Griffen Guizan

CSULB Portfolio

1 November 2023

My Portfolio

Urban Forests Undergraduate Research Poster, Summer 2023: Story Map

OpenCanopy: Leveraging aerial imagery and deep learning to delineate California's urban tree canopy

N. Nguyen[1], J. Baiza[4], G. Guizan[3], C. Pawlak[2], S. Arnold[3], R. Chase[5], L. Crocker[3], J. Yost[2], M. Ritter[2], G.A. Fricker[3], J. Ventura[1]

[1]Dept. of Computer Science & Software Engineering, [2]Dept. of Biological Sciences, [3]Dept. of Social Sciences, [4]Dept. of Architecture & Environmental Design, [5]Dept. of Natural Resources Management & Environmental Sciences

CAL POLY

Figure 1: Qualitative results of running our model on aerial imagery from Bakersfield, CA. *Left: Input image. Right: Predicted canopy cover map in yellow overlaid on image.*

Introduction

The goal of the OpenCanopy project is to create an open-source platform for scalable canopy cover mapping. Accurate canopy cover maps provide useful information about the geographic distribution of trees in a city and the level of benefit they provide to the city's residents. We present our preliminary investigation into building a deep learning pipeline that will leverage public multispectral aerial imagery and LiDAR data where available to produce high-resolution and accurate canopy cover maps for the entire urban forest of California and support longitudinal analysis.

Background

Urban trees positively impact people's health, happiness, and well-being by providing tangible benefits such as reducing the urban heat island effect and saving on energy bills by alleviating extreme temperatures [1, 7]. In addition to these tangible benefits, urban trees provide services such as sequestering carbon, reducing particulate air pollution, increasing property values, supporting job creation and business growth, and improving urban residents' mental health [1, 6, 7, 8]. Measuring and managing urban forests to understand and maximize their benefits is an expensive and time-intensive process. One way to measure an urban forest is by mapping urban canopy cover. The California Urban Forestry Act created a goal to achieve a statewide "10-percent increase of tree canopy cover in urban areas by 2035, with priority for increasing tree canopy cover in disadvantaged and low-income communities and low-canopy areas" [1]. To achieve this goal, the state needs an accurate and repeatable method for estimating canopy cover in California's urban areas.

Chico, CA	Ground-truth			Precision	Recall	F1-score
Classified	No Canopy Cover	Canopy Cover	Total			
No Canopy Cover	1,317,095	34,859	1,351,954			
Canopy Cover	195,802	452,244	648,046	0.698	0.928	0.797
Total	1,512,897	487,103	2,000,000			

Table 1: Confusion Matrix of our U-Net model on imagery of Chico, CA. Ground-truth canopy cover is obtained from LIDAR

Bakersfield, CA	Ground-truth			Precision	Recall	F1-score
Classified	No Canopy Cover	Canopy Cover	Total			
No Canopy Cover	240	7	247			
Canopy Cover	15	228	243	0.938	0.970	0.954
Total	255	235	490			

Table 2: Confusion Matrix of our U-Net model on imagery of Bakersfield, CA. Ground-truth canopy cover is obtained from hand annotation

References

[1] 1978. California Urban Forestry Act of 1978. https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=PRC&division=4.&title=&part=2.5.&chapter=2.&article=
[2] 2017. U.S. Geological Survey. 2017. USGS 3DEP Lidar Point Cloud dataset. https://usgs.entwine.io/
[3] John Francis and Stephen Law. 2022. Estimating Chicago's tree cover and canopy height using multi-spectral satellite imagery. In NeurIPS Workshop on Tackling Climate Change with Machine Learning.
[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770-778.
[5] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014). Forestry Reports (2023), 1-22
[6] Laura Kleerekoper, Marjolein Van Esch, and Tadeo Baldiri Salcedo. 2012. How to make a city climate-proof, addressing the urban heat island effect. Resources, Conservation and Recycling 64 (2012), 30-38.
[7] SJ Livesley, E Gregory McPherson, and Carlo Calfapietra. 2016. The urban forest and ecosystem services: Impacts on urban water, heat, and pollution cycles at the tree, street, and city scale. Journal of environmental quality 45, 1 (2016), 119-124.
[8] David J Nowak, Daniel E Crane, and Jack C Stevens. 2006. Air pollution removal by urban trees and shrubs in the United States. Urban forestry & urban greening 4, 3-4 (2006), 115-123.
[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 234-241.
[10] Ben G Weinstein, Sergio Marconi, Stephanie Bohlman, Alina Zare, and Ethan White. 2019. Individual tree-crown detection in RGB imagery using semi supervised deep learning neural networks. Remote Sensing 11, 11 (2019), 1309.

Methods

We use high-resolution 60 cm aerial imagery from the United States National Agriculture Imagery Program (NAIP), which consists of four spectral bands: red, blue, green, and near-infrared. We have collected imagery of urban areas in California from 2016, 2018, and 2020. Our goal is to create a convolutional neural network model to automatically label each pixel in the input imagery as tree or not tree. Because it would be too time-consuming and difficult to create enough training annotations for this task by hand, we automatically created training annotations from LiDAR [3, 10]. We collected LiDAR data where available from the USGS 3DEP Lidar Point Cloud dataset [2]. The LiDAR data for each city was processed into a canopy height model (CHM). To produce the target canopy cover map, we thresholded the CHM at two meters and the NDVI value from the aerial imagery at 0.05. As a preliminary investigation, we trained a U-Net model [9] on training data from Santa Cruz. We used Resnet34 [4] as backbone for our U-Net model. With a batch size of 20, and the Adam optimizer [5], we trained the model on 50 epochs and saved the best model obtained during training.

Results

To test the model's generalization ability, we evaluated its performance in estimating the canopy cover of Chico, a city that is geographically and climatically distinct from the training set. After running inference on imagery of the entire city, we sampled two million random validation points within the urban area to compute the confusion matrix (Tab. 1). With 0.928 recall, 0.698 precision, and 0.797 F1 score, the evaluation results demonstrate a high level of agreement between the predicted and ground truth canopy cover maps obtained from LIDAR, which we believe could be further improved through fine-tuning. We also ran the model on imagery from Bakersfield, CA, an area for which little urban forest data is available (Fig. 1). It took around 6 minutes to produce a canopy cover map for the city. The confusion matrix from 490 assessment points verified manually by a student (Tab. 2) demonstrated high accuracy of our model. To increase the accuracy and generalization capability of our model, we intend to add more training data from sites across California where LiDAR data is available. We are also preparing a dataset of hand- annotated image tiles to use for a more rigorous validation of the model's accuracy. We have selected 500 256×256pixel tiles across urban California based on random sampling stratified by climate zone, city area, city population, tree species diversity, image sensor type, date, and time. We will hand annotate each tile by outlining the canopy of each individual tree with a polygon (Fig. 2).

Figure 2: Hand Annotated NAIP image tile in Millbrae, CA. Tile 34/500. *Top Left: Input image. Top Right: Image after hand annotation ready for post processing. Bottom Left: False-Color Infrared image of Tile 34 (vegetation identified in orange). Bottom Right: False-Color Infrared image of Tile 34 after hand annotation.*

Assessing Bakersfield's Urban Canopy Coverage: A Deep (Learning) Dive into Canopy Coverage Classification, Sociodemographics, and Urban Planning

Story Map

Introduction:

Urban areas represent 5% of the land area of California, yet 95% of the state's population resides in them, making urban forests all the more essential for good human health in cities. Urban forests provide many ecosystem services, such as urban cooling, carbon sequestration, reducing urban stormwater runoff, and reducing air pollution. California aims to increase urban tree canopy cover by 10% by 2035 (California Assembly Bill 2251: Urban Forestry: Statewide Strategic Plan). However, a reliable and repeatable method for understanding how much canopy cover the state currently has is not established. Without a baseline, measuring increase or decrease is impossible. Healthy urban forests bring various ecological benefits by reducing surface heating, improving air quality, and reducing stormwater runoff (O'Briena et al., 2022; Tyrväinen et al., 2005). With that said, a deep learning model capable of accurately detecting urban tree canopies will be immensely helpful in monitoring the urban forests to maximize these benefits.

While models classifying urban canopy coverage do exist (Hogland et al., 2018; King & Locke, 2018; Liu, 2008; Ma et al., 2013), they are typically behind paywalls and are privately owned and developed, leading to a lack of transparency or reliability in how their model was created, the data they used, or the model's efficacy. This project aims to produce and analyze a deep learning model created by an interdisciplinary team of computer science students and analyzed by GIS students at California Polytechnic State University San Luis Obispo (Cal Poly), in turn comparing it against one of the private for-profit models developed by EarthDefine, a geospatial data and technology corporation. The project's area of interest surrounds Bakersfield, California, as Bakersfield is perpetually under-served and under-represented when it comes to environmental management, sustainability, and policy. The city is consistently reported as the most polluted city in the United States and will continue to be if nothing is done (Cisneros et al., 2017). Further, an analysis of the trends and relationships between urban canopy coverage in Bakersfield and the sociodemographic characteristics will be conducted, helping to provide solutions for the high levels of pollution burden within the city. Sociodemographic data from CalEnviroScreen 4.0, which provides "California community environmental health screening tools," is used for the analysis.

Urban trees are vital in enhancing the quality of urban environments and providing the ecosystem services mentioned above (Martins et al., 2021). Accurate identification and mapping of urban tree canopies are essential for urban planning, environmental management, and assessing the health and distribution of urban green spaces. However, manual detection and classification of trees in urban areas are time-consuming and labor-intensive (Wang et al. 2021).

In recent years, advancements in deep learning and computer vision techniques have revolutionized image analysis and object detection tasks. Deep learning models, specifically convolutional neural networks (CNNs), have remarkably succeeded in various computer vision applications. The deep learning model used in this project, U-Net CNN, was initially created for biomedical image segmentation but has since been used to perform numerous pixel-based classification tasks. (Wang et al. 2021).

