

Abordagem multiagente para modelagem da progressão de estudantes

Victor Lima , Julia Farias, Renato Lopes, Matheus Souza e Alim Gonçalves

Resumo— Uma gestão escolar eficaz pode se beneficiar de modelos precisos de progressão estudantil para otimizar a alocação de professores e espaços de ensino. No entanto, as pesquisas existentes concentram-se predominantemente em modelos markovianos para programas baseados em anos, deixando uma lacuna na literatura no que diz respeito a cursos de graduação baseados em créditos. Este artigo propõe uma estrutura para modelar a progressão de grupos de estudantes em programas baseados em créditos usando cadeias de Markov, sem a necessidade de treinamento extensivo. Para elaborar o modelo, analisamos dados de progressão de estudantes em cursos de engenharia para extrair distribuições de probabilidade de reprovação, matrícula e evasão. Essas distribuições foram então utilizadas para desenvolver modelos de cadeia de Markov, os quais foram simulados por meio de uma abordagem multiagente. Nossos resultados indicam que um modelo de cadeia de Markov pode emular efetivamente a progressão dos estudantes em cursos de engenharia, e que os resultados da simulação multiagente correspondem de forma bastante próxima aos dados reais de progressão. Essa estrutura oferece uma ferramenta valiosa para a gestão escolar e para a previsão da demanda por disciplinas em programas de graduação baseados em créditos.

Palavras-Chave— Cadeias de Markov, Simulação Multiagente, Gestão Escolar

I. INTRODUÇÃO

Um grande desafio no contexto da administração escolar é prever adequadamente a progressão dos estudantes ao longo de um programa. Se os gestores se baseiam exclusivamente na proposta de andamento do curso formulada pela escola para o planejamento, essa disparidade pode levar ao desperdício de recursos como abertura de turmas sem demanda dos alunos, à falta de vagas em determinados cursos e a outros problemas que levam à retenção exagerada e à evasão escolar. Esta progressão é afetada por diversas variáveis ambientais e pedagógicas que causam uma baixa adesão às propostas de cumprimento dos currículos.

Diversos trabalhos na literatura modelam a matrícula e a progressão de estudantes utilizando abordagens baseadas em cadeias de Markov. Por exemplo, em “Markov Analysis of Students’ Performance and Academic Progress in Higher

Education” [1], os autores apresentam um modelo estocástico para estimar e monitorar continuamente diversos indicadores de qualidade e eficácia de um programa de ensino superior. O progresso dos alunos no programa é modelado por uma cadeia de Markov finita com cinco estados transitórios (um para cada ano do programa) e dois estados absorventes que modelam a formatura e a desistência. A matriz de transição de probabilidades é estimada a partir dos dados, e as características quantitativas da cadeia de Markov absorvente, como o tempo esperado até a absorção e as probabilidades de absorção, são utilizadas para avaliar indicadores selecionados do programa. O modelo é aplicado para investigar o padrão de matrícula dos estudantes e seu desempenho acadêmico em uma instituição de ensino superior eslovena. A progressão dos estudantes para a próxima etapa do programa de estudos foi estimada, juntamente com as probabilidades de abandono. Essa abordagem é similar à apresentada no artigo “Planning models in higher education” [2].

Em “Simulating the Effect of Student Profile Changes on Retention and Graduation Rates: A Markov Chain Analysis” [3], os autores propõem um método para avaliar o impacto de características estudantis, como média ponderada (GPA) e carga horária de créditos, na taxa de graduação. Este estudo baseou-se em modelos de Markov para avaliar a probabilidade de progressão e concluiu que mudanças relevantes nas notas têm baixo impacto nas taxas de graduação.

No artigo “Markov Analysis of Academic Performance of Students in Higher Education: a case study of an engineering institution”[4], os autores também focam no desenvolvimento e análise de um modelo de Markov com estados absorventes para avaliar o desempenho acadêmico dos estudantes em um programa de engenharia de quatro anos em uma instituição localizada na Índia. Os pesquisadores desenvolveram uma matriz de transição de probabilidades para o progresso acadêmico dos estudantes, utilizando a matriz de frequência obtida a partir de dados históricos. As probabilidades de progressão dos estudantes entre anos consecutivos do programa acadêmico foram calculadas usando essas probabilidades de transição. Por meio do processo Markoviano, determinou-se o número esperado de anos em um estado transitório especificado antes da absorção e as probabilidades de absorção. A análise dos resultados revelou que aulas de reforço e atividades de mentoria desempenharam um papel significativo no aumento das taxas de graduação e na redução das probabilidades de evasão.

