficiência de estimadores em estudos de simulação [22]. Esse tipo de amostragem é recorrente em análises de incerteza e sensibilidade baseadas em simulação [23], [24], e comparações recentes continuam avaliando variantes de LHS em estudos

probabilísticos baseados em simulação [25]. Como linha de base amostral, utiliza-se *Monte Carlo* (MC) simples, com sorteios uniformes independentes de altitude e velocidade nas mesmas faixas.

A geração LHS foi feita por estratificação nas duas dimensões normalizadas, seguida de escala para as faixas físicas de altitude e velocidade. No caso MC, usou-se o mesmo gerador pseudoaleatório para produzir pontos uniformes independentes no hipercubo normalizado. Cada campanha recebeu uma semente de campanha distinta, de 2026 a 3025. A partir dela, foram gerados o plano amostral e as sementes de amostra usadas pelo cenário para vento e ruído. A referência empírica LHS usou semente distinta, igual a 3026.

A execução é organizada em lotes de 3 amostras. Como $n_{\min} = 100$, a primeira avaliação elegível dos critérios adaptativos ocorre em $n = 102$. O estudo executa 1000 campanhas para cada combinação de método de amostragem e critério de parada. O intervalo t é aplicado diretamente às observações acumuladas de consumo. Para LHS, esse uso é tratado como estimador operacional da incerteza acumulada, não como garantia formal de cobertura nominal.

C. Estratégias, Referência e Métricas

Para explicitar o desenho experimental, a comparação considera quatro critérios de parada: orçamento fixo, FWCI elementar, FWCI confirmado e uma heurística de estabilização relativa. O valor $n_{\min} = 100$ foi adotado como patamar mínimo pragmático para evitar decisões de parada baseadas em estatísticas acumuladas ainda muito instáveis. O limite $n_{\max} = 300$ foi definido igual ao orçamento fixo, de modo que todos os critérios fossem comparados sob o mesmo teto operacional de amostras. A Tabela I resume os parâmetros específicos de cada critério, enquanto os parâmetros globais do desenho experimental são consolidados na Tabela II.

TABELA I
ESTRATÉGIAS DE PARADA AVALIADAS

Estratégia	Parâmetros principais
Orçamento fixo	$N = 300$
FWCI elementar	95%, $\epsilon_r = 0,02$
FWCI confirmado	95%, $\epsilon_r = 0,02$, $k_{\text{confirm}} = 2$
Estabilização relativa	$\tau = 0,02$, tamanho dos lotes = 3

A referência empírica foi gerada por uma campanha LHS de alta amostragem, com $N_{\text{ref}} = 100000$, semente 3026 e média da métrica sintética `fuel_consumed` igual a 1115,61,

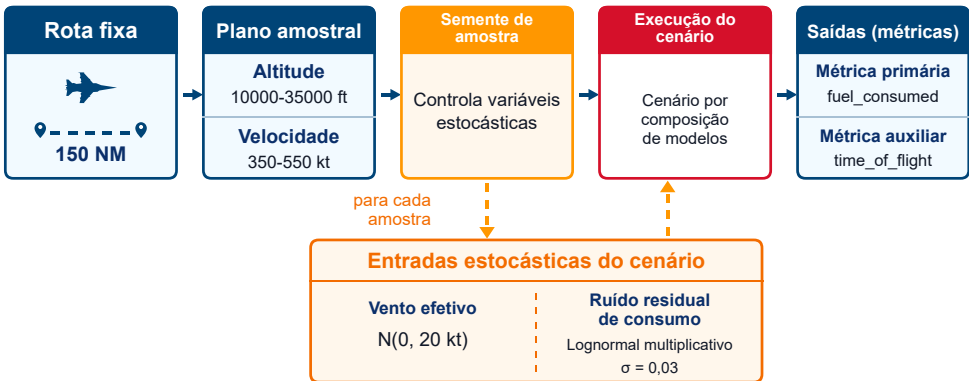


Fig. 2. Cenário sintético de rota econômica: rota fixa de 150 NM, altitude e velocidade do plano amostral, semente para vento e ruído residual, e saídas de consumo e tempo de voo.

no mesmo espaço contínuo de altitude e velocidade. Ela não é valor verdadeiro analítico, mas aproximação prática para avaliar erro, cobertura empírica e proximidade dos métodos. A incerteza da própria referência não é propagada nas taxas de cobertura reportadas; portanto, a cobertura é sempre cobertura contra essa referência empírica fixa.