The project involves the following key steps: dataset preparation, model training, inference, and post-processing. We assemble a comprehensive dataset comprising aerial images of urban areas with annotations indicating the presence or absence of tree canopies. Using this dataset, the U-Net model is trained to learn urban tree canopies' distinctive features and spatial characteristics. During the inference phase, the trained model is applied to new, unseen urban images to generate pixel-wise predictions of tree canopies. The resulting classification maps provide detailed information about the location and extent of urban tree canopies, enabling efficient tree inventory management and supporting evidence-based decision-making for urban green infrastructure.

The outcomes of this project will offer valuable insights into the distribution, composition, and health of urban tree canopies. The accurate detection and mapping of urban trees using deep learning techniques have the potential to transform the way cities monitor and manage their green spaces, facilitating sustainable urban planning.

Methods:

How to create and train a Deep Learning model in ArcGIS Pro:

To preface, the student-created deep learning model used in this project was trained prior to its application and analysis on Bakersfield and the comparison between EarthDefine's private for-profit model. With that being said, the model did not need to go through the digitization and training process discussed below, but only the analysis of the two models.

One must train a model to carry out and create a deep learning artificial intelligence (AI) model, as models are only as good as the data they are trained on. In Layman's terms, training a model is like teaching a computer to recognize patterns or make predictions. Just like we learn from examples and experience, the model learns from examples and experiences in a large set of training data. During training, the model looks for patterns and relationships in the data and adjusts its internal parameters to optimize its performance. Training the model gives the model practice questions and answers, so it can learn how to solve similar problems in the future. The more diverse and representative the training data is, the better the model can generalize and make accurate predictions on new, unseen data (Abdi, 2020). In this case, identifying canopy coverage in urban areas based on aerial imagery.

To train the deep learning model, we had to create training data. To create training data, we digitized canopy coverage in NAIP (National Agriculture Imagery Program) imagery. NAIP imagery can be found on USGS Earth Explorer Website and is taken every three years, with approximately one-third of the United States photographed annually. The imagery utilized in this project is from 2018, as EarthDefine's private for-profit model was applied to a mosaic raster dataset of NAIP imagery taken in 2018 covering California. NAIP imagery boasts a 1-meter spatial resolution (very high resolution) and consists of four bands: red (1), green (2), blue (3), and near-infrared (4).

The process of digitizing canopy coverage in ArcGIS Pro is simple. Download and import the NAIP imagery about to be digitized into ArcGIS Pro. Next, create an empty folder to store the shapefile about to be made through the digitization process. Then, right-click the folder just created, and create a new shapefile. Be sure to name the shapefile the same name as the NAIP image to keep track of what shapefile belongs to which image. Right-click on the new shapefile and choose “Edit Features” then “Create” to start the editing session. In the “Create Features” pane, select “polygon” as the construction type to create as many vertices as needed to outline canopy coverage. Start digitizing canopy coverage found within the NAIP image. Use the sketch tools and editing options to create and modify the shapefile features accurately. Once you have finished creating or modifying features, you can save your edits by clicking the "Save" button in the “Edit” tab. To end the editing session, click the "Stop Editing" button in the “Edit” tab. You will be prompted to save your changes before exiting.

To accurately digitize canopy coverage, make sure to digitize any form of a tree that boasts a shadow. However, ensure to exclude the canopy's shadow within your polygons, as that will lead to inaccuracies within the model. Something that helped the team to precisely digitize the full extent of the canopy of a tree was to right-click band one (red) of the NAIP image and change it to band four (near-infrared) in order to create a false infrared image which highlights vegetation in an orange/red color, easily identifiable to the naked eye.

Repeat the digitization steps for all NAIP imagery used in the digitization process, creating a separate shapefile for each image. Make sure to provide a wide variety of geographic locations, tree species, and pixel values when creating training data to encompass the variety of canopy coverage in urban areas fully.

After digitizing, it is time to turn the shapefiles into training data the deep learning model can interpret and train itself with. First, import the data (the shapefiles and NAIP imagery used in the digitization process. Next, run the “Create a Mosaic Dataset” tool and name it something appropriate for the project. Then, run the “Add Raster to Mosaic Dataset” tool and add all the raster NAIP images from every city used in the digitization process. Click run. Following adding the rasters to the mosaic dataset, run the “Reclassify” tool using the mosaic dataset “image” as the input, and change all the raster values to “1” and “NODATA” to “NODATA.” Next, using the “Merge” tool, merge all of the shapefiles created in the digitization process for every city into one polygon layer. Run the “Polygon to Raster” tool on the new merged shapefile layer, followed by the “Reclassify” tool, reclassifying all values to “1” and “NODATA” to “NODATA.” Next, with the reclassified mosaic dataset layer and the reclassified polygon to raster layer of the merged shapefiles, run the “Mosaic to New Raster” tool. Sequentially, run the “Raster to Polygon” tool with the new raster layer created, and a polygon layer where canopy coverage is identified as “2” in the “gridcode” and non-canopy coverage identified as “1” should be created. Finally, run the “Export Training Data for Deep Learning” tool. Use the polygon layer just created as the “Input Feature Class,” “gridcode” as the “Class Value Field,” “Process all raster items separately” for “Processing Mode,” and “Classified Tiles” for “Metadata Format.” Click run. The training data has been created.

Next, the model will be created and trained using “Notebooks” and Python in ArcGIS Pro. A copy of the code used is provided below:

Figure 1: The Code.

```python
from arcgis.gis import GIS
gis = GIS("home")
import os
import zipfile
from pathlib import Path

from arcgis.gis import GIS
from arcgis.learn import prepare_data, UnetClassifier
folder_path = r"G:\OneDrive - Cal Poly\GIS441\UrbanTrees_GroupProj\OUTPUT\Deep_learning_demo"
training_zip_path = (folder_path+ '\MERGE_training.zip')
with zipfile.ZipFile(training_zip_path, 'r') as zip_ref:
    zip_ref.extractall(folder_path)
data_path = folder_path+ '\MERGE_training'
data = prepare_data(data_path, batch_size=8, imagery_type='naip')
data.show_batch(rows=3)

unet = UnetClassifier(data, backbone= 'resnet34')
learning_rate = unet.lr_find()

print(learning_rate)
7.585775750291836e-05
unet.fit(30, learning_rate)
epoch     train_loss  valid_loss  accuracy  dice      time
0         0.112988    0.122035    0.950622  0.919409  00:05
1         0.107768    0.120387    0.949696  0.916776  00:04
2         0.110868    0.123054    0.947681  0.911375  00:04
3         0.110453    0.131294    0.943577  0.904826  00:04
4         0.106307    0.136823    0.947256  0.916545  00:04
5         0.114925    0.122206    0.951195  0.921866  00:04
6         0.113832    0.141526    0.941405  0.908909  00:04
7         0.112226    0.111518    0.953277  0.922258  00:04
8         0.109189    0.112470    0.953092  0.921508  00:04
9         0.106897    0.117242    0.951886  0.923294  00:04
10        0.107586    0.127569    0.949138  0.919625  00:04
11        0.106375    0.107698    0.954561  0.924615  00:04
12        0.109208    0.112179    0.952304  0.923646  00:04
13        0.109863    0.113535    0.950291  0.921925  00:04
14        0.109524    0.121460    0.946671  0.909133  00:04
15        0.109255    0.112057    0.954462  0.927029  00:04
16        0.106778    0.103222    0.956267  0.928165  00:04
17        0.108288    0.107585    0.956477  0.930036  00:04
18        0.107914    0.106430    0.954433  0.926571  00:04
19        0.103702    0.106275    0.955340  0.928191  00:04
20        0.101556    0.103788    0.956170  0.929822  00:04
21        0.095286    0.100023    0.957258  0.931544  00:04
22        0.093100    0.099463    0.957078  0.930458  00:04
23        0.089602    0.099988    0.957446  0.931736  00:04
24        0.085867    0.099758    0.956924  0.930133  00:04
25        0.084135    0.098733    0.957892  0.931893  00:04
26        0.082332    0.098671    0.957410  0.930462  00:04
27        0.088928    0.098745    0.957575  0.931381  00:04
28        0.087395    0.099018    0.957488  0.931352  00:04
29        0.084519    0.098925    0.957504  0.931603  00:04
unet.per_class_metrics()
NoData  1       2
precision       0.999466        0.950018        0.769893
recall  0.999969        0.967104        0.681165
f1      0.999718        0.958485        0.722816
unet.show_results()

unet.save('30e_UrbanTrees_Model')
Computing model metrics...
WindowsPath('G:/OneDrive - Cal Poly/GIS441/UrbanTrees_GroupProj/OUTPUT/Deep_learning_demo/MERGE_training/models/30e_UrbanTrees_Model')
```

Once the code is run and the U-net deep learning model is created, return to ArcGIS Pro to perform the canopy coverage classification. Open the "Classify Pixel Using Deep Learning" tool. Use NAIP images that cover your area of interest as the input raster. Navigate to the U-net model just created and select it as the model definition. Click run. The model has been run, and the output is a raster layer that classifies canopy coverage.

Creating and Analyzing Results:

The area of interest is Bakersfield, California. NAIP imagery was downloaded, encompassing the entirety of the city limits of 2018 and 2022, as EarthDefine's model was applied to NAIP imagery encompassing California from 2018. Once the imagery was downloaded, the Bakersfield 2018 and 2022 city boundary was downloaded from the Bakersfield GIS Portal. Then the NAIP imagery of Bakersfield was meshed together using the "Create a Mosaic Dataset" tool, followed by the "Add Raster to Mosaic Dataset" tool, then sequentially clipped to the 2018 city boundary, then the 2022 boundary using the "Clip Raster" tool to classify canopy coverage in the same area EarthDefine did in 2018, giving our area of interest pictured in Figure 2. The result is the NAIP imagery the student-created model will be applied to, and the imagery the EarthDefine private for-profit model has already been applied to.