O artigo intitulado “The College Completion Puzzle: A Hidden Markov Model Approach” [5] aborda a questão das baixas taxas de conclusão universitária nos Estados Unidos, especialmente entre estudantes de instituições menos seletivas.

Victor Lima (vlima@decom.fee.unicamp.br), Julia Farias (j200130@dac.unicamp.br) e Renato Lopes (rlopes@unicamp.br) são do Departamento de Comunicações da Faculdade de Engenharia Elétrica e de Computação - Universidade Estadual de Campinas, Campinas-SP; Matheus Souza (matsouza@unicamp.br) e Alim Gonçalves (alimiped@unicamp.br) são do Departamento de Engenharia de Computação e Automação da Faculdade de Engenharia Elétrica e de Computação - Universidade Estadual de Campinas, Campinas-SP. Este trabalho foi parcialmente financiado por CNPQ (310824/2021-4) e FAEPEX (2466/23). O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento 001.

Os pesquisadores analisam dados de transcrição longitudinal para examinar os processos que levam à graduação, utilizando modelos ocultos de Markov. Eles identificam vários estados latentes associados aos padrões de escolha de disciplinas e mostram que um modelo oculto de Markov treinado pode prever a probabilidade de graduação ou não graduação de um aluno específico com base em apenas alguns semestres de dados de histórico. Os resultados das trajetórias do modelo oculto de Markov indicam que tanto os estudantes graduados quanto os não graduados cursam disciplinas matemáticas e técnicas mais difíceis em taxas iguais. No entanto, os estudantes que concluem o bacharelado em até seis anos são mais propensos a alternar entre semestres com carga horária pesada e aqueles com carga horária mais leve.

O artigo “Enrollment Projection Using Markov Chains: Detecting Leaky Pipes and the Bulge in the Boa” [6] explora a aplicação de cadeias de Markov para prever a progressão de matrículas no ensino superior. Diferentemente dos modelos tradicionais que utilizam o nível do estudante como variável de classificação, este estudo emprega créditos estudantis acumulados para acompanhar o movimento dos estudantes de um período acadêmico para outro, dividindo-os por porcentagem de conclusão do programa em vez de períodos. Essa abordagem permite uma melhor identificação de quando os estudantes entram e saem da instituição como desistentes ou graduados. Os resultados sugerem alta previsibilidade de matrícula de ano para ano e revelam que a matriz de probabilidades de transição pode ajudar os administradores a identificar tendências e irregularidades no processo de gestão de matrículas. O estudo também aborda o problema de evasão e retenção dos estudantes, que ocorre quando eles abandonam ou deixam de frequentar devido a vários motivos. Ao isolar onde ocorrem as saídas do programa, as instituições podem identificar causas e implementar estratégias para reter estudantes e manter um fluxo constante ao longo do programa.

Como uma alternativa às abordagens baseadas em Markov, o artigo “Predicting Students’ Progression in Higher Education by Using the Random Forest Algorithm” [7] propõe o uso de um algoritmo Random Forest (RF) para identificar as variáveis que mais precisamente preveem a progressão dos estudantes. Os autores utilizaram dados da Manchester Metropolitan University, incorporando conjuntos de dados do ambiente virtual de aprendizagem (VLE), sistemas de informações gerenciais e registros estudantis. O algoritmo RF foi considerado apropriado para este estudo devido ao grande volume de dados disponível para análise. O artigo também discute resultados iniciais com base em dados do ano acadêmico de 2007–2008. Os resultados sugerem que variáveis como o horário de acesso ao VLE, a última vez em que os estudantes fizeram login no VLE e o número de acessos a documentos por parte dos professores são os principais preditores de progressão dos estudantes.

