Classificação Semissupervisionada do Uso e Cobertura do Solo baseada em Dados Multiespectrais no Sul do Piauí

Bruno Vicente Alves de Lima[1], **Elayne da Silva Figueredo**[2]

[1]Instituto Federal do Maranhão (IFMA)
Coelho Neto – MA – Brasil

[2]Universidade Federal do Rio Grande do Norte (UFRN)
Natal – RN – Brasil

brunovicente.lima@ifma.edu.br, elaynefigueredoo@gmail.com

Abstract. *In this article, a study was conducted to map agricultural activities in the Uruçuí region, employing semi-supervised learning techniques. Through this approach, a partially labeled dataset was utilized, enabling a more comprehensive analysis of the region. The applied models demonstrated satisfactory results, proving to be efficient in identifying and classifying the various predefined classes within the study area.*

Resumo. *Neste artigo foi realizado um trabalho para mapear atividades agrícolas na região de Uruçuí, utilizando técnicas de aprendizado semissupervisionado. Por meio desse método, um conjunto de dados parcialmente rotulados foi utilizado, o que possibilitou uma análise mais abrangente da região. Os modelos utilizados demonstraram resultados satisfatórios, mostrando-se eficientes na identificação e classificação das diversas classes definidas na área de estudo.*

1. Introdução

A integração de grandes volumes de dados orbitais em aplicações de Sensoriamento Remoto (SR) voltadas para estudos ambientais, vem ganhando espaço na literatura [Liu et al. 2018] [Adam et al. 2014][Feranec et al. 2010]. Esse grande volume de dados tem permitido a utilização de técnicas de Aprendizado de Máquina (AM) para realizar a classificação supervisionada de dados multiespectrais [Li et al. 2017] [Yao et al. 2019] [Zhu et al. 2019] [Koda et al. 2019] para obter informações do uso e cobertura do solo.

[Colditz et al. 2011] destaca a importância de se obter informações sobre a cobertura da terra como um dos principais fatores para a compreensão dos processos que ocorrem na superfície terrestre. Isso se deve ao fato da maioria dessas alterações está diretamente relacionada a aspectos ambientais, econômicos e sociais, que podem ocorrer em diferentes escalas espaciais.

As informações geoespaciais provenientes de Sistemas de Sensoriamento Remoto (SR) estão se tornando cada vez mais essenciais na tomada de decisões de políticas públicas. Essas decisões abrangem diversos setores, como a gestão ambiental e territorial, além do monitoramento ambiental, com foco especial na identificação e resposta a desastres naturais.

Através do uso de dados geoespaciais obtidos por meio do Sensoriamento Remoto, é possível aumentar a eficiência e precisão na avaliação dos impactos das mudanças no

uso e cobertura da terra. Dessa forma, torna-se viável desenvolver estratégias adequadas para proteger e conservar o meio ambiente, além de implementar medidas que contribuam para o desenvolvimento sustentável e a melhoria da qualidade de vida das comunidades afetadas.

Desta forma, neste trabalho, destaca-se aplicação do aprendizado semissupervisionado no mapeamento de atividades agrícolas na Região Geográfica Intermediária (RGI) de Uruçuí, no Sudoeste do Estado do Piauí, Brasil. A região faz parte da área chamada MATOPIBA, estratégica para expansão agrícola e preservação da biodiversidade brasileira, cujo nome é em referência a zona formada pelos estados brasileiros Maranhão (Ma), Tocantis (To), Piauí (Pi) e Bahia (Ba).

2. Área de Estudo

A área de estudo deste trabalho é a Região Geográfica Imediata (RGI) de Uruçuí, composta por sete municípios, são eles: Antônio Almeida, Baixa Grande do Ribeiro, Bertolínia, Manoel Emídio, Ribeiro Gonçalves, Sebastião Leal e Uruçuí, reconhecidos por intensa atividade agrícola. Localizados na porção noroeste da mesorregião sudoeste piauiense, com variações de 152 a 651 metros em relação ao nível do mar (Figura 1).

