ISSN 2318-535X

# COMPARAÇÃO DE ALGORITMOS DE REGRESSÃO LOGÍSTICA E ÁRVORES DE DECISÃO PARA A CLASSIFICAÇÃO DA COBERTURA DO SOLO URBANO A PARTIR DE IMAGEM WORLDVIEW-3

**Mateus Henrique Pereira Rosseto[1], Paulo Roberto da Silva Ruiz[2], Cláudia Maria de Almeida[3]**
[1]Graduando em Tecnologia em Ciência de Dados na Faculdade de Tecnologia - FATEC Adamantina, mateus.rosseto@fatec.sp.gov.br
[2]Professor na Faculdade de Tecnologia - FATEC Adamantina, paulo.ruiz2@fatec.sp.gov.br
[3]Pesquisadora na Divisão de Observação da Terra e Geomática – Instituto Nacional de Pesquisas Espaciais – INPE, claudia.almeida@inpe.br

## RESUMO

O presente trabalho tem como objetivo explorar as limitações e potencialidades do sensor WorldView-3 (WV-3) com resolução espacial de 30 centímetros e 16 bandas espectrais para mapeamentos da cobertura do solo. Os testes foram realizados na tarefa de classificação de alvos urbanos em um nível de legenda com 11 classes de cobertura. Para isso, foram utilizados diferentes algoritmos de aprendizado de máquina, sendo eles: o *SimpleLogistic*, que é baseado em regressão logística simples e o método *Hoeffding Tree*, o qual é baseado em árvores de decisão. O modelo de classificação gerado pelo *SimpleLogistic* obteve o maior acerto global, chegando a 76.49%, já o *Hoeffding Tree* obteve acurácia de 72.32%. Os modelos obtiveram acurácias classificadas como muito boas, segundo os coeficientes *Kappa* os resultados obtidos foram respectivamente de 0.7149 e 0.6720. Dessa forma, os resultados encorajam o uso destes algoritmos para a classificação da cobertura do solo urbano a partir de imagens WorldView-3.

**Palavras-chave:** Classificação. Regressão. Árvores de decisão. Aprendizado de Máquina.

## 1 INTRODUÇÃO

Por possuir o maior contingente populacional, as cidades necessitam de melhorias constantes em seu planejamento para atender as necessidades de seus cidadãos. Qualidade de vida, acesso a bens públicos, áreas de lazer, conforto térmico e ambiental impactam na qualidade de vida, além das questões econômicas relacionadas à distribuição equitativa de bens e serviços. Atualmente, a inclusão socioeconômica da população urbana torna-se necessária devido ao seu rápido crescimento, verificando-se a superação do índice de população rural mundialmente. Em países subdesenvolvidos, esse crescimento ocorre de forma acelerada e desordenada, sendo acompanhado pela ocupação irregular em loteamentos clandestinos e favelização, reflexo do avanço da miséria e desigualdade social nos centros urbanos periféricos (SANTOS, 2020).

As administrações públicas municipais necessitam de dados diversificados para desenvolver projetos de urbanização. Por exemplo, planos de arborização em áreas com deficiência de árvores em relação à sua população, criação de áreas de lazer em terrenos públicos ociosos, implantação de novos bairros em áreas de vazios urbanos, dentre outros (DENALDI; FERRARA, 2018). Nesse sentido, o sensoriamento remoto aliado à mineração de dados, ganha destaque por possibilitar o mapeamento e a atualização cartográfica do ambiente urbano. A melhoria sistemática da qualidade geométrica dos sensores orbitais de alta resolução espacial e o imageamento em vários canais multiespectrais permitem detalhar os alvos urbanos. Esses recursos são requeridos para o mapeamento urbano, pelo fato de as cidades possuírem uma grande diversidade de alvos, geralmente de pequeno e médio porte (JENSEN, 2011). Além disso, os algoritmos de mineração de dados permitem classificar de maneira automática os alvos urbanos com alta acurácia, possibilitando diversas aplicações (RUIZ, 2017).