Todos esses artigos tentam modelar matrícula e progressão de estudantes. Exceto pela tentativa com Random Forest, todos os modelos baseados em processos de Markov utilizam um modelo de programa serializado, no qual um estudante que não é aprovado em um ano inteiro deve refazer todos os componentes do ano mencionado. A maioria dos mo-

delos também não abrange disciplinas eletivas e taxas de evasão. Como também afirmado em “Recruitment, Retention and student Flow: A Comprehensive Approach to Enrollment Management Research” [8], modelos baseados em cadeias de Markov são sensíveis ao período de tempo em que os dados estatísticos que alimentam o modelo foram coletados. Este artigo também detectou que diferenças nas populações são ignoradas por modelos de Markov, que as agrupam como uma única população.

Para tratar essa lacuna na literatura, elaboramos um modelo baseado em cadeia de Markov desenhado para tratar programas baseados em créditos. O objetivo é prover uma ferramenta que, a partir de dados passados como taxas de reprovação e de evasão, obtenha um bom estimador para a demanda de disciplinas no futuro e os efeitos que mudanças de currículo exercerão sobre essas demandas.

II. MODELAGEM DO PROGRAMA E PROGRESSÃO

Para emular a progressão de um estudante ao longo de um programa de ensino superior, simulamos a conclusão dos créditos necessários para a graduação, que é o modelo adotado na maioria das universidades brasileiras. Vamos utilizar para a elaboração do modelo os critérios de formação da UNICAMP, mas ele pode ser adaptado facilmente para outras universidades. A maioria desses créditos é dividida em m disciplinas obrigatórias; os créditos restantes, denotados por o , devem ser concluídos com disciplinas eletivas, que podem ser organizadas em grupos. Assim, para se formar, um estudante deve ter concluído todas as disciplinas obrigatórias e, simultaneamente, ter completado os créditos necessários para cada grupo de disciplinas eletivas. Os créditos obrigatórios não podem ser substituídos por outra disciplina semelhante, e os grupos eletivos podem restringir as disciplinas que contam para a conclusão.

O núcleo do modelo proposto é uma estrutura que simula o histórico acadêmico de um único estudante. Como a progressão agregada dos créditos depende exclusivamente da matrícula e da taxa de sucesso em cada disciplina, podemos visualizar um modelo de estados dependente das disciplinas necessárias para a graduação. No entanto, os estudantes só podem ser aprovados em disciplinas nas quais tenham se matriculado, então precisamos de estados intermediários para armazenar informações sobre a matrícula, um para cada disciplina. Também devemos levar em consideração o período de cada estudante, pois a lista de disciplinas sugeridas para matrícula depende desse parâmetro. Na UNICAMP, os programas são divididos em semestres. O modelo final corresponde a uma cadeia de Markov com $3^{K(m+o)} + 3$ estados, onde m é o número de disciplinas obrigatórias, K é o número máximo de períodos que um estudante pode permanecer no programa, e o é o número de créditos eletivos que o estudante deve obter para se formar. Há três estados absorventes adicionais: um para modelar a evasão, como apresentado na seção III, um para modelar os estudantes que excedem o limite de tempo máximo para se formar e são desligados do curso e o estado absorvente que representa os estudantes que concluíram o programa. O tamanho desse problema indica uma necessidade

não escalável de recursos computacionais; no entanto, muitos desses estados são inalcançáveis (muitas disciplinas têm pré-requisitos, bloqueando estados intermediários correspondentes a disciplinas que o estudante não pode cursar), e um estudante não pode alcançar um estado correspondente à aprovação em uma disciplina na qual ele não tenha previamente se matriculado. Esses estados inalcançáveis tornam nossa matriz de transição de Markov altamente esparsa. De fato, as únicas entradas não nulas da matriz de transição serão:

- Se o estado atual está no período k antes da matrícula em disciplinas:
 - A entrada que leva ao estado absorvente de tempo máximo excedido se $k > K$ — nesse caso, todas as outras transições têm probabilidade 0;
 - A entrada que leva ao estado absorvente de evasão com a probabilidade correspondente detalhada na seção III;
 - As entradas que levam aos estados que modelam a matrícula nas disciplinas disponíveis no período k , dada o histórico de aprovações, com distribuição de probabilidade detalhada na seção IV.
- Se o estado atual está no k -ésimo período após a matrícula:
 - As entradas que levam a todas as possibilidades de aprovação nas disciplinas matriculadas no período $k + 1$, incluindo o estado final de graduação. As probabilidades são dadas pelas taxas históricas de aprovação ou reprovação nas disciplinas matriculadas.