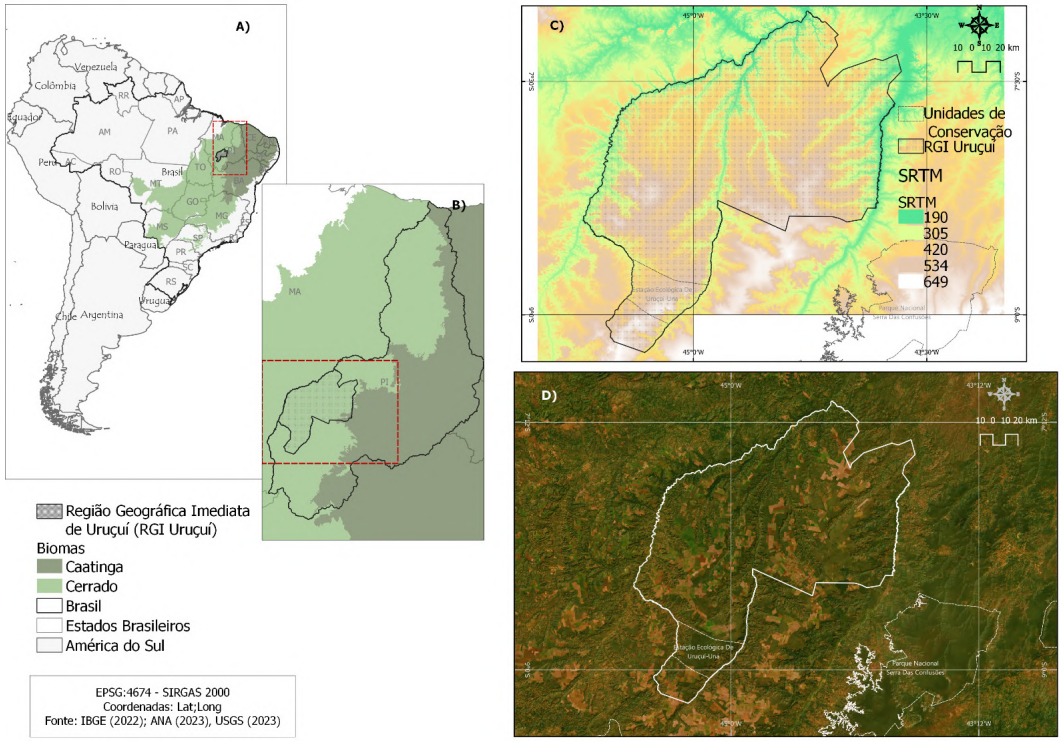


Figura 1. Área de Estudo do MATOPIBA

A Região Geográfica Imediata (RGI) de Uruçuí tem como principal referência a rede urbana que conecta os municípios de Antônio Almeida, Baixa Grande do Ribeiro, Bertolínia, Manoel Emídio, Ribeiro Gonçalves e Sebastião Leal a Uruçuí. Essa conexão visa atender às necessidades imediatas das populações, como acesso a serviços de saúde e educação, além da oferta de serviços públicos, como atendimento do Instituto Nacional do Seguro Social (INSS) e do Ministério do Trabalho, bem como serviços judiciários. A

integração dessa rede urbana tem um papel fundamental na promoção do desenvolvimento e no bem-estar das comunidades envolvidas [Aguiar and Gomes 2004] [IBGE 2017].

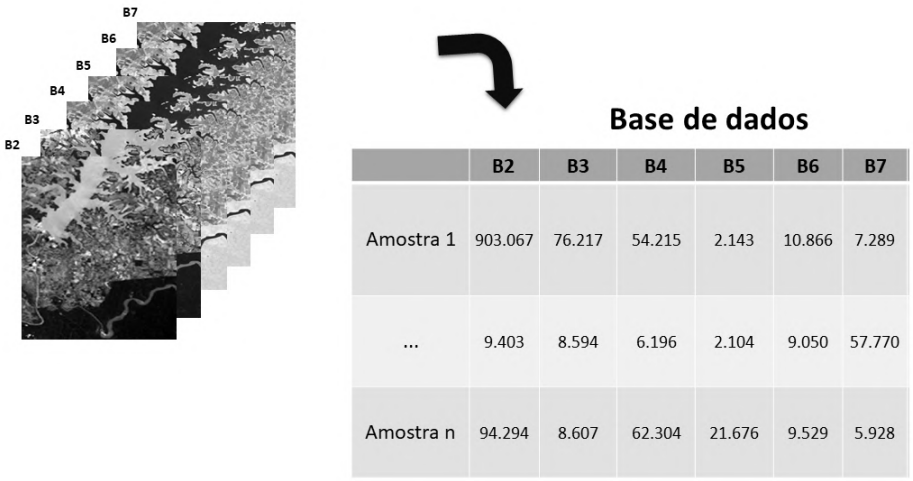
A área de estudo é composta por uma variedade de estruturas produtivas, abrangendo cultivos diversos como soja, milho, algodão, arroz, cana-de-açúcar, entre outros. Essa extensão foi dividida em 6 classes distintas, incluindo Água, Vegetação Cerrado, Atividade Agrícola, Solo Exposto, Mata Ciliar e Encosta de Serra. Com uma área total de 2.718.986,40 hectares, diante da importância da região, esta é considerada uma amostra representativa ideal para aplicação da metodologia experimental proposta.