Assim, o presente trabalho objetiva explorar as potencialidades e limitações do sensor WorldView-3 (WV-3) para a classificação de alvos urbanos em um nível de legenda com 11 classes de cobertura do solo. Para isso, foram utilizados comparativamente diferentes métodos de classificação de imagens baseados em regressão logística simples e árvores de decisão. A área de estudo localiza-se no interior de São Paulo, em um setor do campus da Universidade Estadual de Campinas (UNICAMP), a qual contém grande diversidade de materiais de cobertura do solo urbano.

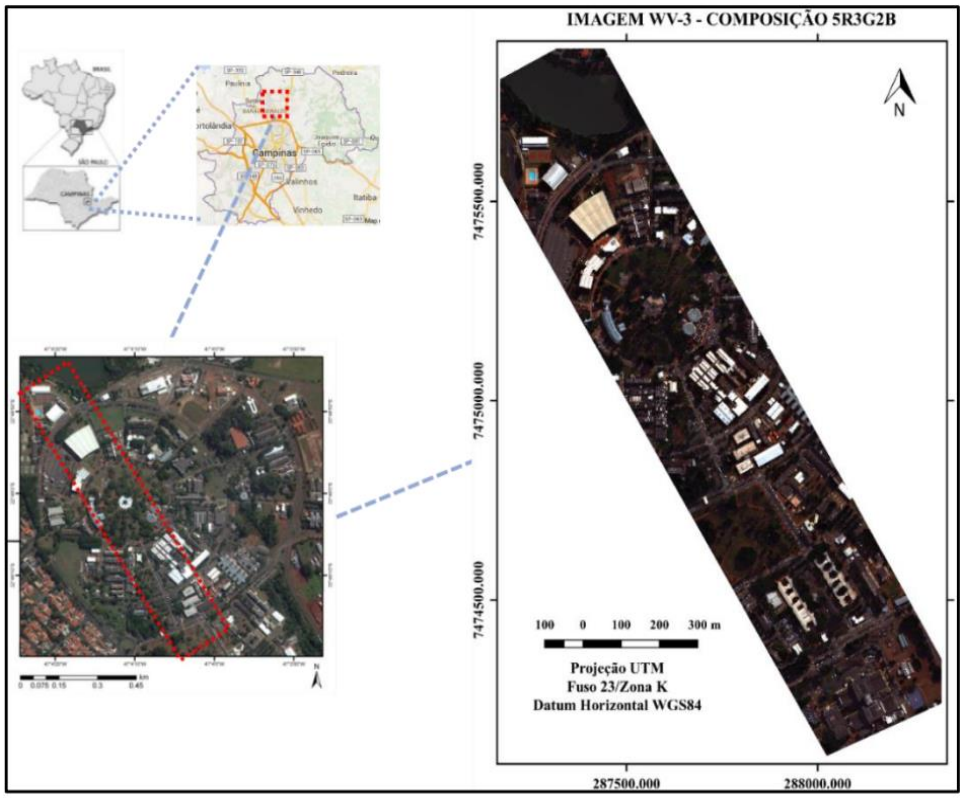
## 2 MATERIAL E MÉTODOS

A área de estudo localiza-se na cidade de Campinas – SP, a qual possui coordenada central de 22°54’3’’S e 47°3’26’’W com altitude média de 685 metros (Figura 1). Trata-se de um transecto do campus da Universidade Estadual de Campinas (UNICAMP), sendo escolhida por possuir uma grande diversidade de alvos urbanos. O trabalho foi realizado a partir de uma imagem orbital WV-3, que possui resolução espacial de 30 centímetros e 16 bandas espectrais. A imagem foi obtida em 24 de julho de 2015, com angulação de 6,52° *off nadir*, 40,6° de elevação solar e 0% de cobertura de nuvens.

A partir da imagem original, inicialmente é realizado seu pré-processamento, consistindo na conversão dos dados de nível de cinza para radiância, a qual consiste em uma transformação física que obtém o valor da incidência da radiação solar refletida pelos objetos para o espaço, os quais foram captados pelos sensores do satélite. A seguir é

realizada a correção atmosférica, que considera o quanto a contaminação da atmosfera, com gases e partículas, afetou os dados captados pelos sensores. Dessa forma, o dado é transformado para a grandeza física de reflectância de superfície. Em seguida é realizada a fusão de bandas, a qual combina a melhor resolução espacial da banda pancromática com as bandas multiespectrais para sintetizar uma nova imagem multiespectral com melhor resolução espacial (JENSEN, 2011).