Dada a esparsidade da matriz de transição, em nossa implementação consideramos apenas as probabilidades não nulas e utilizamos a cadeia de Markov resultante para simular possíveis progressões de estudantes no curso, que correspondem a possíveis caminhos na cadeia de Markov. Para implementar nosso modelo, a cada temporada de admissão, N estudantes são criados com históricos acadêmicos vazios e um catálogo de progressão de disciplinas que dita os requisitos de graduação. Essa flexibilidade dos catálogos nos permite modelar mudanças nos mesmos e observar os efeitos da transição entre catálogos.

Para cada um dos N estudantes, o estado atual é armazenado como um vetor de codificação, onde cada uma das m entradas corresponde às disciplinas obrigatórias e aos grupos de eletivas. O valor de cada entrada codifica se o estudante não foi aprovado nem está matriculado, está atualmente matriculado ou foi aprovado na respectiva disciplina. Para cada disciplina obrigatória, estabelecemos a partir do histórico escolar uma taxa de aprovação.

Em cada período, é feito um sorteio usando as probabilidades de transição para avaliar se um dado estudante será aprovado em uma disciplina matriculada. O processo de matrícula será apresentado com mais detalhes na seção IV. O modelo de transição atualizará cada um dos componentes dos requisitos, e o estudante se formará quando todas as disciplinas obrigatórias forem concluídas e os grupos de disciplinas eletivas forem completados. Esse estado final é absorvente e não há transição a partir dele. A figura 1 apresenta um esquema simplificado

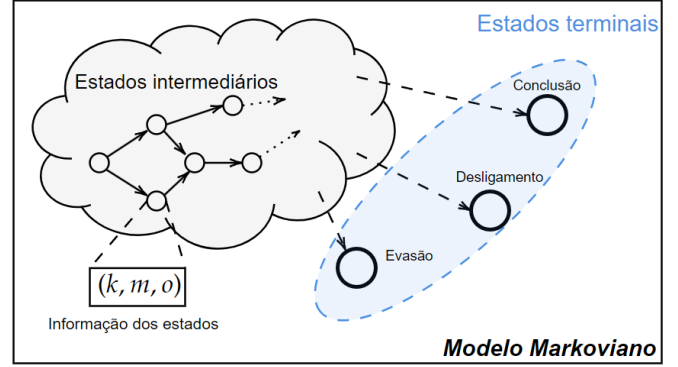


Fig. 1: Modelo de cadeia de Markov da progressão dos estudantes. A cadeia simula a progressão e conclusão do programa com três estados terminais: conclusão, evasão e desligamento por tempo. Estados intermediários representam disciplinas matriculadas e concluídas, enquanto a existência de pré-requisitos criam uma matriz de transição esparsa.

do espaço de estados da cadeia de Markov do modelo.

III. MODELAGEM DE EVASÃO

Embora a evasão acadêmica seja uma das maiores preocupações em programas de ensino superior, ela é geralmente atribuída a fatores externos que não estão diretamente relacionados ao desempenho acadêmico. Assim, modelamos a evasão acadêmica como uma probabilidade fixa d_k de um estudante desistir do curso em qualquer período k antes de atingir o limite máximo K . Essas probabilidades podem variar entre períodos (por exemplo, podem ser maiores nos primeiros períodos, onde a adaptação ao programa ocorre).

Para capturar a evasão acadêmica, incluímos um estado absorvente na cadeia de Markov, que representa os estudantes que abandonaram o programa. A transição para esse estado ocorre no início de cada período e a probabilidade de evasão é extraída de distribuições históricas para o programa estudado. Portanto, a taxa de evasão observada em nosso modelo dependerá não apenas das probabilidades d_k , mas também do número de estudantes que permanecem matriculados em cada período.

IV. MATRÍCULA E PROGRESSÃO ACADÊMICA

A matrícula em disciplinas no período k é determinada pelas regras do catálogo de cursos e pelas restrições de pré-requisitos. O processo de matrícula para um estudante em particular é simulado como uma distribuição probabilística dependente do estado atual do estudante. Consideramos as seguintes condições para a matrícula:

- O estudante só pode se matricular em disciplinas que ainda não concluiu e que têm seus pré-requisitos satisfeitos;
- O número total de créditos em que um estudante pode se matricular em um único período é limitado por $C_{\max}$;
- O estudante seleciona aleatoriamente as disciplinas a serem cursadas com base na disponibilidade e na preferência histórica dos estudantes.

