CONSTRUÇÃO DE UMA BASE DE DADOS DE BENCHMARK PARA O PROBLEMA DE HORÁRIO ESCOLAR A PARTIR DE DADOS REAIS DO SIGAH

Paulo Guilherme Schnaufer (PIBIC/CNPq/UEM), Ademir Aparecido Constantino (Orientador). E-mail: aaconstantino@uem.br.

Universidade Estadual de Maringá, Centro de Tecnologia, Departamento de Informática, Maringá, PR.

Área e subárea do conhecimento: Ciência da Computação/Matemática da Computação.

Palavras-chave: BENCHMARK; HORÁRIO ESCOLAR; XHSTT.

RESUMO

O problema da grade de horários escolares (*High School Timetabling* — HSTT) é NP-difícil, e bases de dados de *benchmark* padronizadas são de muita importância para comparar algoritmos de resolução, assim como fonte de dados para métodos *Deep Learning*. Este trabalho tem como objetivo construir uma base de dados de *benchmark* para tal problema a partir de dados reais de escolas brasileiras, extraídos do sistema SIGAH-UEM, usado em todos os estados do país. A metodologia consiste em um *pipeline* que converte dados brutos em formato *.csv*, para o padrão internacional XHSTT em *.xml*. Foram geradas 3.421 instâncias, com tamanhos de XML variando entre 156 e 23.417 linhas. Na etapa de validação pelo HSEval, 3.106 instâncias foram confirmadas como sintaticamente válidas, enquanto 315 foram descartadas devido a inconsistências cadastrais de origem (como falhas de integridade referencial nos registros das escolas). Das instâncias válidas, 3.058 foram processadas pelo resolvedor KHE, resultando em 994 soluções ótimas (custo zero) e 2.064 soluções com penalidades. A base construída supera em quase duas ordens de grandeza os repositórios públicos existentes. Conclui-se que o *pipeline* é eficaz para a construção e validação de uma base de *benchmark* pública de grande porte para o problema de horário escolar.

INTRODUÇÃO

O problema da grade de horários escolares consiste em alocar aulas, turmas, professores e horários de forma a atender um conjunto de restrições rígidas e flexíveis das instituições de ensino. Trata-se de um problema de otimização combinatória classificado como NP-difícil, o que torna inviável a busca exaustiva por soluções ótimas em instâncias de tamanho realista. Por conta disso, bases de dados de *benchmark* públicas e padronizadas são essenciais, permitindo que diferentes algoritmos sejam testados sobre os mesmos dados e comparados de forma justa, o que impulsiona o avanço da pesquisa na área (POST et al., 2014). Entretanto, as instâncias reais disponíveis na literatura são tanto limitadas em número (reunindo

cerca de 40 instâncias), quanto nem sempre seguem um formato padrão (SAVINIEC; CONSTANTINO, 2017). Este trabalho, de natureza aplicada, tem como objetivo construir uma base de dados de *benchmark* para o problema de horário escolar a partir de dados reais de escolas brasileiras, utilizando como fonte o sistema SIGAH-UEM, empregado por mais de 3500 escolas no país (SIGAH-UEM, 2026). O escopo do trabalho é a construção do *pipeline* de extração e padronização dos dados no formato internacional XHSTT.

MATERIAIS E MÉTODOS

A fonte de dados primária é o sistema SIGAH-UEM (SIGAH-UEM, 2026). Os dados de cada uma (turmas, professores, disciplinas, disponibilidades e restrições) são extraídos em *dumps* .csv compostos pelas tabelas: *cenario_dias*, *cenario_periodos*, *cenario*, *disciplina*, *evento*, *indisponibilidade*, *professor*, *programacao* e *turma*.
O *pipeline* desenvolvido realiza a leitura, o cruzamento relacional e a validação prévia dos dados sem intervenção manual. Então um serializador converte cada instância válida para a especificação XML do padrão XHSTT (POST et al., 2014). Por fim, as instâncias geradas são submetidas à validação sintática e semântica por meio do avaliador HSEval (KINGSTON, 2012) e do resolvedor genérico KHE.

RESULTADOS E DISCUSSÃO

O *pipeline* processou um total de **3.421 instâncias**. A variação no tamanho dos arquivos XML gerados (entre 156 e 23.417 linhas) evidencia a diversidade estrutural e de porte entre as instituições processadas. Na fase de validação via HSEval (KINGSTON, 2012), **315 instâncias** (aproximadamente 9,2% do total) foram classificadas como sintaticamente inválidas e descartadas. A análise diagnóstica revelou que os motivos do descarte residem em inconsistências nos dados de origem (ponta da escola):

- Cadastros incompletos no SIGAH-UEM: turmas ou disciplinas sem professores vinculados;
- Erros de integridade referencial nos CSVs: IDs citados em horários que não constavam no cadastro principal;
- Incompatibilidade XML: eventos com duração nula ou restrições conflitantes com a estrutura do XHSTT.

Das **3.106 instâncias** válidas, 3.058 foram executadas com sucesso pelo resolvedor KHE. Dentre estas, **994 instâncias** (32,5%) atingiram solução ótima comprovada (Custo = 0); **2.064 instâncias** (67,5%) foram resolvidas com penalidades (Custo > 0), representando problemas com maior densidade de restrições ou concorrência de recursos.
A base final obtida supera amplamente o maior repositório público de instâncias reais em XHSTT disponível na literatura (UNIVERSITY OF TWENTE, 2026), estabelecendo um novo patamar de escala para experimentos na área.

CONCLUSÕES

A automação do *pipeline* de extração e serialização provou ser plenamente viável para a construção de uma base de dados de *benchmark* de grande porte para o problema de horário escolar a partir do SIGAH-UEM. Com mais de 3.000 instâncias validadas, o trabalho disponibiliza um acervo representativo da realidade das escolas brasileiras. Como trabalho futuro, pretende-se disponibilizar a base publicamente para a comunidade científica.

AGRADECIMENTOS

Os autores agradecem ao CNPq pela concessão da bolsa PIBIC e à Universidade Estadual de Maringá pelo apoio institucional ao desenvolvimento da pesquisa.

REFERÊNCIAS

KINGSTON, J. H. *HSEval: High School Timetabling Evaluator*. 2012. Disponível em: http://jeffreykingston.id.au/cgi-bin/hseval.cgi. Acesso em: 28 ago. 2026.

POST, G.; KINGSTON, J. H.; AHMADI, S.; DASKALAKI, S.; GOGOS, C.; KYNGAS, J.; NURMI, C.; MUSLIU, N.; PILLAY, N.; SANTOS, H.; SCHAERF, A. XHSTT: an XML archive for high school timetabling problems in different countries. *Annals of Operations Research*, v. 218, n. 1, p. 295-301, 2014.

SAVINIEC, L.; CONSTANTINO, A. A. Effective local search algorithms for high school timetabling problems. *Applied Soft Computing*, v. 60, p. 363-373, 2017.

SIGAH-UEM. *Sistema de Gestão de Horários Escolares da Universidade Estadual de Maringá*. 2026. Disponível em: https://www.sigah.uem.br/. Acesso em: 28 ago. 2026.

UNIVERSITY OF TWENTE. *XHSTT Datasets: High School Timetabling Archive*. Data Management and Mathematical Programming Group (DMMP). 2026. Disponível em: https://www.utwente.nl/en/eemcs/dmmp/hstt/datasets/. Acesso em: 28 ago. 2026.