After the area of interest was defined, the student-created model was run on the NAIP imagery within the area of interest resulting in a raster layer classifying urban canopy coverage. Next, the EarthDefine canopy coverage classification raster layer was downloaded. Please note: a paywall restricts the model. Access was granted through Cal Poly faculty. Then, the EarthDefine model was clipped to the Bakersfield city limits from 2018 using the "Clip Raster" tool, not the

2022 city limits, as no canopy coverage classification data does not exist for some areas of the 2022 city limits due to the constant reshaping of city boundaries year after year.

Following the two canopy coverage classification raster layers added to the project in ArcGIS Pro, the private for-profit layer downloaded from EarthDefine, and the student-created layer output by the student-created model, an accuracy assessment was conducted, resulting in a confusion matrix. The first step to carrying out the accuracy assessment is to run the “Raster to Polygon” tool on the raster layer output by the student-created model identifying canopy coverage with the “Field” set to “Value” and the “Simplify Polygons” box unchecked. Next, open the “Create Accuracy Assessment Points” tool and select the raster to polygon layer just created for the student-created model. Create 500 points that are “Equalized Stratified Random.” Click run. Following making these points, open the attribute table for the new layer of points, zoom into each of the 500 generated points, and input the value of the class that the NAIP imagery displayed in the “GrndTruth” column. For bias mitigation, the “Classified” column was hidden not to sway the interpretation of each point/pixel. Following the input of all “GrndTruth” values into the attribute table, the “Compute Confusion Matrix” tool was run to provide results on the model's accuracy (Figure 4). Once the confusion matrix is output, the kappa coefficient will be provided. A statistic that, after accounting for chance agreement, provides a quantitative measure of the agreement between the classification and the ground truth data (Congalton & Green, 2019; Foody, 2020). In order to compare and contrast our deep learning model versus the private for-profit canopy coverage classification model, another accuracy assessment was conducted on the private model from EarthDefine, repeating the steps above, starting with the “Raster to Polygon” tool (Figure 6).

Continuing analysis after the accuracy assessments and confusion matrices, download CalEnviroScreen 4.0 (CES) data to look into sociodemographic relationships associated with urban canopy coverage. After the data is downloaded and imported into ArcGIS Pro, using the "Clip Raster" tool, clip the CES layer to the area of interest in Bakersfield. Next, combine the demographic and environmental data of the CES layer with the raster to polygon layer of the student-created canopy coverage classification model layer using the "Pairwise Intersect" tool. Under "Input Features," select the CES layer and the raster to polygon layer for the student-create model. Under "Join Attributes," select "All Attributes," and under "Output Type," select "Same as Input." Running this tool will add and count the canopy coverage polygons to the data for each census tract in Bakersfield, allowing for analysis of the relationships between the amount of canopy coverage and sociodemographic and environmental data.

Lastly, using the intersected CES and canopy coverage layer, the "Hot Spot Analysis (Getis-Ord Gi*)" tool was run to identify relative hot spots of poverty in Bakersfield, followed by relative hot spots of trees or canopy coverage in Bakersfield. The output of the Hot Spot Analysis tool produces a visualization of poverty hot spots (Figure 8) and canopy coverage hot spots (Figure 9).

Results:

Figure 2: Bakersfield, California Canopy Coverage Assessment Area of Interest. The area of interest for the project. Due to EarthDefine's private for-profit model utilizing 2018 urban city limits, the Bakersfield NAIP imagery is clipped to the 2018 boundary and then clipped to the 2022 boundary to find the overlap between the two.

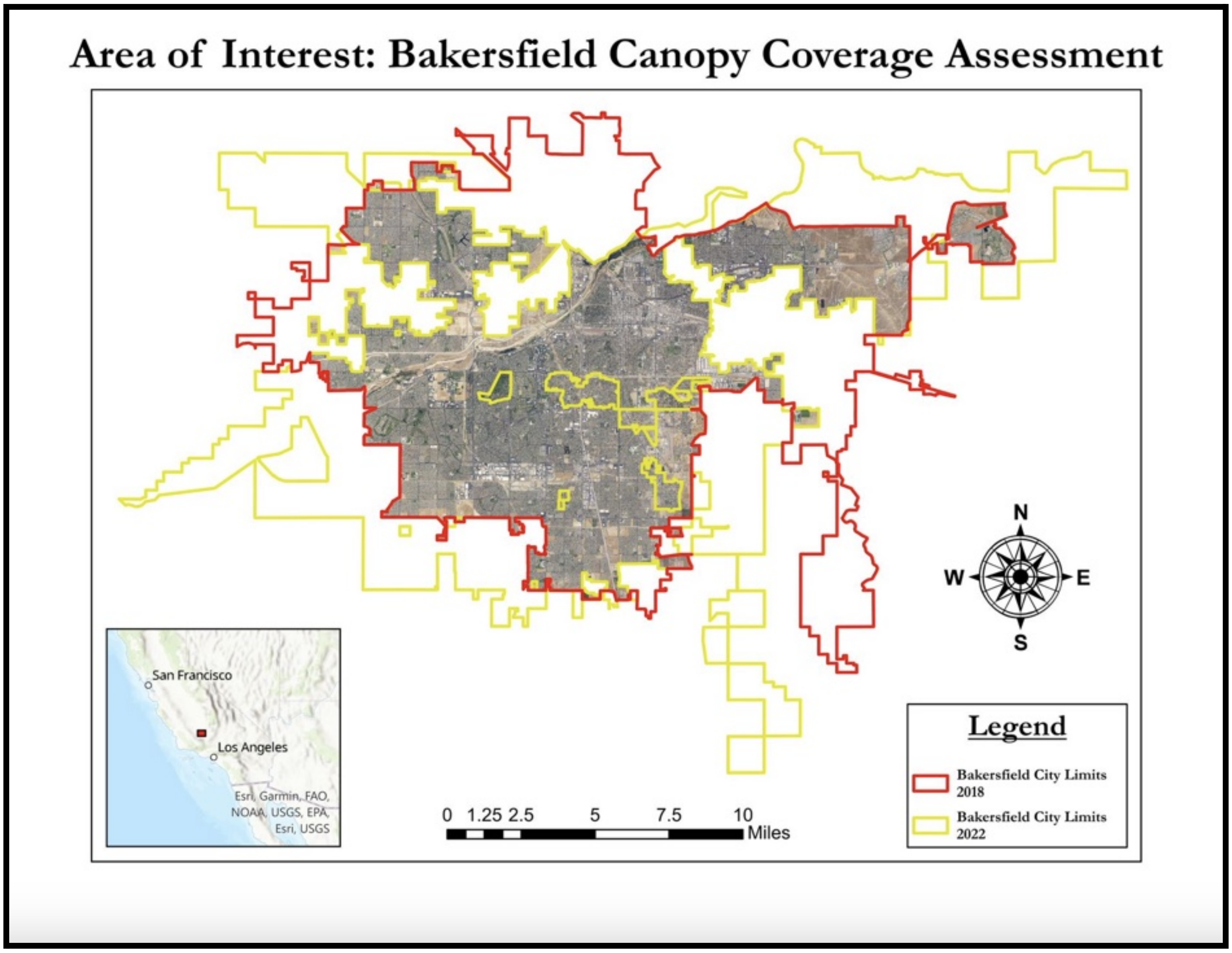


Figure 3: Student-Created Deep Learning Canopy Coverage Classification Model. On the left is NAIP imagery of Bakersfield, a town in the San Joaquin Valley area of California. Imagery from Bakersfield was used to test the model after training. On the right is the imagery with the results of the student-created model overlaid, identifying canopy coverage.