Figura 1: Localização da área de estudo




Fonte: Ruiz (2017).

Neste trabalho foi adotada a classificação supervisionada por regiões, para isso, após o pré-processamento a imagem foi segmentada por meio do algoritmo multirresolução (*Multiresolution Segmentation*), disponível no software eCognition Developer 8.7 (TRIMBLE, 2011). Foram adotadas classes de cobertura urbana em um nível de legenda baseado em Ruiz (2017), o qual possui as seguintes classes: coberturas cerâmicas, coberturas metálicas, lago, materiais mistos, pavimentação não viária, pavimentação viária, piscina, solo exposto, sombra, vegetação arbórea e vegetação rasteira. A seguir, procede-se à extração de amostras das classes, que consiste na obtenção de regiões na imagem segmentada representativas de cada classe de cobertura, que foram utilizadas para treinamento dos algoritmos. Os atributos das classes são extraídos das bandas espectrais da imagem e por meio de aritmética entre as bandas, denominados

Atributos Customizados (ACs). Dentre eles destacam-se os índices de vegetação, como o NDVI (*Normalized Difference Vegetation Index*) e o SAVI (*Soil Adjusted Vegetation Index*), que são responsáveis pela identificação da vegetação arbórea e rasteira (FRANCISCO et al., 2020).

Os algoritmos de regressão logística simples e indução de árvores de decisão utilizados neste trabalho estão implementados no software Weka 3.8.6 (*Waikato Environment for Knowledge Analysis*), trata-se de um conhecido software de aprendizado de máquina (*Machine Learning*) escrito em Java, desenvolvido na Universidade Waikato na Nova Zelândia, contendo uma coleção de ferramentas de visualização e diversos algoritmos para solucionar problemas que demandam mineração e predição de dados (KULKARNI; KULKARNI, 2016).

A regressão é um processo estatístico utilizado para medir a relação entre uma variável dependente e uma ou mais variáveis independentes. O algoritmo *SimpleLogistic* insere-se nesse contexto, ao treinar ou aprender as características de um conjunto de dados e em cada interação adicionar um modelo de regressão linear simples por classe no modelo de regressão logística (NICK; CAMPBELL, 2007). O algoritmo para de adicionar modelos de regressão linear quando o erro da validação cruzada se estabiliza (AKHTER et al., 2020).

Algoritmos de árvores de decisão constroem uma árvore a partir dos dados fornecidos, seus nós representam atributos ou características de um exemplo com sua importância de classificá-los nos nós terminais, chamados de folhas, as quais indicam as classes a serem identificadas. Comumente são fáceis de interpretar, mas complexo e demorado quanto maior for o conjunto de dados para treinar e construir o modelo de classificação (QUINLAN, 1986). O algoritmo *Hoeffding Tree* diferencia-se dos demais métodos de árvores de decisão por utilizar um limite denominado *Hoeffding* para calcular um certo nível de pontuação de confiança e decidir quantos exemplos são necessários para alcançar esse nível, que está ligado com a acurácia esperada do modelo (AKHTER et al., 2020).

