

Análise Estatística Descritiva e Regressão da Inserção das Mulheres nos Cursos de TI nos Anos de 2009 a 2018

Ariana R. Cursino¹, Juliana F. P. Martinez¹

¹Faculdade de Tecnologia de São José dos Campos - Professor Jessen Vidal.
Av. Cesare Mansueto Giulio Lattes, 1350. Eugênio de Melo.
São José dos Campos/SP, 12247-014, Brasil.

{ariana.cursino,pasquini.juliana01}@fatec.sp.gov.br

Abstract. *The courses in the Technology area are mostly occupied by men. In order to provide metrics for the coordination of courses and communities that work with women in the field of technology, a descriptive statistical analysis of data from the Higher Education Census from 2009 to 2018 is necessary. The present work aims to verify the number of women in relation to the number of men who entered IT courses, as well as a regression analysis to verify the equation of growth rate of men and women in the courses. The scope of this work is to provide metrics for future decision-making, with the aim of making the environment more inclusive and with gender equality.*

Resumo. *Os cursos da área de Tecnologia, são majoritariamente ocupados por homens. Com o intuito de fornecer métricas para a coordenação dos cursos e comunidades que trabalham com mulheres na área da tecnologia, uma análise descritiva estatística dos dados do Censo de Educação Superior dos anos de 2009 a 2018, se faz necessário. O presente trabalho, visa verificar o número de mulheres em relação ao número de homens que ingressaram nos cursos de TI, bem como uma análise de regressão para verificar a equação da taxa de crescimento de homens e mulheres nos cursos. O escopo do presente trabalho é fornecer métricas para futuras tomadas de decisão, com o objetivo de tornar o ambiente mais inclusivo e com igualdade de gênero.*

1. Introdução

As mulheres tiveram papéis importantes no desenvolvimento da Tecnologia da Informação (TI). Ada Lovelace, por exemplo, foi uma pioneira, escrevendo o primeiro algoritmo a ser processado por uma máquina. Até meados de 1970, os cursos de ciências da computação do Instituto de Matemática da Universidade de São Paulo (IME-USP), eram compostos majoritariamente por mulheres, porém, entre 1970 e 1980, esse quadro sofreu uma inversão, que está relacionada com o fato dos computadores pessoais (PC – Personal Computer) se tornaram mais populares [Santos 2018].

De acordo com [Amaral et al. 2017] não é possível excluir o desenvolvimento tecnológico, dos valores sociais, o contrário também é válido. Há consequências, ao criar estereótipos de gênero. Por exemplo, os PC's começaram a ser utilizados para jogos lúdicos e a partir desse momento, ele começou a ser utilizado pelos meninos, o que acabou afastando as meninas da interação com o computador e consequentemente, nas possíveis escolhas futuras em suas carreiras.

Antes dos PC's, havia grandes máquinas para processamento de dados e realização de cálculos, o que era ligado ao secretariado. Os cursos da área de computação também eram muito ligados à licenciatura (matemática, estatística), o que atraíam mais mulheres para a área [Santos 2018].

[Louzada et al. 2014], mostra que questões como a influência de mulheres na história da computação e seu baixo reconhecimento e representatividade, estão interligadas e o estudo sobre uma questão sempre implicará em estudar outra. Não há apenas um fator determinante, mas vários que surgiram desde o advento da computação.

Como as mulheres ainda têm pouca participação no mercado de tecnologia, existem várias iniciativas independentes, que tem o intuito de incentivar, empoderar e apoiar mulheres na área de TI. No repositório do Github [git 2018], é possível visualizar uma lista com informações sobre essas iniciativas.

Mas, apesar de todo o esforço das comunidades, é necessário que haja um engajamento maior na coordenação dos cursos para que o ambiente na área de TI se torne mais inclusivo. De acordo com [Moreira et al. 2018], o cenário futuro no curso de Ciências da Computação da Universidade Federal da Paraíba (UFPB), segue na contramão dos Objetivos de Desenvolvimento Sustentável, no requisito de equidade de gênero da ONU (Organização das Nações Unidas), pois seu estudo mostra uma prospecção da presença feminina e constatam que o número de mulheres na área reduz-se a uma taxa de 0,4% ao ano e que, sem intervenções para reverter essa tendência, não haverá mais mulheres em 2050.