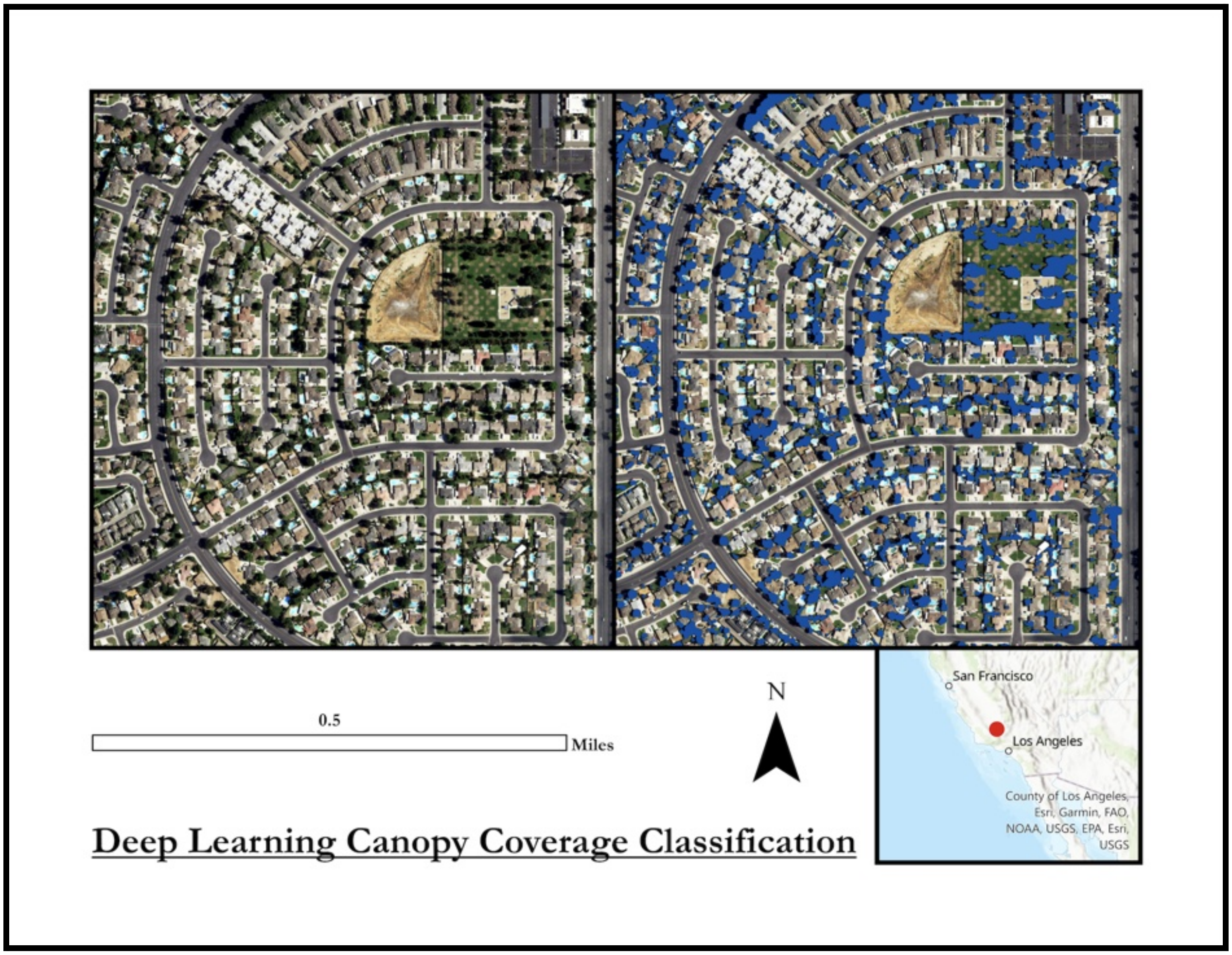


Figure 4: Confusion Matrix, Student-Created Model. The accuracy assessment using 490 randomly generated points yielded a 95.5% accuracy rate. The student model yielded a Kappa coefficient of .91, meaning the canopy coverage classification is 91.0% better than would have occurred strictly by chance. 468 of the 490 accuracy assessment points were correctly classified into one of the two categories. Two hundred forty points were identified as "No Canopy Coverage" and were "No Canopy Coverage," while 7 points were identified as "No Canopy Coverage" and were "Canopy Coverage." Fifteen points were identified as "Canopy Coverage" but were actually "No Canopy Coverage," while 228 points were identified as "Canopy Coverage" and were "Canopy Coverage."

Open-Source Model	No Canopy Coverage	Canopy Coverage	Total	U_Accuracy	Kappa
No Canopy Coverage	240	7	247	0.971659919	0
Canopy Coverage	15	228	243	0.938271605	0
Total	255	235	490	0	0
P_Accuracy	0.941176471	0.970212766	0	0.955102041	0
Kappa	0	0	0	0	0.910174

Figure 5: Private For-Profit Deep Learning Canopy Coverage Classification Model. On the left is NAIP imagery of Bakersfield, CA. On the right is the imagery with the results of the private for-profit model overlaid, identifying canopy coverage. EarthDefine, a private corporation selling a "Tree Map" of the United States, created this model. There is no transparency or reliability in how their model was created, the data they used, or its efficacy.

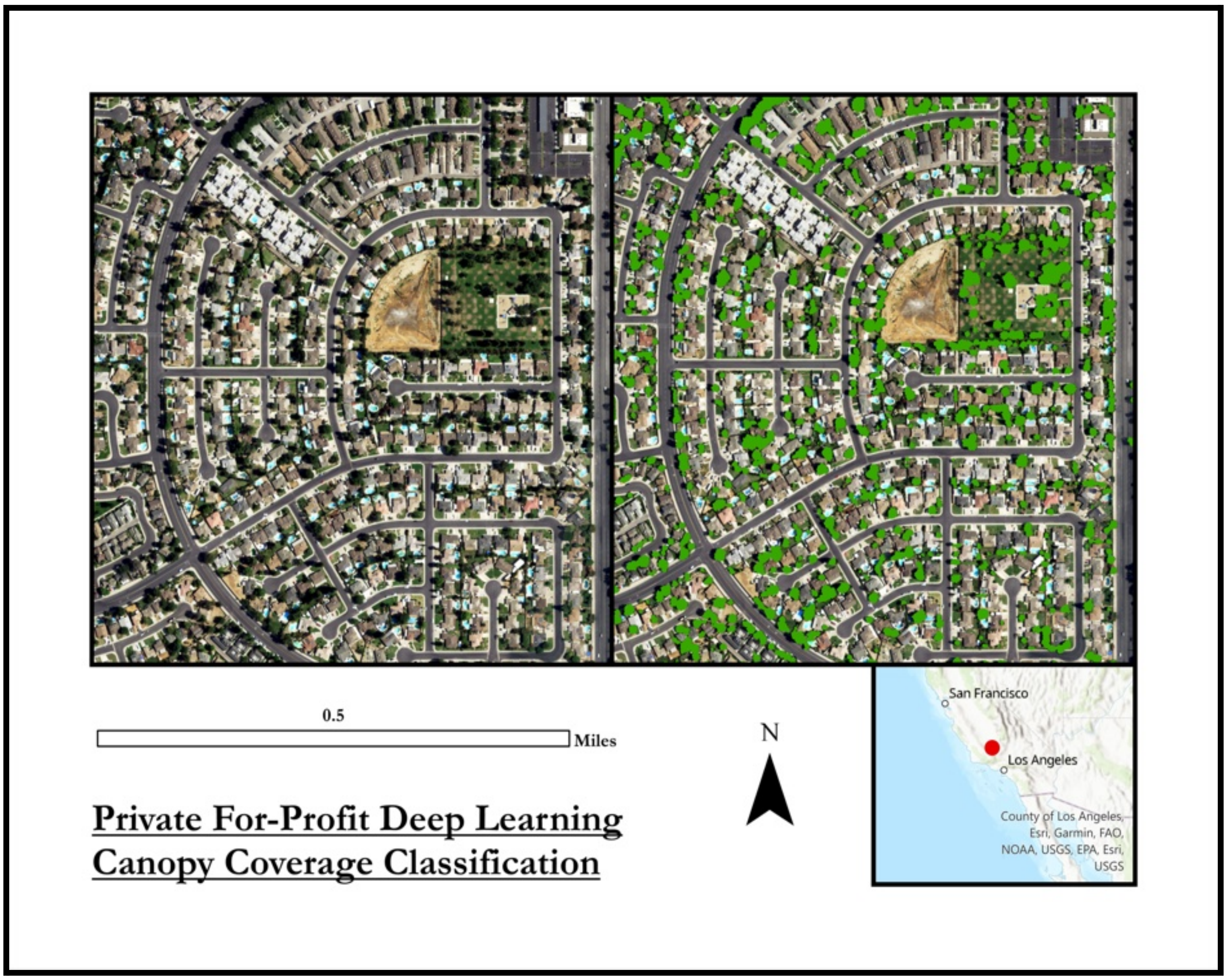


Figure 6: Confusion Matrix, Private For-Profit Model. The accuracy assessment using 495 randomly generated points yielded an 88.5% accuracy rate. The private model yielded a Kappa coefficient of .769, meaning the canopy coverage classification is 76.9% better than would have occurred strictly by chance. 438 of the 495 accuracy assessment points were correctly classified into one of the two categories. Two hundred forty-two points were identified as "No Canopy Coverage" and were "No Canopy Coverage," while 7 points were identified as "No Canopy Coverage" and were "Canopy Coverage." Fifty points were identified as "Canopy Coverage" but were actually "No Canopy Coverage," while 196 points were identified as "Canopy Coverage" and were "Canopy Coverage." *Note: the private for-profit model misclassifies land cover as canopy coverage.*

Earth Defined	No Canopy Coverage	Canopy Coverage	Total	U_Accuracy	Kappa
No Canpoy Coverage	242	7	249	0.971888	0
Canopy Coverage	50	196	246	0.796748	0
Total	292	203	495	0	0
P_Accuracy	0.828767	0.965517	0	0.884848	0
Kappa	0	0	0	0	0.769446

Figure 7: Student-Created Model vs. Private For-Profit Model. The results are displayed together. The student-created model is visualized in blue, and the private for-profit model is in red. Note: The student-created model has more precise edges and does not overclassify canopy coverage compared to the private for-profit model.