O modelo também permite que um estudante seja impedido de cursar determinadas disciplinas por várias razões, como conflitos de horário ou restrições administrativas. Em cada

período, o nosso modelo prevê que um estudante tenta se matricular em tantas disciplinas quanto possível, respeitando o limite de créditos $C_{\max}$ e as restrições do programa. A probabilidade de matrícula em uma disciplina específica é baseada em taxas históricas de matrícula.

Após a matrícula, a progressão acadêmica é simulada utilizando as taxas de aprovação e reprovação extraídas de dados históricos. Se um estudante reprova em uma disciplina, ele terá a oportunidade de cursá-la novamente em períodos subsequentes, desde que ainda esteja matriculado no programa e não tenha excedido o limite de tempo máximo K . A taxa de reprovação é controlada pelos mesmos fatores que influenciam a matrícula inicial, e o progresso acadêmico global de um estudante depende diretamente de seu desempenho em cada tentativa.

Dessa forma, o modelo é capaz de simular a trajetória acadêmica de um grupo de estudantes, identificando os gargalos no progresso e fornecendo insights sobre o impacto de mudanças nas políticas acadêmicas ou no catálogo de cursos. Nos experimentos realizados neste trabalho consideramos o corpo de estudantes como um único grupo.

V. RESULTADOS OBTIDOS

Para avaliar os resultados, comparamos a simulação com dados anonimizados e compilados do escritório de registros da Universidade Estadual de Campinas, especificamente para alunos de Engenharia Elétrica e de Computação de 2013 a 2023. Os dados de 2020 e 2021 foram excluídos devido a medidas relacionadas à pandemia, que introduziram outliers. As probabilidades de entrada e as taxas de aprovação foram estimadas com base em dados de 2009 a 2019.

Para as taxas de admissão, modelamos a entrada de novos estudantes em cada semestre ímpar da seguinte forma, baseada nos números de ingressantes nos respectivos cursos na Unicamp em 2025: 70 estudantes para o programa integral de Engenharia Elétrica, 30 para o programa noturno de Engenharia Elétrica e 90 para Engenharia de Computação. A simulação utiliza o mesmo currículo que os dados reais, começando com um grupo inicial de 190 estudantes, como se o programa estivesse sendo iniciado do zero.

A primeira característica avaliada foi a estabilidade do modelo. A taxa de novos agentes ativos (por exemplo, alunos ingressantes) deve, em algum momento, atingir um equilíbrio com a taxa de absorção pelos estados terminais (por exemplo, alunos que concluíram o curso). Na Figura 2, observamos que nosso modelo, após uma resposta transitória inicial, alcança uma oscilação estável em torno de um número constante de estudantes ativos. Essa oscilação ocorre porque novos agentes são admitidos anualmente, enquanto as transições para os estados absorventes ocorrem a cada semestre.

Após avaliar a estabilidade do modelo, comparamos suas saídas com os dados reais. Nosso modelo alcançou um estado estacionário após 14 semestres. Dado que os três programas usados para comparação estão em operação há tempo suficiente para gerar estatísticas estáveis, executamos nosso modelo por 24 semestres em 500 simulações, calculando a média do número de estudantes ativos, formados e evadidos ao final