TABELA II
DESENHO EXPERIMENTAL DO ESTUDO EMPÍRICO

Elemento	Valor
Referência empírica	LHS, $N_{\text{ref}} = 100000$
Semente da referência	3026
Média de referência	1115,61
Campanhas por combinação	1000
Métodos amostrais	LHS e MC
Crítérios de parada	4
Orçamento máximo	300 amostras
n_{min}	100 amostras
Tamanho do lote	3 amostras

As métricas principais são número de amostras, semi-largura relativa final, erro contra a referência, cobertura empírica, atendimento da meta de precisão, razão de parada e atingimento de n_{max} . A cobertura empírica é a proporção de campanhas cujo intervalo final contém a média da referência. A meta de precisão é atendida quando $h_n/\bar{x}_n \leq 0,02$. O custo computacional é representado pelo número de amostras executadas, usado aqui como indicador indireto de esforço de simulação. A execução dos 8000 resultados finais, sem reconstruir a referência empírica, levou aproximadamente 12,9 minutos em Windows 11 Pro, Intel Core i7-8550U, 15,9 GiB de RAM e Python 3.12.7.

V. RESULTADOS

O estudo executou 1000 campanhas para cada combinação de método amostral e critério de parada, totalizando 8000 resultados finais. Cada campanha foi comparada contra a mesma referência empírica LHS com $N_{\text{ref}} = 100000$. O orçamento fixo e o limite máximo dos critérios adaptativos seguem o teto operacional de 300 amostras definido na metodologia. A Tabela III resume os resultados agregados. Na tabela, $\bar{n}$ representa o número médio final de amostras por campanha.

Sob LHS, os critérios baseados em orçamento e FWCI produziram cobertura empírica acima do nível nominal de 95%, comportamento compatível com o uso de uma amostragem estratificada que, neste cenário sintético de baixa dimensão, reduziu a variabilidade amostral e tornou os intervalos operacionais mais conservadores em relação à referência empírica adotada. Sob MC, por sua vez, a cobertura permaneceu mais próxima do valor nominal. Essa leitura é local ao experimento: os resultados não constituem garantia geral de cobertura para LHS nem para o uso sequencial do intervalo t .

Conforme a Tabela III, as diferenças entre o FWCI elementar e o FWCI confirmado foram pequenas em termos de cobertura e erro médio. Neste cenário, a vantagem do FWCI confirmado deve ser interpretada principalmente como salvaguarda operacional: ele reduz a dependência de uma decisão tomada no primeiro cruzamento da fronteira de precisão, com custo adicional médio de poucas amostras.

Os critérios FWCI reduziram o custo amostral em relação ao orçamento fixo. Enquanto o orçamento fixo executou sempre 300 amostras, o FWCI elementar e o FWCI confirmado pararam, em média, com aproximadamente 244–248

amostras, o que representa economia de cerca de 17–19%. O custo adicional do critério confirmado foi pequeno, da ordem de poucas amostras em média, em troca da exigência de confirmação em lotes consecutivos. A Fig. 3 complementa a média reportada na Tabela III ao mostrar a dispersão do número final de amostras entre campanhas.