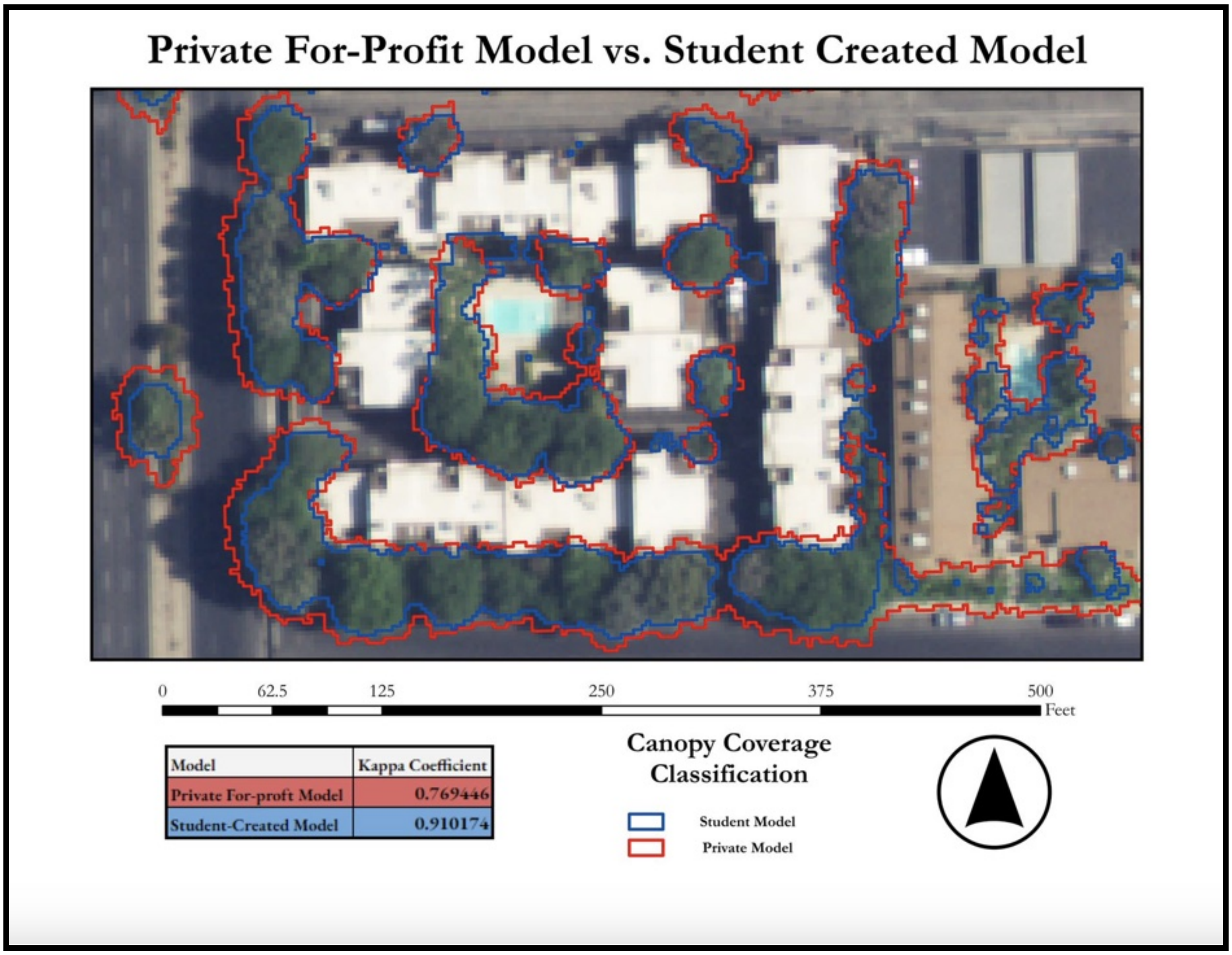


Figure 8: Identified Hot Spots of Poverty in Bakersfield. In the top left, hot spots of poverty relative to neighboring census tracks are displayed. More red areas experience higher rates of poverty, while more blue areas experience lower poverty rates. On the bottom left, NAIP imagery from the blue area is displayed, while on the bottom right, NAIP imagery from the red area is displayed. Note the abundance of urban canopy coverage in the blue area compared to the red area.

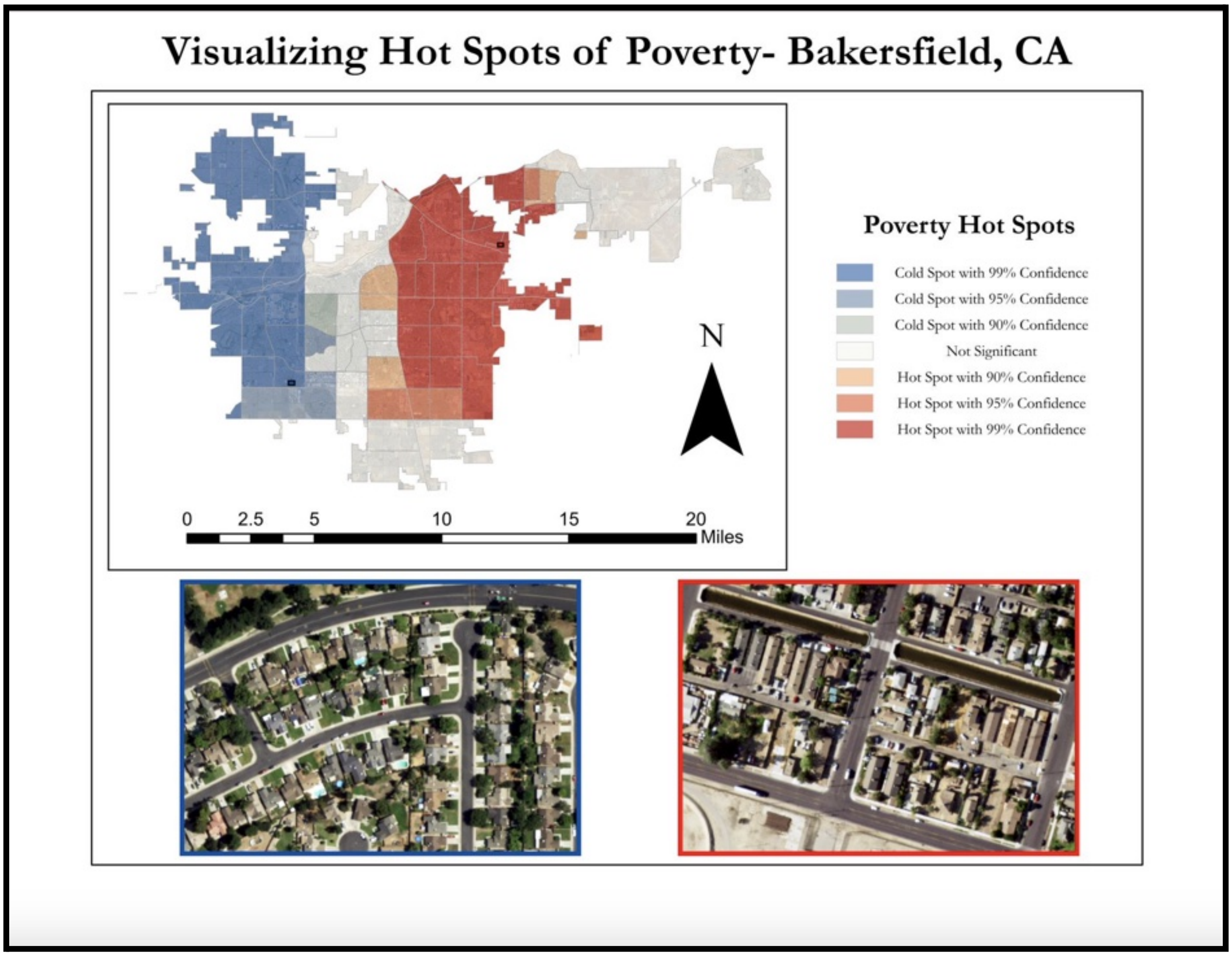


Figure 9: Identified Hot Spots of Canopy Coverage in Bakersfield. In the top left, hot spots of canopy coverage relative to neighboring census tracks are displayed. More red areas experience higher rates of canopy coverage, while more blue areas experience lower rates of canopy coverage. On the bottom left, NAIP imagery from the red area is displayed, while on the bottom right, NAIP imagery from the blue area is displayed.