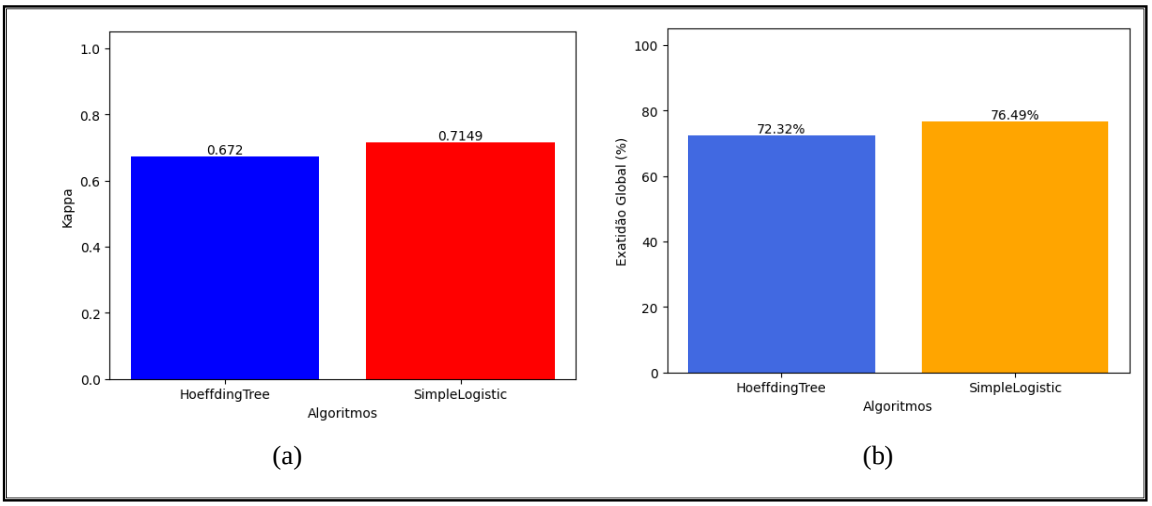
A avaliação da qualidade das classificações foi realizada por meio de comparações com o mapa de verdade de solo, a partir de Ruiz (2017). Para realizar a avaliação foram gerados pontos aleatórios nos mapas classificados e de verdade de solo. A partir deles, foi realizado o cruzamento espacial de dados para obter a classe presente no mapa de verdade e nas classificações realizadas. Foram utilizados de 2500 a 3000

pontos aleatórios, variando em cada classificação devido à exclusão do retângulo envolvente da imagem. O resultado é uma tabela de classes associando cada ponto à classe presente no mapa verdade e na classificação. De posse destas tabelas, foram geradas as matrizes de confusão das classificações. Por meio da matriz de confusão, é possível obter o cálculo da exatidão global, coeficiente *Kappa*, coeficiente *Kappa* condicional, bem como os erros de omissão e de comissão de cada classe. Os erros de omissão referem-se à relação entre o número de amostras classificadas corretamente da classe k e o número total de amostras de referência da classe k. Já os erros de comissão, correspondem à relação entre o número de amostras classificadas corretamente da classe k e o número total de amostras classificadas da classe k (RUIZ et al., 2018).

## 3 RESULTADOS E DISCUSSÕES

Para realizar as classificações foram extraídos 123 atributos da imagem WV-3. A Figura 2 apresenta os coeficientes *Kappa* (Fig. 2A) e a exatidão global (Fig. 2B) alcançados pelas classificações realizadas a partir dos dois algoritmos avaliados. É possível verificar que o algoritmo *SimpleLogistic* apresenta os melhores coeficientes *Kappa* e as maiores exatidões globais, sendo respectivamente de 0.7149 e 76.49%. Conforme a categorização de Landis e Koch (1977), os coeficientes *Kappa* alcançados pelas classificações inserem-se na categoria muito boa.

Figura 2 – Coeficiente Kappa e Exatidões Globais do nível 1 de legenda.




(a) (b)