O escopo do presente trabalho é realizar uma análise descritiva dos dados do Censo da Educação Superior dos anos de 2009 a 2018, para criar métricas de apoio às coordenações e também para as comunidades que incentivam as mulheres, bem como validar o estudo com o uso de regressão, para traçar a tendência da taxa de crescimento da inserção dos alunos de TI.

1.1. Gênero

[Louro 1997] cita numa nota de rodapé:

No Dicionário do Aurélio [Ferreira 1997], aparece uma série de definições para gênero, desde seus significados no âmbito da Lógica, da Biologia, da Gramática, até usos mais correntes como "maneira, modo, estilo", "classe ou natureza do assunto abordado por um artista" ou, ainda, expressões classificadas como gíria: "fazer gênero" ("fingir ser o que não é") e "não fazer o gênero de" ("não estar conforme a opinião ou gosto de 'alguém; não agradar a'").

A partir de 2010, ao procurar o termo "gênero" no dicionário [Ferreira 2010], temos as mesmas definições de 1994, porém uma nova definição é acrescentada: "*Antrop.* a forma como se manifesta social e culturalmente, a identidade sexual dos indivíduos".

A revisão da edição de [Ferreira 2010] vai ao encontro do que Simone de Beauvoir, no seu Livro Segundo Sexo [De Beauvoir 1967]:

Ninguém nasce mulher: torna-se mulher. Nenhum destino biológico, psíquico, econômico define a forma que a fêmea

humana assume no seio da sociedade; é o conjunto da civilização que elabora esse produto intermediário entre o macho e o castrado que qualificam de feminino.

Esta definição, porém, é atual, devido aos questionamentos que vem surgindo na modernidade sobre qual é o papel da mulher e qual o papel do homem na sociedade. Mas desde 1967, a filósofa francesa Simone de Beauvoir, já trabalhava na desconstrução dessa ideia de natural e buscava a igualdade entre os gêneros, demonstrando que mulheres e homens devem ter os mesmos direitos e acesso aos mesmos ambientes.

1.2. Desigualdade de gênero na área de Tecnologia da Informação (TI)

Por muito tempo a diferença entre os gêneros serviu para justificar as mais variadas distinções entre mulheres e homens. Atributos físicos, biológicos, psíquicos e comportamentais, foram utilizados para indicar diferentes habilidades sociais, talentos, possibilidades e os destinos de cada gênero, [Louro 1997].

[de Souza 2017] apud [Abbate 2010]; [Hayes 2010], mostra que entre a década de 1960 e 1970 houve um crescimento dos cursos na área da computação e a crescente participação das mulheres. Para as autoras, o fato de o início da computação ter incorporado um número significativo de mulheres, se deve a esta área ser muito nova e ainda não estar associada a um gênero específico, não possuindo imagens, estereótipos ou identidades públicas. Com a consolidação da área e a visibilidade pública tanto econômica e social da computação a partir de 1980, iniciou-se a construção de estereótipos e identidades que associavam a computação ao universo masculino e a cultura geek. O período de crescimento desses estereótipos e identidades públicas, 1980, coincide com o momento que iniciou a redução da participação das mulheres nos cursos de graduação e no mercado de trabalho, tendência que permanece até os dias atuais.

Houve uma redução da participação das mulheres nos cursos relativos à área de tecnologia da informação no Brasil, sobretudo nos anos mais recentes. [Maia 2016] ao realizar uma série histórica do curso com o maior número de concluintes na área de tecnologia da informação, ciências da computação, entre os anos de 1991 e 2013, constata não apenas a baixa participação das mulheres em todo o período analisado, mas acentuada sub-representação de mulheres nos períodos mais recentes.

O campo do trabalho a partir da categoria gênero no Brasil revela um cenário caracterizado pela permanência de desigualdades entre homens e mulheres no mercado de trabalho, evidenciada na remuneração, no acesso a cargos de prestígio e a determinadas ocupações [de Souza 2017].