3. Metodologia dos Experimentos

3.1. Base de Dados

Para a realização deste trabalho, tornou-se necessário a ajuda de um especialista cartográfico para coletar os dados para a metodologia dos experimentos. Para extrair os dados para formar a base de dados, utilizou-se a ferramenta *Google Earth Engine* [1].

A base de dados é composta por metadados extraídos das bandas Espectrais 2 a 7 (Figura 2, das faixas espectrais do azul, verde, vermelho, infravermelho de ondas curtas 1 e infravermelho de ondas curtas 2, respectivamente, estas em níveis de refletância calibrada para o Top of Atmosphere (TOA) - camada 1, do satélite Landsat-8, que possui resolução espacial de 30 x 30 m nestas respectivas bandas, adquiridas no período de agosto do ano 2019.



	B2	B3	B4	B5	B6	B7
Amostra 1	903.067	76.217	54.215	2.143	10.866	7.289
...	9.403	8.594	6.196	2.104	9.050	57.770
Amostra n	94.294	8.607	62.304	21.676	9.529	5.928

Figura 2. Base de dados formada pelas bandas espectrais

Foram coletadas pelo especialista um total de 11966 amostras para treinamento utilizando a Ferramenta *Google Earth Engine*. Devido a quantidade excessiva de amostra necessária, foram rotuladas apenas 1200 das amostras em 6 classes possíveis, são elas: Água, Vegetação Cerrado, Atividade Agrícola, Solo Exposto, Mata Ciliar e Encosta de Serra. Cada amostra é formada por um total de 6 características, sendo estas formadas pelos valores em cada uma das bandas espectrais (B2 à B7), como mostra na Figura 2. Ou seja, a imagem é formada por pixels, e para cada pixel existe um total de 6 valores, cada um referente à uma banda espectral.

[1] https://code.earthengine.google.com/

3.2. Metodologia dos Experimentos

A metodologia dos experimentos neste trabalho foi realizada em duas etapas: Teste dos Algoritmos e Geração dos Mapas de Classificação.

3.2.1. Testes dos Algoritmos

Como dito anteriormente, foram utilizados algoritmos de aprendizado de máquina. Como coletar as amostras é um processo oneroso para uma pessoa, e não existe uma base de dados totalmente rotulada sobre a região de estudo, optou-se por coletar apenas 1200 amostras rotuladas, e o restante não rotulado.

Com esse cenário, torna-se viável a utilização de algoritmos de aprendizado semissuperivionado, pois este tipo de algoritmo de AM, treinam utilizando tanto dados rotulados quanto não rotulados.

Neste trabalho utilizou-se três modelos preditivos semissupervionados amplamente utilizados na literatura, são eles *Self-training*[Yarowsky 1995], *Co-training* [Blum and Mitchell 1998] e *Label Propagation* [Zhu and Ghahramani 2002]. Os algoritmos *Self-training* e *Co-training* funcionam basicamente treinando internamente outro modelo supervisionado, assim, adotamos quatro modelos supervisionados, são eles Máquina de Vetor de Suporte (SVM), Random Forest (RF), K-Vizinhos Mais próximos (KNN) e Rede Neural Multicamadas (MLP).