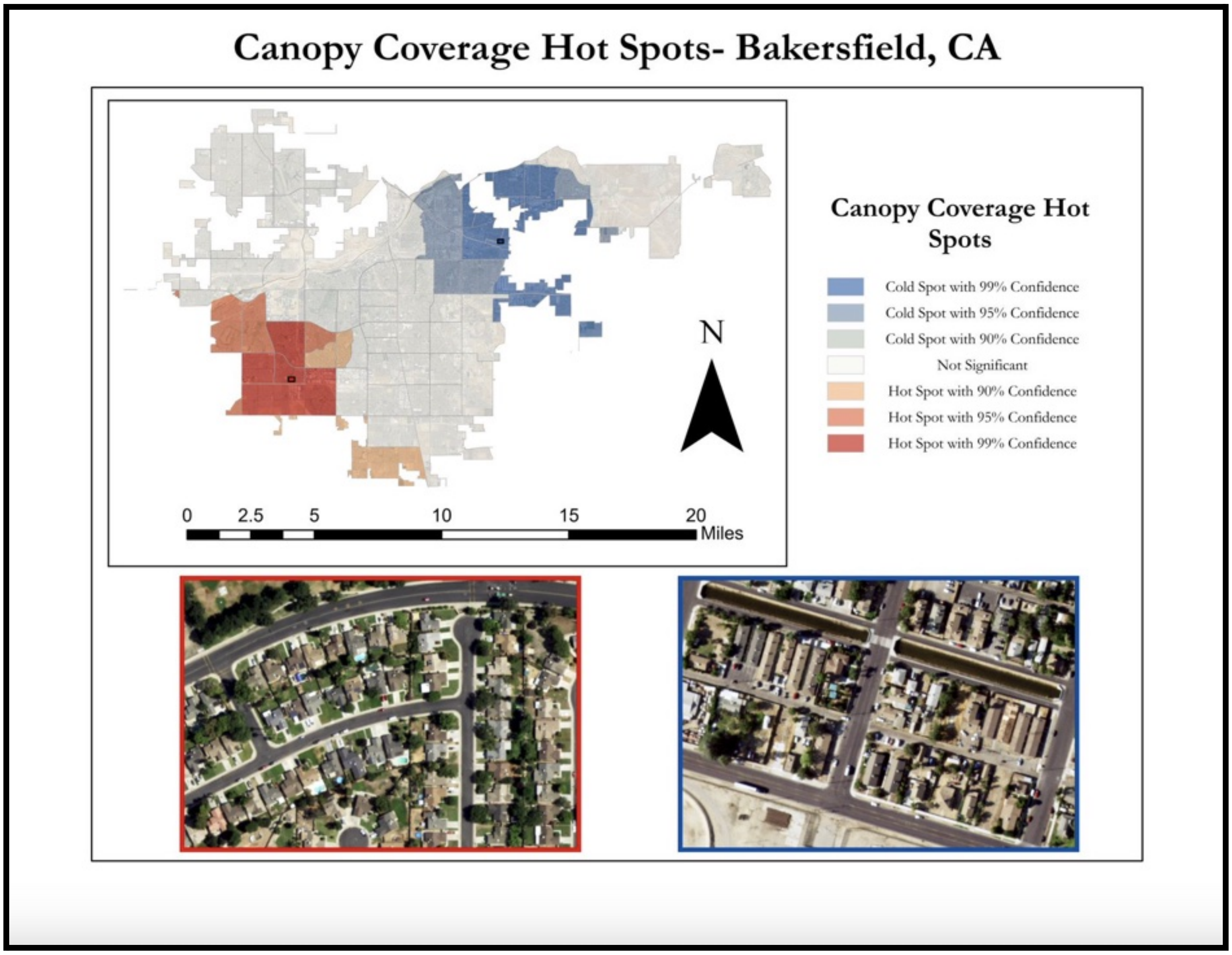


Figure 10: Visualized Relationship Between Poverty and Canopy Coverage in Bakersfield. In the top right, a bivariate map displays the relationship between poverty and canopy coverage in different Bakersfield census tracts. Areas with a combination of low canopy coverage and low poverty are displayed in white. Areas with a combination of high canopy coverage and low poverty are displayed in blue. Areas with a combination of low canopy coverage and high poverty are displayed in orange. Areas with a combination of high canopy coverage and high poverty are displayed in brown.

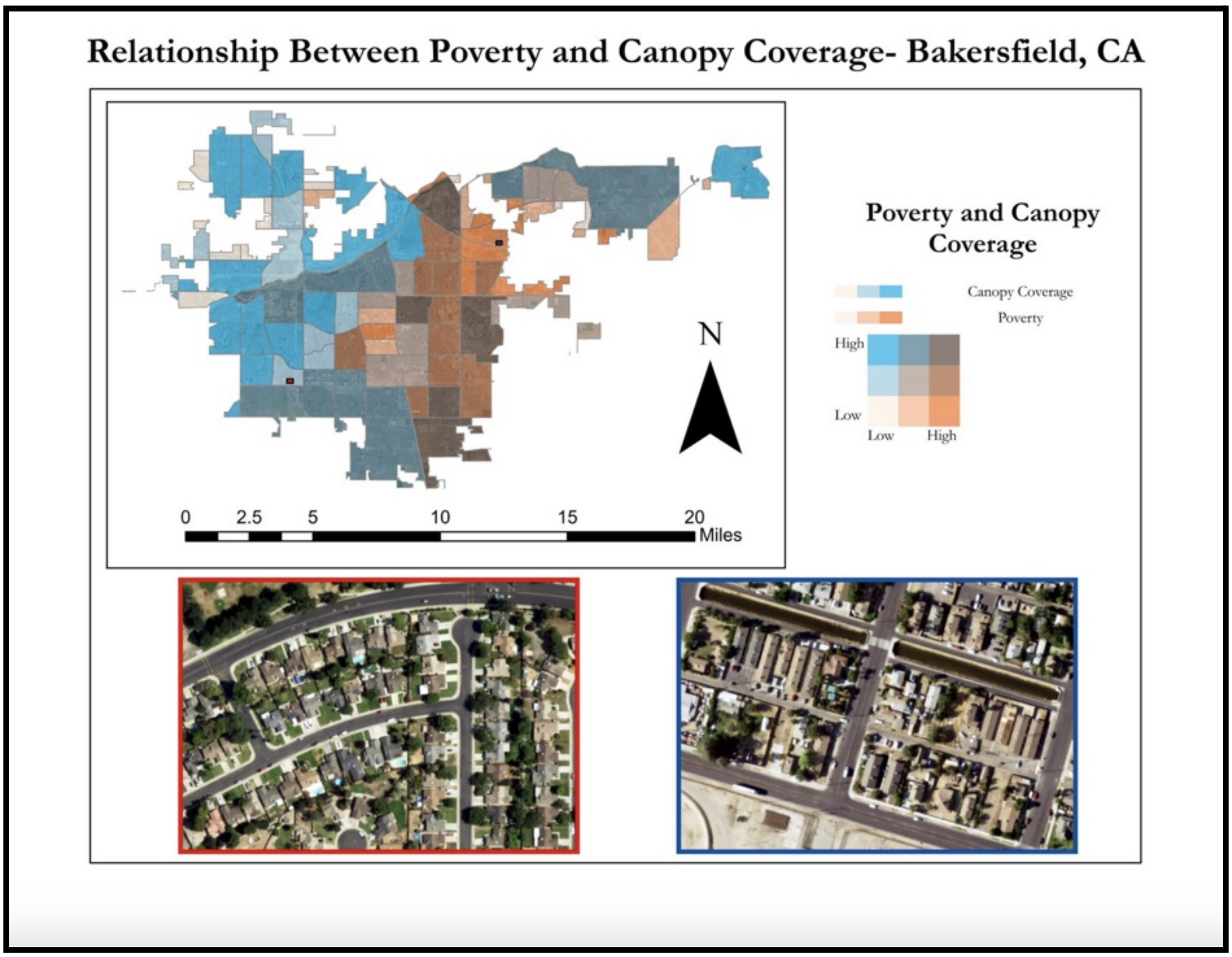
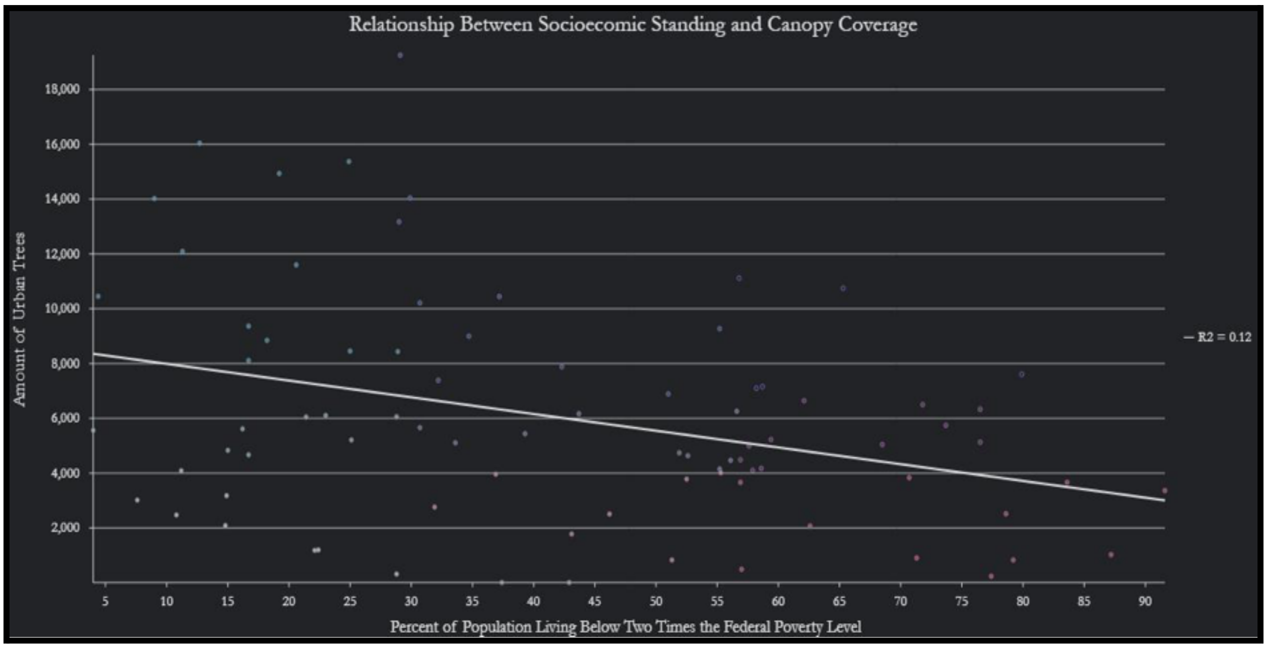


Figure 11: Histogramatic Relationship Between Poverty and Abundance of Canopy Coverage. On the x-axis is the percent of the population living below two times the federal poverty level, CalEnviroScreen 4.0's definition of poverty. On the y-axis is the amount of urban trees. The relationship yields a y = .12 r-squared value, meaning the percentage of the population living below two times the federal poverty level explains only 12% of the urban tree variability.



Discussion:

The student-created model yielded a higher kappa coefficient, 91.0%, and accuracy of 95.5%, than the private for-profit model created by EarthDefine, which had a 76.9% kappa coefficient and 88.5% accuracy. Compared to the private for-profit model, the student-created model was able to identify urban canopy coverage more accurately and precisely. As shown in Figure 7, the student-created model classifies canopy coverage more precisely, as the private for-profit model creates a sort of buffer around canopy coverage that the student-created model does not have. Further, according to the accuracy assessment and confusion matrix output from the accuracy assessment, the private for-profit model misclassifies land cover as canopy coverage frequently more than the student-created model. Overall, the student-created model yielded a more accurate and precise urban canopy coverage classification.

In addition to the positive results assessing our model against EarthDefine's model, interesting sociodemographic trends were identified. There is an inverse relationship between poverty levels and the number of urban trees within the census tracts in Bakersfield, California. In other words, the higher the percentage of poverty in a census tract, the lower the number of urban trees they will have. Moreover, when hot spots for poverty and hot spots for canopy coverage were visualized, the areas identified as poverty hotspots were also identified as canopy coverage cold spots. An interesting trend is posed that, when visualized together in Figure 10, identifies areas primed for an urban tree increase. The census tracts in the northeast quadrant of the area of interest suffer high poverty levels and low canopy coverage relative to the neighboring census tracts. While this project yielded positive results, it also highlighted the disparities in our backyard.