[Silva et al. 2019], cita uma pesquisa realizada pela empresa de consultoria Yocotoo, 2019, uma empresa responsável por recrutamento de profissionais na área de TI, aponta que:

82,8% das mulheres entrevistadas relata ter vivido, ou ainda vivenciar, preconceito de gênero dentro do seu ambiente de trabalho. Em relação à área acadêmica, 61,8% assegura ter vivido ou vivência este preconceito. O estudo realizado pela empresa de consultoria é bastante revelador, trazendo outros importantes números: 91% das entrevistadas afirmam que ainda existe

preconceito dentro das empresas e que essas ainda estão dando os primeiros passos direcionados à implementação de políticas de diversidade e inclusão; 72% afirmam que o ambiente familiar não costuma estimular meninas a gostarem de brincadeiras ou carreiras ligadas à tecnologia; 42% das participantes afirmam que o maior desafio é ter de provar a todo tempo que são profissionais competentes; e 36% diz ser urgente ações que incentivem o interesse pela tecnologia começando pela infância.

2. Metodologia

Para atingir os objetivos dessa pesquisa, optou-se pela análise descritiva dos dados do Censo Superior – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), seguindo a metodologia Knowledge Discovery in Databases (KDD) para identificação de padrões nos dados dos alunos, Figura 1.



Figura 1. Etapas da metodologia proposta. Adaptado [Fayyad et al. 1996]

2.1. Dados e Seleção

Na fase da seleção, os dados brutos dos anos de 2009 a 2018, foram baixados do site do INEP [ine 2020], foi escolhido os arquivos que continham os dados dos alunos, dos cursos e o dicionário das variáveis com as explicações sobre o significado e conteúdo das colunas dos arquivos baixados. Os dados dos anos anteriores de 2009 não continha as informações sobre gênero de uma forma estruturada como os dados posteriores e até o momento da execução da análise havia apenas pesquisas até 2018.

Ao analisar o dicionário, é possível verificar que o arquivo possui cerca de 105 colunas (o número de colunas altera dependendo do ano), com diversas informações sobre processo seletivo, qualidade da escola ingressada, necessidades específicas do aluno, entre outras. Como o objetivo é verificar o número de alunos por curso, seu respectivo gênero e idade, as colunas escolhidas para a análise anual foram: “TP_SEXO” (que indica o sexo dos alunos), “NU_IDADE” (indica a idade dos alunos) e “CO_CURSO” (indica o curso dos alunos).

O nome do curso se encontra em um arquivo diferente. Novamente foi observado o dicionário de variáveis, e o arquivo com as informações sobre os cursos, contém cerca de 110 colunas, porém, para construção de uma biblioteca com o código do curso e seu respectivo nome, apenas as colunas escolhidas foram: “CO_CURSO” e “NO_CURSO” (indica o nome do curso).

2.2. Pré-Processamento e Formatação

Para fazer limpeza e transformação de dados, análise estatística e visualização de dados foi utilizado a linguagem de programação Python 3, as bibliotecas de suporte Pandas, Numpy, Matplotlib, Seaborn, Statsmodels e o Jupyter Notebook, que é um aplicativo da web de código aberto que permite criar e compartilhar documentos que contêm código ativo, equações, visualizações e textos narrativos.

Como o volume de dados de 10 anos da educação superior é muito grande, para sua análise é exigido um grande poder computacional. Foi adotado a estratégia de fazer a limpeza dos dados primeiramente anual e seus dados exportados para arquivos Comma-Separated-Values (.csv), e depois foi realizado a concatenação dos arquivos em um só dataframe em um novo notebook.

No dataframe com todos os arquivos anuais concatenados, o arquivo com as informações sobre os cursos, foi importada e em seguida uma biblioteca para padronização da escrita para relacionar o código com o nome do curso, foi criada. Foi aplicado um filtro com os termos: “Computação”, “Tecnologia da Informação”, “Informática” e “Análise de Desenvolvimento de Sistemas”, para a escolha dos cursos, a ser analisados.

2.3. Mineração e Análise dos Dados

Após isso, foi realizado a fase de mineração e análise dos dados, com a contagem de alunos por curso, com as medidas de tendência central, cruzamento dos dados divididos por gênero e por idade, assim como a porcentagem de mulheres e homens nos cursos. Por fim, foi realizado o cruzamento dos dados por gênero e por ano, para verificar o comportamento da inserção das mulheres na área ao longo dos anos analisados, para verificar o comportamento da inserção das mulheres na área ao longo dos anos analisados. De modo a verificar a relação dos dados e sua causalidade, a correlação e a covariância foi realizada.