Para testar os algoritmos usou-se a técnica da Validação Cruzada. Neste caso, dividiu-se apenas os dados rotulados em 10 grupos, tomando um como teste, e o restante como treino. Os dados não rotulados foram utilizados totalmente em todos os testes. Para avaliar os algoritmos, utilizou-se as métricas de avaliação Acurácia, Precisão, *Recall* e *F1-Score*. Como foi utilizada a Validação Cruzada, calculou-se a média de 10 execuções para cada métrica citada em cada um dos algoritmos.

Para a implementação dos algoritmos e todos testes, foi utilizada a Biblioteca para a Linguagem Python, *Sklearn*[2], e o ambiente de programação chamado *Google Colaboratory*. Os hiperparâmetros dos algoritmos foram definidos de forma empírica, mostrados na Tabela 1.

Tabela 1. Hiperparâmetros utilizados nos algoritmos de classificação

KNN	Quantidade de vizinhos = 10
Random Forest	Quantidade de árvores = 100
SVM	Kernel = rbf
MLP	1 Camada oculta com 10 Neurônios Função de Ativação Relu nas camadas ocultas e Softmax na saída Taxa de aprendizado 0.1 Otmizador SGD

[2]https://scikit-learn.org/stable/index.html

3.2.2. Geração dos Mapas de Classificação

Quando trabalhamos com aprendizado de máquina no contexto de Sensoriamento Remoto, além de saber se os algoritmos estão com valores satisfatórios em termos de acurácia, torna-se necessário a visualização dos mapas de classificação.

Estes mapas contribuem para a interpretação visual da classificação realizada pelos modelos preditivos, mostrando onde na região estão cada uma das classes definidas no problema. Ou seja, pode-se observar visualmente se o modelo está realmente detectando corretamente todas as classes do problema.

Além disso, o mapa de classificação é um dos produtos utilizados por gestores e entidades públicas para tomar decisões estratégicas acerca do problema em questão. Pra gerar esse mapa de classificação, foram coletadas todas as amostras que formam a área de estudo e submetidas à classificação por parte dos algoritmos de aprendizado de máquina, que foram treinados anteriormente na fase de Testes dos Algoritmos. Desta foram podemos obter o mapa de classificação para cada um dos modelos.

4. Resultados

Esta sessão apresenta os resultados obtidos nos experimentos.

4.1. Resultado da Avaliação dos Algoritmos

A Tabela 2 apresenta os resultados dos algoritmos na Fase de Avaliação. Podemos ver na Tabela 2, os valores de Acurácia, Precisão, *Recall* e *F1-Score* pra cada um dos algoritmos escolhidos.

Tabela 2. Resultados dos Testes dos Algoritmos Semissupervisionados

MODELO	**Acurácia**	**Precisão**	***Recall***	***F1-Score***
Self-training (KNN)	0,948	0,952	0,952	0,947
Self-training (MLP)	0,933	0,942	0,942	0,934
Self-training (RF)	**0,961**	0,963	0,963	0,961
Self-training (SVM)	0,940	0,947	0,947	0,940
Co-training(KNN)	0,930	0,935	0,935	0,929
Co-training(MLP)	0,922	0,932	0,932	0,922
Co-training(RF)	**0,940**	0,946	0,946	0,940
Co-training(SVM)	0,928	0,937	0,937	0,928
Label Propagation	**0,955**	0,958	0,958	0,955

Pode-se observar na Tabela 2, que o algortimo *Self-training* obteve uma melhor acurácia quando executado com o algoritmo Random Forest. Da mesma forma, o algoritmo *Co-training* obteve melhor acurácia com o Random Forest.

Em relação aos três modelos semissupervisionados, o que mais se destacou em relação à acurácia foi o *Self-training*, com 0,961 de acurácia, enquanto o *Co-training* e o *Label Propagation* obteve acurácia 0,940 e 0,955 respectivamente.

Observando também os valores de Precisão, nota-se que todas as classes estão com boa taxa de classificação correta. Assim como também, podemos observar nos valores de

recall, ambos possuem valores acima de 0.930. A taxa de f1-score, que é uma relação da precisão e *recall*, mostra com valores acima de 0,920, que o resultado dos algoritmos são satisfatórios, ou seja, os modelos conseguiram generalizar o problema e obter bons resultados.