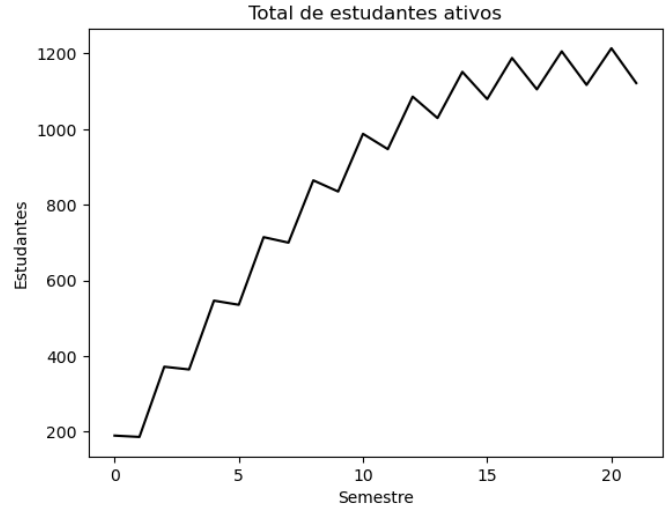


Fig. 2: Média de estudantes ativos por semestre

TABELA I: Estatísticas da simulação e dos dados reais

	Simulação	Dados reais
Engenharia Elétrica (integral)		
Ativos	409	358
Formados	328	320
Evasão	103	163
Engenharia Elétrica (noturno)		
Ativos	200	178
Formados	109	115
Evasão	51	80
Engenharia de Computação		
Ativos	171	153
Formados	98	83
Evasão	32	18

TABELA II: Tempo médio de conclusão do curso

	Simulação	Dados reais
EE (integral)	13.13	11.85
EE (noturno)	15.06	12.62
EC	14.55	11.96

desse intervalo. A Tabela I compara as estatísticas do estado estacionário do modelo com os dados reais ao fim do segundo semestre de 2024, fornecidos pela diretoria acadêmica da UNICAMP. Uma pequena discrepância no número de evasões é provavelmente devido à ao baixo número de alunos evadidos, o que dificulta a estimativa das probabilidades de transição e ao ruído estatístico inerente. As taxas de conclusão de cursos, no entanto, são previstas com maior precisão, já que o número de estudantes formados é maior.

A Tabela II compara o tempo médio, em semestres, que cada estudante levou para se formar. Agentes simulados pelo modelo de Markov consistentemente demoraram mais para se formar do que os estudantes reais. No entanto, as curvas de progressão mostradas nas Figuras 3 a 5 indicam que, embora os agentes demorem mais para completar todos os créditos, sua progressão ao longo do programa se assemelha bastante à dos estudantes reais. Esses resultados indicam que a diferença no tempo médio de conclusão se deve ao fato de os agentes não se matricularem em créditos eletivos, uma consequência direta do modelo estocástico para créditos matriculados por semestre. Ajustar os parâmetros de distribuição pode ajudar a resolver essa discrepância.

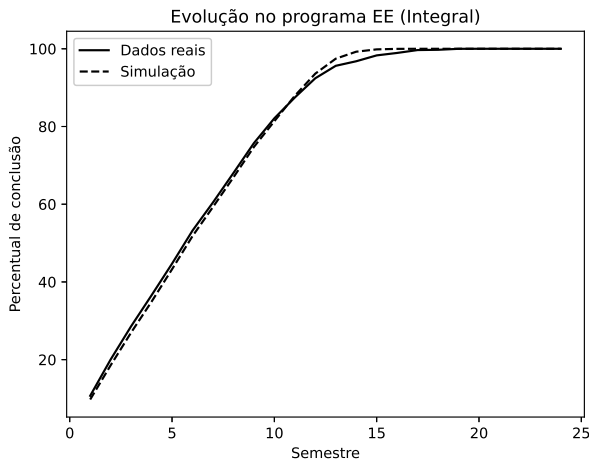


Fig. 3: Progressão média no curso - EE (integral)