Com esses dados tabelados, para construir o gráfico de dispersão e com o apoio da biblioteca Statmodels, gerando os gráficos de regressão da taxa de crescimento dos homens e das mulheres nos cursos de TI, nos anos de 2009 a 2018. Para validar o modelo, foi utilizada a biblioteca sklearn.metrics, para o cálculo dos erros médios.

2.4. Interpretação/Avaliação e Conhecimento

A fase de interpretação e Avaliação pode ser observada na Sessão 3 – Resultados, enquanto a etapa conhecimento será abordada na conclusão.

A análise empreendida nesse artigo encontra-se no repositório do Github [git 2020].

Na Figura 2, é possível verificar o fluxo de tratamento, transformação, análise, interpretação e avaliação.

3. Resultados

O objetivo do trabalho é fazer uma análise descritiva dos dados disponíveis pelo Censo de Educação Superior, o dataframe final conta com os dados de 1.217.117 alunos, sendo

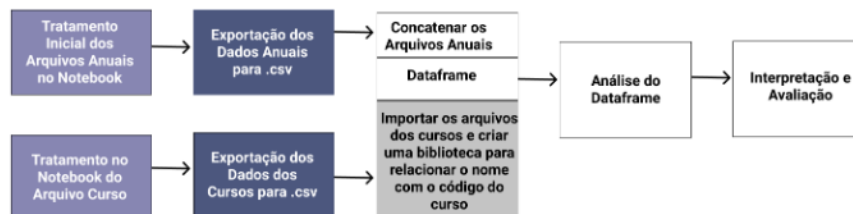


Figura 2. Fluxo de implementação

a idade mínima de 13 anos e a máxima de 84 anos. Apesar dessas idades serem discrepantes, os dados não foram retirados da análise, pois é um número muito pequeno não atrapalhando o resultado (cerca de 10 alunos). Ao analisar as medidas de tendências centrais. Pode-se observar, que em relação à idade a diferença é bem pequena entre os homens e mulheres, Tabela 1.

Tabela 1. Medidas de Tendência Centrais da idade dos alunos.

Sexo	Média	Mediana	Moda
Masculino	27 anos	25 anos	21 anos
Feminino	26 anos	25 anos	20 anos

O histograma da idade dos alunos foi plotado para verificar a distribuição das idades nos cursos, ao longo dos anos (Figura 3). É possível verificar que o volume de dados está mais concentrado na região que compreende as idades de 20 a 30 anos (75%). O gráfico é assimétrico, pois a distribuição tem uma cauda e média à direita. Pode-se observar que a sua curtose é leptocúrtica, ou seja, a curva da função de distribuição é mais afunilada com um pico mais alto do que uma distribuição normal.



Figura 3. Distribuição da idade dos alunos ao longo dos anos.

Ao se fazer o recorte de gênero, é possível verificar que 85% dos alunos são homens e 15% mulheres. Houve uma redução de 5% no número de mulheres ao longo dos 10 anos, enquanto houve um aumento de cerca de 5% no número de homens. Na Tabela 2, pode-se observar ainda, que a porcentagem de mulheres ao longo dos anos só decai, enquanto dos homens segue aumentando.

Tabela 2. Número de alunos por sexo e sua porcentagem.

Ano/Sexo	Feminino	Masculino	Feminino(%)	Masculino(%)
2009	10164	40309	20.14	79.86
2010	13887	58140	19.28	80.72
2011	16777	73014	18.68	81.32
2012	19052	86917	17.98	82.02
2013	20876	98899	17.43	82.57
2014	23483	115072	16.95	83.05
2015	23559	121969	16.19	83.81
2016	23639	129352	15.45	84.55
2017	24430	136797	15.15	84.85
2018	27247	153534	15.07	84.93

A Figura 4, é possível observar, os mesmos dados da Tabela2, porém agora em forma de gráfico para facilitar a observação da inserção das mulheres em relação aos homens nos cursos de TI entre os anos de 2009 a 2018.