O método de amostragem também influenciou o erro contra a referência empírica. Nos critérios principais, o erro médio sob LHS foi menor que sob MC, tanto no orçamento fixo quanto nos critérios FWCI. Essa diferença indica que, para a resposta sintética de baixa dimensão avaliada, o LHS foi mais eficiente em aproximar a referência empírica. Como a própria referência empírica foi gerada por LHS, essa comparação de erro pode favorecer campanhas LHS em relação a MC. A conclusão deve permanecer restrita à ferramenta e ao cenário estudados, sem implicar superioridade universal do LHS sobre MC.

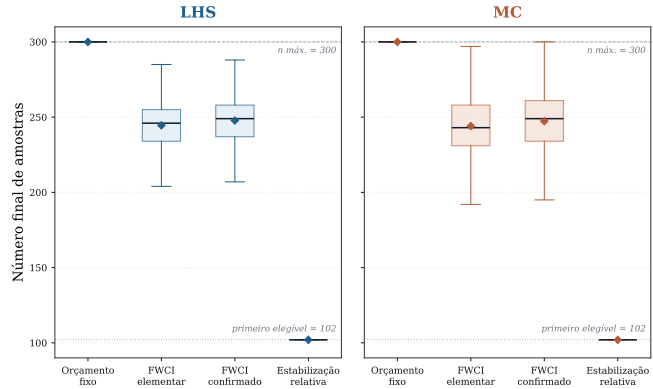


Fig. 3. Distribuição do número final de amostras por critério. Diamantes indicam médias; linhas centrais, medianas; horizontais, $n = 102$ e $n = 300$.

A heurística de estabilização relativa apresentou comportamento distinto. Como a configuração usa $n_{\text{min}} = 100$ e lotes de 3 amostras, a primeira avaliação elegível ocorre em $n = 102$; nesse ponto, a condição de estabilização relativa entre estimativas acumuladas já foi satisfeita em todas as campanhas, como evidencia a Fig. 3. Esse resultado mostra que a estabilização local da média pode ocorrer antes de a incerteza estatística ser reduzida até a meta externa de precisão. De fato, embora a cobertura empírica da heurística tenha permanecido próxima ou acima de 95%, a taxa de atendimento da meta de semi-largura relativa foi 0% em ambos os métodos amostrais. Essa cobertura deve ser interpretada com cautela: a parada precoce em $n = 102$ produz médias finais mais variáveis e com maior erro contra a referência, ao mesmo tempo em que o menor tamanho amostral aumenta a semi-largura dos intervalos. Assim, a cobertura observada resulta do balanço entre erro da média final e largura do intervalo: a referência ainda é contida na maior parte das campanhas, mas isso não implica atendimento da meta de precisão nem cobertura superior à do orçamento fixo. Assim, neste cenário, a heurística funciona como ponto de corte agressivo para estabilização da média, não como substituto dos critérios baseados em intervalo de confiança.

VI. CONCLUSÃO

Este artigo apresentou o *HaltSim* como uma ferramenta modular para avaliação preliminar de critérios de parada

TABELA III
RESULTADOS AGREGADOS POR MÉTODO AMOSTRAL E CRITÉRIO DE PARADA

Amostragem	Critério	Cob. (%)	Meta prec. (%)	Erro abs. médio	Semi-larg. rel. (%)	$\bar{n}$
LHS	Orçamento fixo	99,8	100,0	4,57	1,803	300,0
LHS	FWCI elementar	99,5	100,0	6,00	1,991	244,6
LHS	FWCI confirmado	99,6	100,0	5,87	1,978	247,8
LHS	Estabilização relativa	97,3	0,0	12,36	3,116	102,0
MC	Orçamento fixo	95,1	99,9	8,32	1,798	300,0
MC	FWCI elementar	94,2	99,9	9,33	1,992	244,0
MC	FWCI confirmado	94,3	99,9	9,30	1,978	247,3
MC	Estabilização relativa	95,6	0,0	14,22	3,107	102,0

em simulações estocásticas. Embora a demonstração use um cenário aeroespacial sintético, a contribuição de engenharia de *software* está na arquitetura que separa cenário, amostragem, execução, parada e análise para gerar cenários por composição de modelos. O recorte aeroespacial funciona, portanto, como aplicação inicial e caso de validação, não como limite conceitual da proposta.