4.2. Mapas de Classificação

Após o treino e teste dos algoritmos semissupervisionados, selecionou-se três modelos para gerar os mapas de classificação. A Figura 3 apresenta a imagem real da região de estudo. Ao lado é possível observar uma região separada que foi ampliada para melhor visualização. Logo abaixo apresenta os mapas de classificação gerados pelo Algoritmo *Self-training* e *Co-training*, ambos utilizando o classificador Random Forest, e o algoritmo *Label Propagation*.

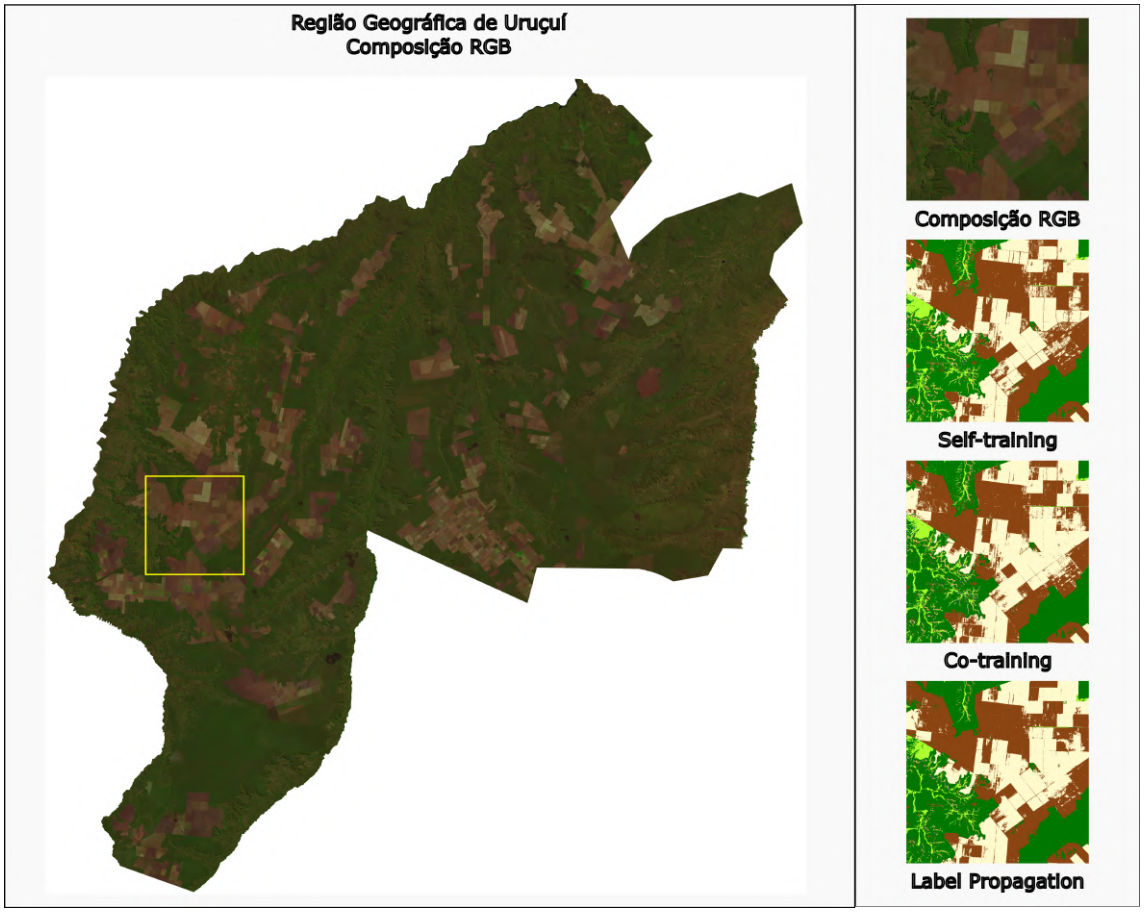


Figura 3. Nessa Figura pode-se visualizar o mapa de classificação de uma região específica da área de estudos.

Na Figura 3, pode-se observar que o modelo foi capaz de detectar na imagem Mata Ciliar, Encosta da Cerra de forma eficiente. Entretanto, apesar da acurácia de 0.961, percebe-se que o modelo confunde-se ao classificar Solo Exposto e Atividades Agrícolas.