Bakersfield experiences an intensely high pollution burden, and these specific areas of higher poverty and lower canopy coverage identified in Bakersfield may feel the impacts more. In order to support these communities and offer a solution to help reduce pollution in these low-income neighborhoods in Bakersfield, we must plant more trees in the urbanscape. As mentioned earlier, California aims to increase urban tree canopy cover by 10% by 2035 per the California Assembly Bill 2251: Urban Forestry: Statewide Strategic Plan, as urban canopy coverage has many environmental and health benefits such as reducing pollution levels, increasing carbon sequestration, and more. The areas with higher poverty levels and lower canopy coverage in the northeast census tracts of Bakersfield need to be the first areas targeted in California's initiative to increase urban canopy coverage by 10%.

Works Cited

Abdi, Abdulhakim Mohamed. "Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data." GIScience & Remote Sensing 57.1 2020: 1-20.

Cisneros, Ricardo, et al. "Understanding public views about air quality and air pollution sources in the San Joaquin Valley, California." Journal of Environmental and Public Health 2017 (2017).

Congalton, Russell G., and Kass Green. "Assessing the accuracy of remotely sensed data: principles and practices." CRC Press, 2019.

Foody, Giles M. "Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification." Remote Sensing of Environment 239 2020: 111630.

Hogland, J., Anderson, N., St. Peter, J., Drake, J. and Medley, P., 2018. Mapping forest characteristics at fine resolution across large landscapes of the southeastern United States using NAIP imagery and FIA field plot data. ISPRS International Journal of Geo-Information, 7(4), p.140.

King, Kristen L., and Dexter H. Locke. "A comparison of three methods for measuring local urban tree canopy cover." Arboriculture & Urban Forestry 39.2 (2013): 62-67.

Liu, X., 2008. Airborne LiDAR for DEM generation: some critical issues. Progress in physical geography, 32(1), pp.31-49.

Ma, Q., Su, Y. and Guo, Q., 2017. Comparison of canopy cover estimations from airborne LiDAR, aerial imagery, and satellite imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(9), pp.4225-4236.

Martins, Gabriela Barbosa, et al. "Deep learning-based tree species mapping in a highly diverse tropical urban setting." Urban Forestry & Urban Greening 64 2021: 127241.

O'Briena, L.E., Urbaneka, R.E. and Gregory, J.D., 2022. Ecological functions and human benefits of urban forests. Urban Forestry & Urban Greening, p.127707.

Tyrväinen, Liisa, et al. "Benefits and uses of urban forests and trees." Urban forests and trees: A reference book (2005): 81-114.

Wang, Zhe, Chao Fan, and Min Xian. "Application and evaluation of a deep learning architecture to urban tree canopy mapping." Remote Sensing 13.9 2021: 1749.

Ground Truthed Land Use and Land Coverage Classification Using NAIP Imagery of Greater San Luis Obispo Area

Abstract:

Using imagery from the National Agriculture Imagery Program or NAIP, I set out to classify the land coverage in the greater San Luis Obispo area. Confined to a desk all quarter, I wanted to get into the field and explore different methods used in the field of geographic information systems to verify supervised classifications of satellite imagery. I carried out a land use and land cover classification of the Greater San Luis Obispo area, and using an Eos Arrow 100 GNSS device, I ground-truthed 30 randomly generated points around San Luis Obispo to assess the accuracy of the supervised land use and land cover classification and to create a confusion matrix analyzing the results. I then carried out a visual accuracy assessment using the NAIP mosaic image and 111 randomly generated points to verify the supervised classification, creating a confusion matrix and comparing it against the ground-truthed results to see which accuracy assessment method is more accurate. The results suggest the computer is more accurate at classifying satellite imagery than would have occurred strictly by chance compared to ground truthing.

Introduction:

Land use and land cover classifications are widely used to classify land in many fields of study. Land use and land cover classification are two related but distinct concepts in the field of remote sensing and geographic information systems. Land use refers to the human activities and purposes for which land is used, such as urban areas, agriculture, forestry, mining, and recreation. Land use classification is typically based on land use types, and it helps to understand and manage land resources and urban planning. On the other hand, land cover refers to the physical and biological cover of the Earth's surface, such as vegetation, water bodies, bare soil, and human-made surfaces. Land cover classification is typically based on the types of surfaces present on the ground, and it helps to assess ecosystem health and climate change impacts. In practice, land use and land cover classification are often combined to provide a more complete understanding of land resources and their changes over time (Rwanga and Julius, 2017).

When classifying land into different land use and land coverage classes, it is essential that each land use and land class is correctly identified. One way to confirm or verify your computer-generated classification is to ground truth your results by randomly assigning points to your classification and going out into the area of interest to these points with a Global Positioning System or GPS to see whether or not the computer-generated classification has correctly identified the land class (Lillesand et al. 2015). Further, after randomly assigned points are ground-truthed, one can perform an accuracy assessment by creating a confusion matrix to produce a kappa coefficient. In the context of land use and land cover classification, the kappa coefficient can be used to evaluate the accuracy of a classification method by comparing the classification results with ground truth data. After accounting for chance agreement, the kappa coefficient provides a quantitative measure of the agreement between the classification and the ground truth data (Congalton and Green, 2019).

Carrying out land use and land coverage classifications does not always lead to an accurate result. There are many possible areas in which one can go wrong when executing an accuracy assessment of a land use and land cover classification that leads to inaccurate results. According to Foody, “These problems range from issues associated with a failure to satisfy basic underpinning assumptions through to the limited amount of information on map quality that is actually conveyed by a basic accuracy assessment” (Foody, 2002). Overall, carrying out a land classification on any given area is an iterative process, and it may require multiple rounds of data collection and analysis to achieve satisfactory accuracy. It is also important to document your take into account your methods and results to ensure the reproducibility and transparency of your research.

Methods:

To start my ground-truthed land use and land cover classification of the greater San Luis Obispo area using NAIP imagery, I went to the USGS Earth Explorer Website. I searched for the most recent images of San Luis Obispo taken by NAIP, which happens to be 06/07/2020. As individual NAIP images do not cover much surface area (64 Square Miles) as they are taken at a 1-meter spatial resolution, I had to download nine individual images and combine them to encompass San Luis Obispo. I imported the nine individual images into ArcGIS Pro and used the Raster to Mosaic tool to create a 3x3 composite of San Luis Obispo covering 576 square miles. This 3x3 composite mosaic then became my area of study.

Next, I ran a normalized difference vegetation index or NDVI on the 3x3 mosaic to assess the vegetation in the area of interest. As NAIP imagery consists of four bands: red (1), green (2), blue (3), and near-infrared (4), I obtained my results using the equation (Band 4 - Band 1)/ (Band 4 + Band1). This resulted in an image displaying the areas of productive and unproductive vegetation areas that allows for a more accurate land use and land cover classification.

After running an NDVI on the satellite imagery, I chose the land use and land cover class categories into which I wanted to classify the greater San Luis Obispo area. Six classes were created: Water (10), Urban (20), Tree Coverage (40), Shrubland (50), Rangeland (70), and Agriculture (80). Water is constituted as any body of water. Urban is constituted as any developed piece of land, including residential neighborhoods, stores, universities, etc. Tree Coverage is constituted as any land that is covered by trees, including urban environments. Shrubland is constituted as an area of land with low-level vegetation and shrubbery but no tree coverage. Rangeland is constituted as an area of land that is bare and can be used for grazing by animals. Lastly, Agriculture is constituted as any land that is used to produce crops. Subsequently, I then created training samples for each of the six classes mentioned above to train ArcGIS Pro as to what spectral values belong in what land class for the supervised classification. I then carried out a supervised classification using the training samples using the Maximum Likelihood Classification tool. After the supervised classification took place, I carried out some analysis of the produced raster. I calculated how many pixels were in each class, the percentage of pixels each class constitutes, the area each land class covers, etc.

Following the supervised classification, I ran two accuracy assessments. One to run an aerial visual accuracy assessment, using the satellite imagery as the ground truth, and another accuracy assessment that would enable me to go out into the field and ground truth the supervised classification by traveling to randomly generated points using a GPS and confirming whether the land class the point was assigned is accurate.

The following steps are for the latter accuracy assessment: I used the Reclassify tool to isolate all the pixels for each class in the Maximum Likelihood Classification into their own individual layers. For example, the values for the water class were set equal to one, while the rest of the values were set to zero. I repeated the previous step for all remaining classes. After creating a layer for each land class, I created a three-mile buffer, using the Buffer tool, around my college residence in San Luis Obispo to serve as the radius in which randomly generated points would be created. Within this three-mile buffer, I used the Create Spatially Balanced Points tool to create ten points for each class and chose five points from each class to go ground truth based on proximity, ease of access, and logic. This resulted in six layers consisting of five points each, 30 points for me to ground truth in total.