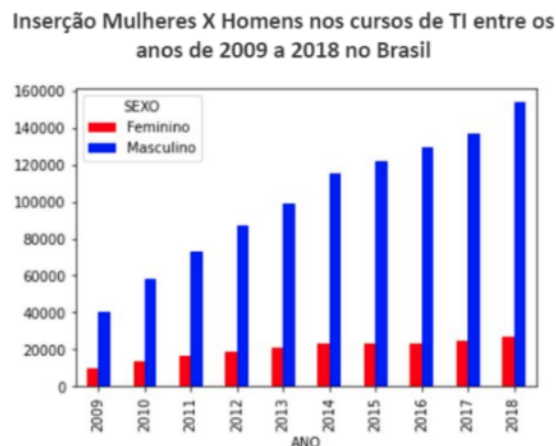


Figura 4. Inserção das mulheres nos Cursos de TI.

A correlação e a covariância em relação à taxa de porcentagem do crescimento da inserção por gênero ao longo dos anos, encontra-se na Tabela 3.

Tabela 3. Correlação e Covariância.

Ano/Gênero	Feminino	Masculino
Correlação	-0.979503	0.979503
Covariância	-5.595556	5.595556

Existe uma forte correlação entre os dados, positiva em relação ao gênero masculino, mostrando que as variáveis se movimentam na mesma direção e proporção. E negativa em relação ao gênero feminino, indicando que as variáveis são inversamente correlacionadas, se movimentam em direções opostas, mas na mesma proporção. O mesmo ocorre com a covariância, positiva para o gênero masculino, indicando que as duas variáveis tendem a variar na mesma direção, indicando que, com o passar dos anos,

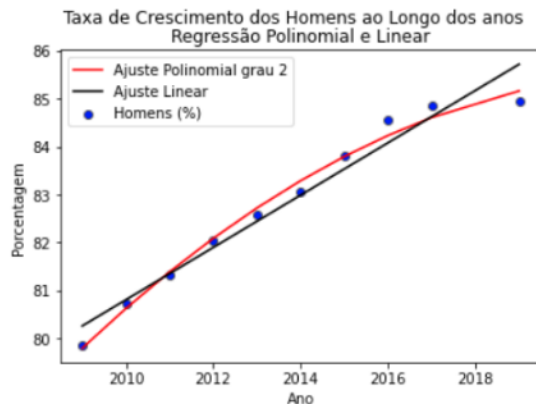


Figura 5. Taxa de Crescimento dos homens nos cursos de TI.

o número de alunos tende a aumentar. Enquanto para o gênero feminino, duas variáveis tendem a variar em direções opostas, indicando que, com o passar dos anos, o número de alunas tende a diminuir.

Para avaliar a tendência da taxa de crescimento, os dados de porcentagem da taxa de crescimento foram utilizados para a geração de um gráfico de dispersão. Ajuste linear e polinomial foi realizado para a verificação do melhor ajuste. Pode-se observar que a taxa de crescimento dos homens na área segue a tendência de uma regressão polinomial de ordem 2, indicando aumento na porcentagem de homens inseridos nos cursos de TI ao longo dos anos, como visto na Figura 5.

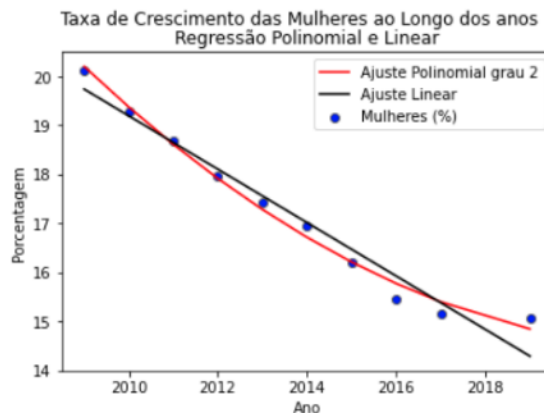


Figura 6. Taxa de Crescimento das mulheres nos cursos de TI.

O mesmo foi realizado para verificar a taxa de crescimento das mulheres e novamente o ajuste polinomial de grau 2 negativo foi o que mais se adequou aos dados reais. Ou seja, a porcentagem de mulheres nos cursos tem diminuído conforme passam-se os anos, como observado na Figura 6.