Os resultados com 1000 campanhas por combinação indicaram que o método de amostragem influencia a interpretação empírica dos critérios. Sob LHS, a cobertura contra a referência empírica foi superior ao nominal e o erro absoluto foi menor; sob MC, a cobertura ficou próxima de 95%. Os critérios FWCI reduziram cerca de 17–19% das amostras em relação ao orçamento fixo, mantendo a meta de precisão nesta ferramenta de avaliação. A heurística de estabilização relativa reduziu drasticamente o custo amostral, mas não atingiu a meta de precisão por intervalo de confiança, devendo permanecer como linha de base operacional.

Esses resultados são locais a um cenário sintético de baixa dimensão e não demonstram garantia geral de cobertura nem fidelidade aeronáutica. As coberturas reportadas são medidas contra uma referência empírica fixa, e o intervalo t sob LHS deve ser lido como estimador operacional da incerteza acumulada. Como próximos passos, recomenda-se calibrar ϵ_r , k_{confirm} e o tamanho de lote em novos cenários sintéticos, comparar o intervalo t com alternativas como *bootstrap* e métodos de análise de saída de simulações, e integrar gradualmente os critérios selecionados a fluxos reais com AsaPy/ASA. No escopo avaliado, os resultados indicam que o *HaltSim* pode servir como ambiente preliminar para comparar critérios de parada antes de sua adoção em simulações de maior custo e complexidade.

REFERÊNCIAS

[1] M. Bicher, M. Wastian, D. Brunmeir, and N. Popper, “Review on monte carlo simulation stopping rules: How many samples are really enough?” *SNE Simulation Notes Europe*, vol. 32, no. 1, pp. 1–8, 2022.

[2] A. M. Law, *Simulation Modeling and Analysis*, 5th ed. New York, NY: McGraw-Hill Education, 2015.

[3] D. I. Singham, “Analysis of sequential stopping rules for simulation experiments,” Ph.D. dissertation, University of California, Berkeley, 2010.

[4] A. M. Law, “Statistical analysis of simulation output data: The practical state of the art,” in *2020 Winter Simulation Conference (WSC)*. IEEE, 2020, pp. 1117–1127.

[5] D. I. Singham and L. W. Schruben, “Analysis of sequential stopping rules,” in *Proceedings of the 2009 Winter Simulation Conference (WSC)*. IEEE, 2009, pp. 723–730.

[6] —, “Finite-sample performance of absolute precision stopping rules,” *INFORMS Journal on Computing*, vol. 24, no. 4, pp. 624–635, 2012.

[7] D. I. Singham, “Selecting stopping rules for confidence interval procedures,” *ACM Transactions on Modeling and Computer Simulation*, vol. 24, no. 3, pp. 1–18, 2014.

[8] Y. Yeh and S. C. Tsai, “Single criterion for rating confidence interval procedures: A generalization to stochastic stopping rules,” *Journal of the Operational Research Society*, vol. 76, no. 7, pp. 1481–1489, 2025.

[9] J. P. A. Dantas, A. N. Costa, V. C. F. Gomes, A. R. Kuroswiski, F. L. L. Medeiros, D. Geraldo, A. R. Samersla, S. R. Silva, A. F. M. Caetano, Y. D. Ferreira, A. O. Arantes, D. S. Santos, and M. R. C. Aquino, “ASA: A simulation environment for evaluating military operational scenarios,” in *Computational Science and Computational Intelligence: 11th International Conference, CSCI 2024, Las Vegas, NV, USA, December 11–13, 2024, Proceedings, Part XII*, ser. Communications in Computer and Information Science. Cham: Springer Nature Switzerland, 2025, vol. 2512, pp. 203–215.