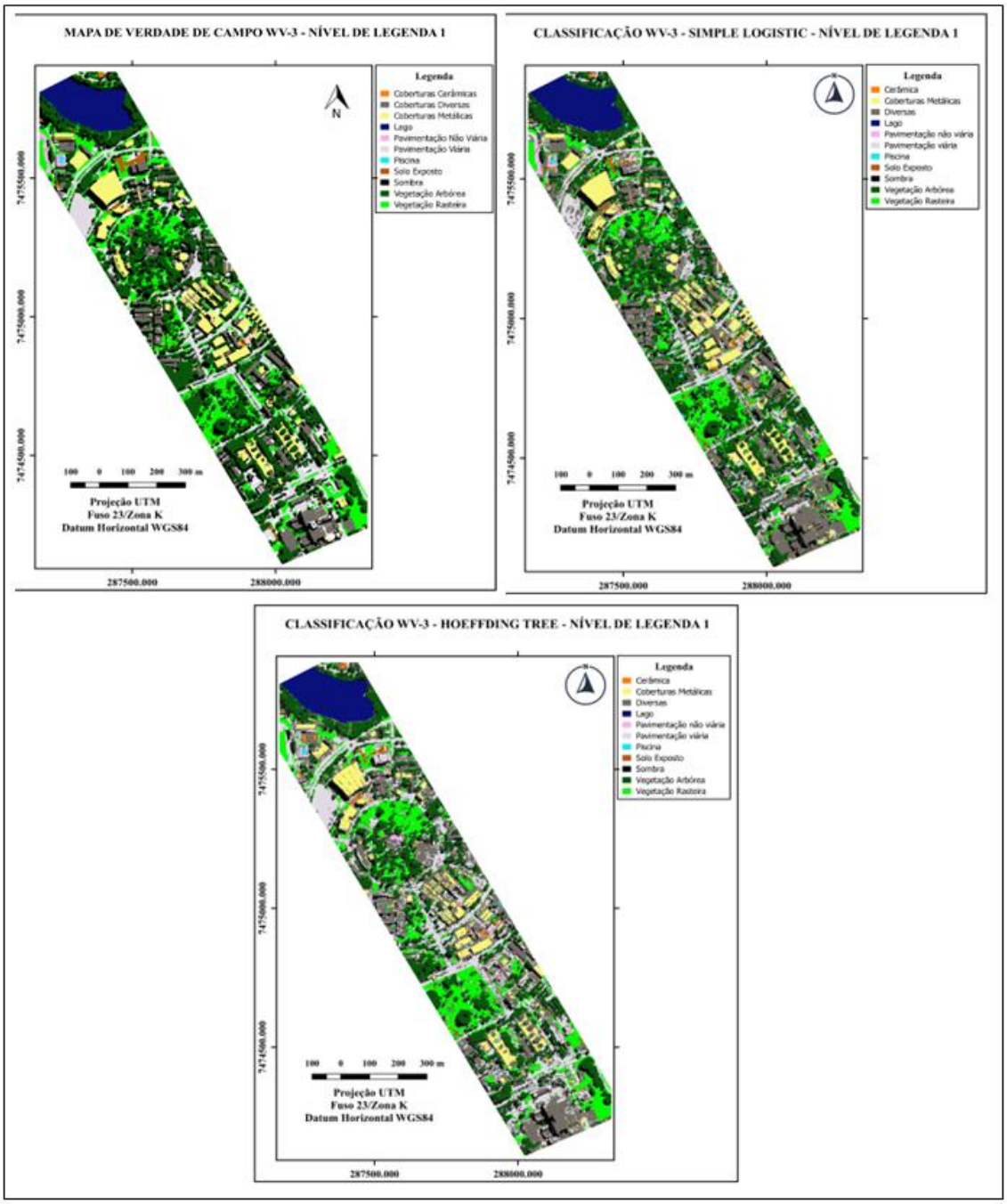
Fonte: Autoria Própria (2023)

A Figura 3 apresenta os mapas de classificação da cobertura urbana obtidos por meio de cada algoritmo utilizado neste trabalho e o mapa de verdade de campo. A partir

ISSN 2318-535X

de uma análise visual, é possível verificar que a melhor classificação foi aquela obtida pelo algoritmo *SimpleLogistic*, quando comparada com o mapa de verdade de solo.

Figura 3 – Mapa de verdade de campo e classificações da cobertura do solo




Fonte: Autoria Própria (2023)

Os algoritmos utilizados neste trabalho apresentaram tempos distintos para a construção de seus modelos de classificação. A verificação dos tempos foi feita por meio do software Weka, onde essas diferenças de tempo são referentes às características específicas de cada algoritmo em seu treinamento e na construção do modelo de

classificação. O algoritmo *Hoeffding Tree* apresentou o menor tempo, sendo de 0,14 segundos. Por sua vez, o algoritmo *SimpleLogistic* apresentou um tempo 6,3 vezes superior, sendo de 0,89 segundos.