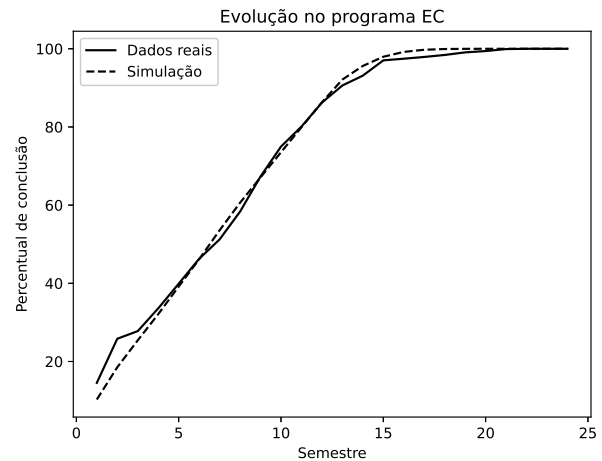


Fig. 5: Progressão média no curso - EC

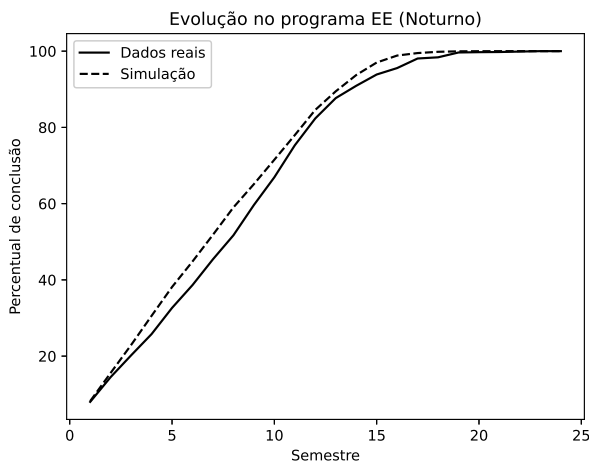


Fig. 4: Progressão média no curso - EE (noturno)

As curvas de conclusão de curso indicam que nosso modelo simula de forma razoável as preferências dos estudantes em relação à carga de créditos por semestre. Nosso modelo não considera atrasos intencionais nas matrículas em disciplinas. Esse comportamento é bastante comum entre estudantes do curso noturno de Engenharia Elétrica, devido à flexibilidade no horário que permite cursar créditos de dia. A Figura 4 ilustra essa diferença, com agentes do modelo cursando mais créditos por semestre do que esse subgrupo, resultando em uma curva de progressão mais íngreme. Embora isso não impacte as taxas de formatura, afeta ligeiramente as estimativas de matrícula no estado estacionário, o que pode influenciar decisões de gestão baseadas na simulação.

VI. CONCLUSÕES

Este artigo apresentou um modelo markoviano multiagente para simular a progressão de estudantes em um programa de ensino superior baseado em créditos. Propusemos distribuições de probabilidade para eventos-chave, como evasão, reprovações em disciplinas e matrículas, e descrevemos os estados e as probabilidades de transição necessárias para a construção de uma matriz de transição de uma cadeia de Markov. Após

o desenvolvimento do modelo, avaliamos sua complexidade computacional e constatamos que ele pode ser escalado de forma eficiente com requisitos razoáveis de recursos.

Para validar o modelo, comparamos seus resultados com dados reais de estudantes matriculados nos cursos de Engenharia Elétrica e de Computação da Universidade de Campinas. Nossos achados indicam que o modelo emula satisfatoriamente o progresso dos estudantes e prediz o comportamento do grupo, fornecendo informações úteis para a gestão escolar e a alocação de recursos. Desenvolvimentos futuros podem aprimorar o modelo ao incorporar características específicas dos programas e refinar a modelagem da evasão, o que aumentaria sua precisão.

REFERENCES

- [1] A. Brezavšček, M. P. Bach, and A. Baggia, "Markov Analysis of Students' Performance and Academic Progress in Higher Education," *Organizacija*, vol. 50, pp. 83–95, May 2017.
- [2] R. H. Stein, "Planning models in higher education," *Higher Education*, vol. 11, no. 6, pp. 725–729, 1982.
- [3] V. M. Borden and J. F. Dalphin, "Simulating the effect of student profile changes on retention and graduation rates: A markov chain analysis," *Annual Forum of the Association for Institutional Research*, 1998.
- [4] K. Eldose, B. Mayureshwar, K. R. Kumar, and R. Sridharan, "Markov analysis of academic performance of students in higher education: a case study of an engineering institution," *International Journal of Services and Operations Management*, vol. 41, no. 1/2, p. 59, 2022.
- [5] D. Witteveen and P. Attewell, "The College Completion Puzzle: A Hidden Markov Model Approach," *Research in Higher Education*, vol. 58, pp. 449–467, June 2017.
- [6] R. Gandy, L. Crosby, A. Luna, D. Kasper, and S. Kendrick, "Enrollment Projection Using Markov Chains: Detecting Leaky Pipes and the Bulge in the Boa," *AIR Professional File*, no. Fall 2019, p. 147, 2019.
- [7] J. Hardman, A. Paucar-Caceres, and A. Fielding, "Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm," *Systems Research and Behavioral Science*, vol. 30, pp. 194–203, Mar. 2013.
- [8] P. T. Ewell, "Recruitment, retention and student flow: A comprehensive approach to enrollment management research," *National Center for Higher Education Management Systems*, 1985.