[10] J. P. A. Dantas, S. R. Silva, V. C. F. Gomes, A. N. Costa, A. R. Samersla, D. Geraldo, M. R. O. A. Maximo, and T. Yoneyama, “AsaPy: A python library for aerospace simulation analysis,” in *Proceedings of the 38th ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*. New York, NY, USA: ACM, 2024, pp. 15–24.

[11] A. N. Costa, J. P. A. Dantas, E. Scukins, F. L. L. Medeiros, and P. Ögren, “Simulation and machine learning in beyond visual range air combat: A survey,” *IEEE Access*, vol. 13, pp. 76 755–76 774, 2025.

[12] N. Mukhopadhyay and S. M. Poruthotage, “Sequential fixed-width confidence interval procedures for the mean under multiple boundary crossings,” *Sequential Analysis*, vol. 32, no. 1, pp. 83–109, 2013.

[13] S. M. Poruthotage and N. Mukhopadhyay, “Multiple crossing sequential fixed-size confidence regions for regression parameters under normality,” *Sri Lankan Journal of Applied Statistics*, vol. 5, no. 4, pp. 147–182, 2014, special Issue: Modern Statistical Methodologies in the Cutting Edge of Science.

[14] G. Bayraktan and P. Pierre-Louis, “Fixed-width sequential stopping rules for a class of stochastic programs,” *SIAM Journal on Optimization*, vol. 22, no. 4, pp. 1518–1548, 2012.

[15] R. Lalehzari, E. Mahmoudi, and A. Khalifeh, “Sequential fixed-width confidence interval for the r th power of the exponential scale parameter: Two-stage and sequential sampling procedures,” *Sequential Analysis*, vol. 37, no. 3, pp. 293–310, 2018.

[16] D. A. Fitts, “Absolute precision confidence intervals for unstandardized mean differences using sequential stopping rules,” *Behavior Research Methods*, vol. 55, no. 4, pp. 1839–1862, 2023.

[17] A. Yousef and H. Hamdy, “The performance of multistage sequential sampling procedures for the mean of a normal population: A monte carlo simulation study,” *Sigma Journal of Engineering and Natural Sciences*, vol. 10, no. 3, pp. 265–276, 2019.

[18] R. R. Barton and L. A. Rhodes-Leader, “Bootstrap confidence intervals for simulation output parameters,” in *2023 Winter Simulation Conference (WSC)*. IEEE, 2023, pp. 421–432.

[19] L. Gong and J. M. Flegal, “A practical sequential stopping rule for high-dimensional markov chain monte carlo,” *Journal of Computational and Graphical Statistics*, vol. 25, no. 3, pp. 684–700, 2016.

[20] E. Torenbeek, “Cruise performance and range prediction reconsidered,” *Progress in Aerospace Sciences*, vol. 33, no. 5–6, pp. 285–321, 1997.

[21] T. M. Young, *Performance of the Jet Transport Airplane: Analysis Methods, Flight Operations, and Regulations*, 1st ed. Hoboken, NJ: Wiley, Nov. 2017.

[22] M. D. McKay, R. J. Beckman, and W. J. Conover, “Comparison of three methods for selecting values of input variables in the analysis of output from a computer code,” *Technometrics*, vol. 21, no. 2, pp. 239–245, 1979.

[23] J. C. Helton and F. J. Davis, “Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems,” *Reliability Engineering & System Safety*, vol. 81, no. 1, pp. 23–69, 2003.

[24] J. C. Helton, J. D. Johnson, C. J. Sallaberry, and C. B. Storlie, “Survey of sampling-based methods for uncertainty and sensitivity analysis,” *Reliability Engineering & System Safety*, vol. 91, no. 10–11, pp. 1175–1209, 2006.

[25] S.-S. Jin, G. Kim, S. Kwag, and S. Eem, “Feasibility study of progressive latin hypercube sampling and quasi-monte carlo simulation for probabilistic risk assessment,” *Geomatics, Natural Hazards and Risk*, vol. 15, no. 1, p. 2425185, 2024.