Quando comparados os resultados das classificações por meio do teste de hipótese Z com nível de significância de 5%, verifica-se que o coeficiente *Kappa* obtido pelo *SimpleLogistic* é significativamente maior que aquele obtido pelo *Hoeffding Tree*.

## 4 CONCLUSÕES

O presente trabalho realizou duas classificações da cobertura urbana a partir de um transecto de uma imagem do satélite WV-3 utilizando dois algoritmos de aprendizado de máquina, um de regressão logística simples e outro de árvore de decisão, sendo eles: o *SimpleLogistic* e o *Hoeffding Tree*. Foi utilizado um nível de legenda com 11 classes de cobertura do solo urbano.

Os resultados revelam uma superioridade do algoritmo *SimpleLogistic*¸ o qual apresentou os melhores coeficientes *Kappa* e índice de acerto global. Apesar desta superioridade de um dos algoritmos, ambos resultados são categorizados como muito bons, quando analisados os coeficientes *Kappa*.

Com a consecução deste trabalho é possível concluir que a imagem e os algoritmos utilizados são adequados para tarefas de mapeamento de ambientes urbanos. A alta resolução espacial e espectral do sensor WV-3 permitiu diferenciar os alvos urbanos em um nível de legenda com 11 classes de cobertura do solo. Por fim, os resultados encorajam o uso dos algoritmos *SimpleLogistic* e *Hoeffding Tree* para a classificação de dados espaciais orbitais.

## 5 REFERÊNCIAS

AKHTER, P. M.; JIANGBIN, Z.; NAQVI, R. I.; ABDELMAJEED,M.; SADIQ, T. S. Automatic Detection of Offensive Language for Urdu and Roman Urdu. **IEEE Access,** v.8, p. 91213-91226, 2020.

DENALDI, R.; FERRARA, L. N. A dimensão ambiental da urbanização em favelas. **Ambiente & Sociedade**, v. 21, 2018.

FRANCISCO, C. N.; RUIZ, P. R. S.; ALMEIDA, C. M.; GRUBER, N. C.; ANJOS, C. S. Análise do impacto da correção atmosférica no cálculo do Índice de Vegetação por Diferença Normalizada a partir de Imagem Landsat 8/OLI. **Revista Brasileira de Geografia Física**, v. 13, n. 1, p. 076-086, 2020.

JENSEN, J. R. **Sensoriamento remoto do ambiente: uma perspectiva em recursos terrestres**. São José dos Campos: Parêntese, 2011. Tradução José Carlos Neves Epiphanio (coordenador).

KULKARNI, E. G.; KULKARNI, R. B. WEKA powerful tool in data mining. **International Journal of Computer Applications** (0975 – 8887). National Seminar on Recent Trends in Data Mining. 2016.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical Data. **Biometrics**, v. 33, n. 1, p.159–174, 1977. DOI 10.2307/2529310.

NICK, T. G.; CAMPBELL, K. M. Logistic regression. **Topics in biostatistics**, p. 273-301, 2007.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v.1, n. 1, p. 81 – 106, 1986.

RUIZ, P. R. S.; ALMEIDA, C. M.; LACERDA, C. S. A. Classificação da cobertura do solo urbano usando árvores de decisão a partir de uma cena WorldView-2 para diferentes níveis de legenda. **Geociências** (UNESP. Impresso), v. 37, p. 597-609, 2018.

RUIZ, P. R. S. **Classificação da cobertura do solo urbano usando árvores de decisão a partir de cenas WorldView-2 e WorldView-3 para diferentes níveis de legenda**. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2017, 181 p.

SANTOS, A. M. S. P. Política urbana no Brasil: a difícil regulação de uma urbanização periférica. **Geo UERJ**, n. 36, p. 47269, 2020.

TRIMBLE. **eCognition developer 8.7 user guide**. Munich, Germany: [s.n.], 2011. 258 p. Disponível em: http://www.ecognition.com/. Acesso em: 18 ago.2023.