Os valores de r^2 , também conhecido como coeficiente de determinação, indica o poder preditivo da reta de regressão. Em ambas regressões linear e polinomial foi de 0,959 e 0,989, respectivamente. Nesse caso afirma-se que 96% e 99%, respectivamente das variações do ano podem ser explicadas com base na variação do número de alunos inseridos nos cursos. Determina que ambas as regressões apresentam ótimo modelo para

predizer a taxa de crescimento dos homens e das mulheres nos anos de 2009 a 2018.

4. Conclusão

A análise empreendida aqui busca analisar a participação das mulheres nos cursos da tecnologia da informação a partir da categoria gênero, realizando para isso uma análise descritiva dos dados. Tal análise evidencia o comportamento e o fluxo da inserção dos alunos ao longo dos anos. Observa-se que o número de mulheres nos cursos da área de TI está diminuindo. [Moreira et al. 2018] teve resultados similares com as análises do curso de Ciências da Computação na UFPB e de [Santos 2018] no IME-USP.

As administrações educacionais devem usar esses dados com a intenção de criar métodos para evitar a redução da procura dos cursos por mulheres e torná-los mais atraentes e inclusivos para o público feminino. Na tomada dessas decisões é necessário criar indicadores e destacar os pontos relevantes para a coordenação do curso.

Para trabalhos futuros, recomenda-se que se utilizem fatores relacionados ao tema, para estabelecer um modelo preditivo com maior consistência e maior precisão, verificando assim a relação da diminuição da inserção das mulheres nos cursos de TI a outros fatores relacionados à área.

Referências

- (2018). Repositório do github. <https://github.com/apdaros/tech-gender-gap>. Accessed: 2020-10-20.
- (2020). Microdados do censo da educação superior. <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>. Accessed: 2020-08-20.
- (2020). Repositório do github. <https://github.com/arcursino/tcc>. Accessed: 2020-12-20.
- Abbate, J. (2010). The pleasure paradox: Bridging the gap between popular images of computing and women's historical experiences. *Gender codes: Why women are leaving computing*, pages 213–227.
- Amaral, M. A., Emer, M. C. F. P., Bim, S. A., Setti, M. G., and Gonçalves, M. M. (2017). Investigando questões de gênero em um curso da área de computação. *Revista Estudos Feministas*, 25(2):857–874.
- De Beauvoir, S. (1967). *O Segundo Sexo, parte II – A experiência vivida*. Difusão Européia do Livro.
- de Souza, T. P. (2017). A desigualdade de gênero no campo da tecnologia da informação.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.
- Ferreira, A. B. d. H. (1997). Novo dicionário aurélio da língua portuguesa. In Louro, G. L., editor, *Gênero, sexualidade e educação*. Petrópolis: vozes.
- Ferreira, A. B. d. H. (2010). *Mini Aurélio: o dicionário de língua portuguesa / Aurélio Buarque de Holanda*. Positivo.

- Hayes, C. C. (2010). Computer science. *Gender codes: Why women are leaving computing*, pages 25–49.
- Louro, G. L. (1997). *Gênero, sexualidade e educação*. Petrópolis: vozes.
- Louzada, C. S., Gomes, W. F., Nunes, M., Salgueiro, E. M., Andrade, B. T., and Lima, P. (2014). Um mapeamento das publicações sobre o ingresso das mulheres na computação. In *CLEI 2014: Conferência Latino-americana em Informática-VI Congresso da Mulher Latino-americana na Computação*. Montevideu, page 16.
- Maia, M. M. (2016). Limites de gênero e presença feminina nos cursos superiores brasileiros do campo da computação. *cadernos pagu*, (46):223–244.
- Moreira, J. A., Silva, R. M., and Carvalho, M. E. P. (2018). Cenários prospectivos: Uma visão do futuro da presença feminina em cursos de ciência da computação de uma instituição de ensino superior. In *Anais do XXVI Workshop sobre Educação em Computação*. SBC.
- Santos, C. M. (2018). Por que as mulheres “desapareceram” dos cursos de computação. *Journal da Universidade de São Paulo*.
- Silva, J., Oliveira, L., and Silva, A. (2019). Meninas na computação: uma análise inicial da participação das mulheres nos cursos de sistemas de informação do estado de alagoas. In *Anais do XXVII Workshop sobre Educação em Computação*, pages 444–452. SBC.