Na Figura 3, pode-se visualizar também o mapa de classificação da mesma região com o algoritmo *Co-training*, também utilizando o classificador *Random Forest*. Em relação ao Algoritmo *Self-training*, o resultado foi similar, com poucas alterações, sendo

o *Co-training* também eficiente em detectar Encosta da Serra de forma satisfatória. Em relação em detectar as atividades Agrícolas, os resultados foram semelhantes ao algoritmo *Self-training*.

E por fim, pode-se ver o mapa de Classificação do Algoritmo *Label Propagation*, que apesar de ter a menor acurácia, obteve um mapa de classificação simular aos outros dois modelos de classificação.

5. Conclusão

Esse trabalho, apresentou um estudo inicial de detecção de uso e ocupação do solo na Região geográfica Intermediária do município de Uruçuí, no sudoeste do Piauí. Foram testados três algoritmos de aprendizado semissupervisionado para realizar esta tarefa, são eles: Self-trainig, *Co-training* e *Label Propagation*.

Os três modelos apresentaram valores de acurácia satisfatórios, sendo o algoritmo *Self-training* utilizando o algortimo *Random Forest* apresentou melhor acurácia.

Com isso, foi possível gerar mapas de classificação de uma região da área de estudo mostrando a eficiência dos modelos de aprendizado semissupervisionado para realizar esta tarefa.

Apesar disso, no mapa de classificação percebeu-se ainda algumas falhas por parte dos modelos preditivos. Assim, como trabalho futuros, pretende-se testar outros modelos de aprendizado semissupervisionados encontrados na literatura. Pretende-se também, ampliar a quantidade de bandas espectrais, e desta foram tornar viável a utilização de *Deep Learning*. Esse enfoque promissor pode contribuir significativamente para aprimorar o monitoramento e planejamento das atividades agrícolas na região, oferecendo informações valiosas para a gestão sustentável e o desenvolvimento econômico dessa importante área agrícola.

Referências

Adam, E., Mutanga, O., Odindi, J., and Abdel-Rahman, E. M. (2014). Land-use/cover classification in a heterogeneous coastal landscape using rapideye imagery: evaluating the performance of random forest and support vector machines classifiers. *International Journal of Remote Sensing*, 35(10):3440–3458.

Aguiar, R. and Gomes, J. (2004). Projeto cadastro de fontes de abastecimento por água subterrânea, estado do piauí: diagnóstico do município de uruçuí. *Serviço Geológico do Brasil. CPRM, Fortaleza (13pp)*.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Colditz, R., Schmidt, M., Conrad, C., Hansen, M., and Dech, S. (2011). Land cover classification with coarse spatial resolution data to derive continuous and discrete maps for complex regions. *Remote Sensing of Environment*, 115(12):3264 – 3275.

Feranec, J., Jaffrain, G., Soukup, T., and Hazeu, G. (2010). Determining changes and flows in european landscapes 1990–2000 using corine land cover data. *Applied Geography*, 30(1):19 – 35.

IBGE (2017). *GEOGRÁFICAS IMEDIATAS E REGIÕES GEOGRÁFICAS INTERMEDIÁRIAS: 2017/IBGE*.

Koda, S., Melgani, F., and Nishii, R. (2019). Unsupervised spectral-spatial feature extraction with generalized autoencoder for hyperspectral imagery. *IEEE Geoscience and Remote Sensing Letters*.

Li, Y., Zhang, Y., Huang, X., Zhu, H., and Ma, J. (2017). Large-scale remote sensing image retrieval by deep hashing neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):950–965.

Liu, P., Di, L., Du, Q., and Wang, L. (2018). Remote sensing big data: theory, methods and applications.

Yao, X., Yang, L., Cheng, G., Han, J., and Guo, L. (2019). Scene classification of high resolution remote sensing images via self-paced deep learning. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 521–524. IEEE.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Zhu, Q., Sun, X., Zhong, Y., and Zhang, L. (2019). High-resolution remote sensing image scene understanding: A review. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 3061–3064. IEEE.

Zhu, X. and Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation. *ProQuest Number: INFORMATION TO ALL USERS*.