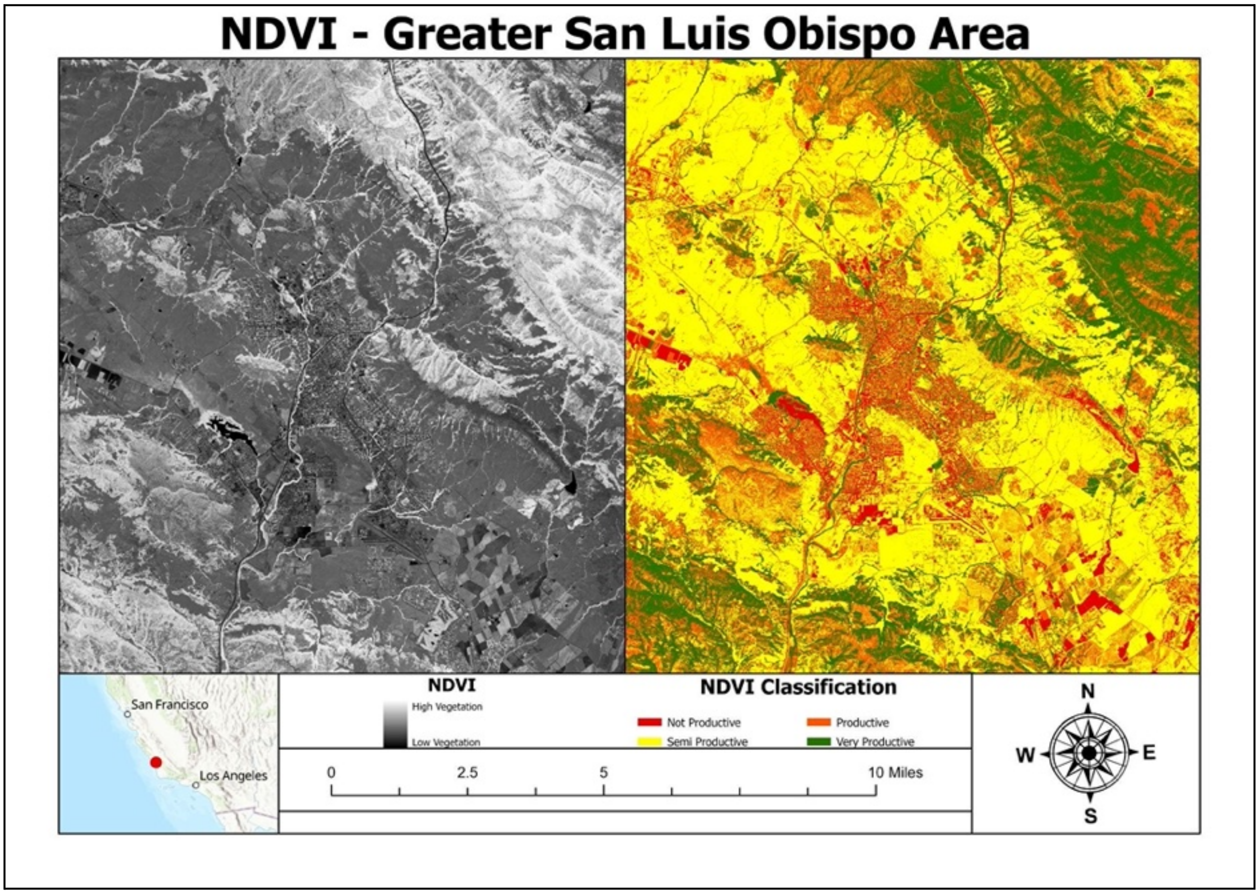
Using ArcGIS Pro Online, I uploaded the map created with the 30 points to ground truth online, creating a web map that can be opened using the Field Maps mobile app provided by ESRI. After uploading the points I needed to ground truth to Field Maps, I acquired an EOS Arrow 100 GPS device that I connected to my phone via Bluetooth and set it as my “Location Provider,” or the GPS that my phone would use to track my location rather than the in-house GPS provided within my phone. The EOS Arrow 100 provides significantly more precise location services (60-centimeter horizontal accuracy) than my mobile telephone (6-meter horizontal accuracy), enabling me to accurately ground-truth the points around San Luis Obispo. After connecting the GPS to my phone, I opened the map I created with the 30 points needed to ground truth in the Field Maps app and traveled to each. Ground-truthing the randomly generated points for each class took several hours of driving, hiking, and trespassing to multiple locations around the town, recording whether the expected class ArcGIS Pro assigned to the location was correct or not. I recorded what I saw and how it differed from the expected class.

Next, using the Merge tool, I merged all of the individual point layers into one layer. I entered the data I collected into the attribute table of the 30 randomly generated points under the "GrndTruth" column I created. I then used the Compute Confusion Matrix tool to assess the accuracy of the supervised classification, comparing the expected class to the ground-truthed class.

In addition to the ground-truthed accuracy assessment, I also conducted an aerial visual accuracy assessment, using the NAIP imagery I had collected as the "ground truth" for the supervised classification. Using the Create Accuracy Assessment Points tool, I created 111 randomly stratified accuracy assessment points. Following making these points, I opened the attribute table for the new layer of points, zoomed into each of the 111 generated points, and input the value of the class that the NAIP imagery displayed in the "GrndTruth" column. For bias mitigation, I hid the "Classified" column so it would not sway my interpretation of each point/pixel. Following the input of all "GrndTruth" values into the attribute table, I ran the Compute Confusion Matrix tool to provide results on the accuracy of my aerial visual accuracy assessment.

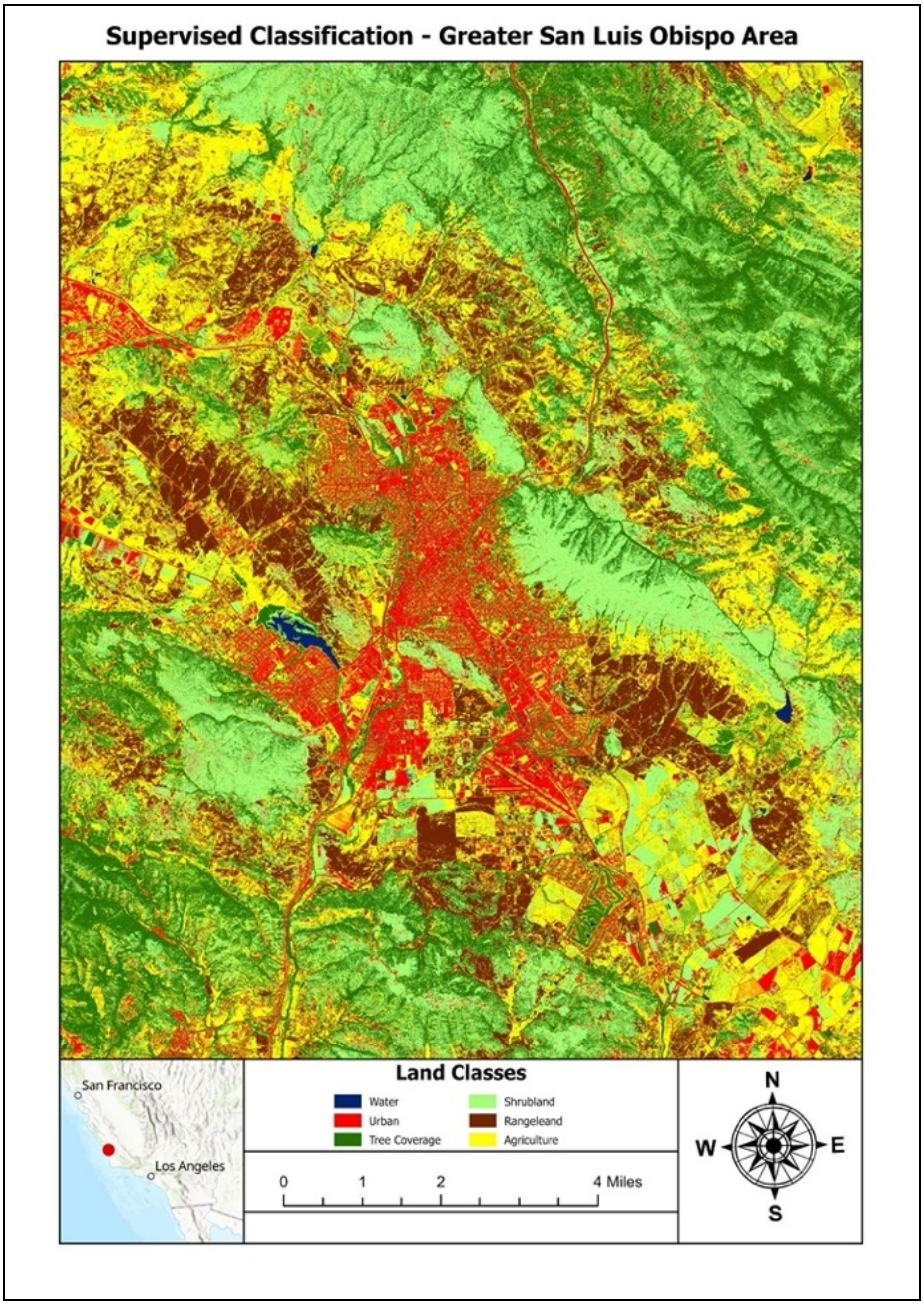
Results:

Figure 1: NDVI - Greater San Luis Obispo Area



First, I ran an NDVI to help facilitate the creation of classes and the overall classification of land in the area. The map portrays the NDVI symbolized in two different manners, allowing for a better understanding of the vegetation in the area of interest. A majority of the productive and dense vegetation in the area is in the northeast and southwest quadrant of the NAIP mosaic.

Figure 2: Supervised Classification - Greater San Luis Obispo Area



The supervised classification of the Greater San Luis Obispo Area is portrayed above. The map divides San Luis Obispo into six classes based on the results from the NDVI above. It classifies the area into either Water, Urban, Tree Coverage, Shrubland, Rangeland, and Agriculture.

Figure 3: Statistics for Land Use and Land Cover Classes

Class	RowID	Pixel Count	% of total	Miles^2
Water 10	0	3554896	0.003400069	1.9584396
Urban 20	1	90277156	0.086345294	49.734889
Tree Coverage 40	2	243565723	0.232957647	134.1836
Shrubland 50	3	260967479	0.2496015	143.77046
Rangeland 70	4	165975174	0.158746418	91.437937
Agriculture 80	5	281196072	0.268949073	154.91467

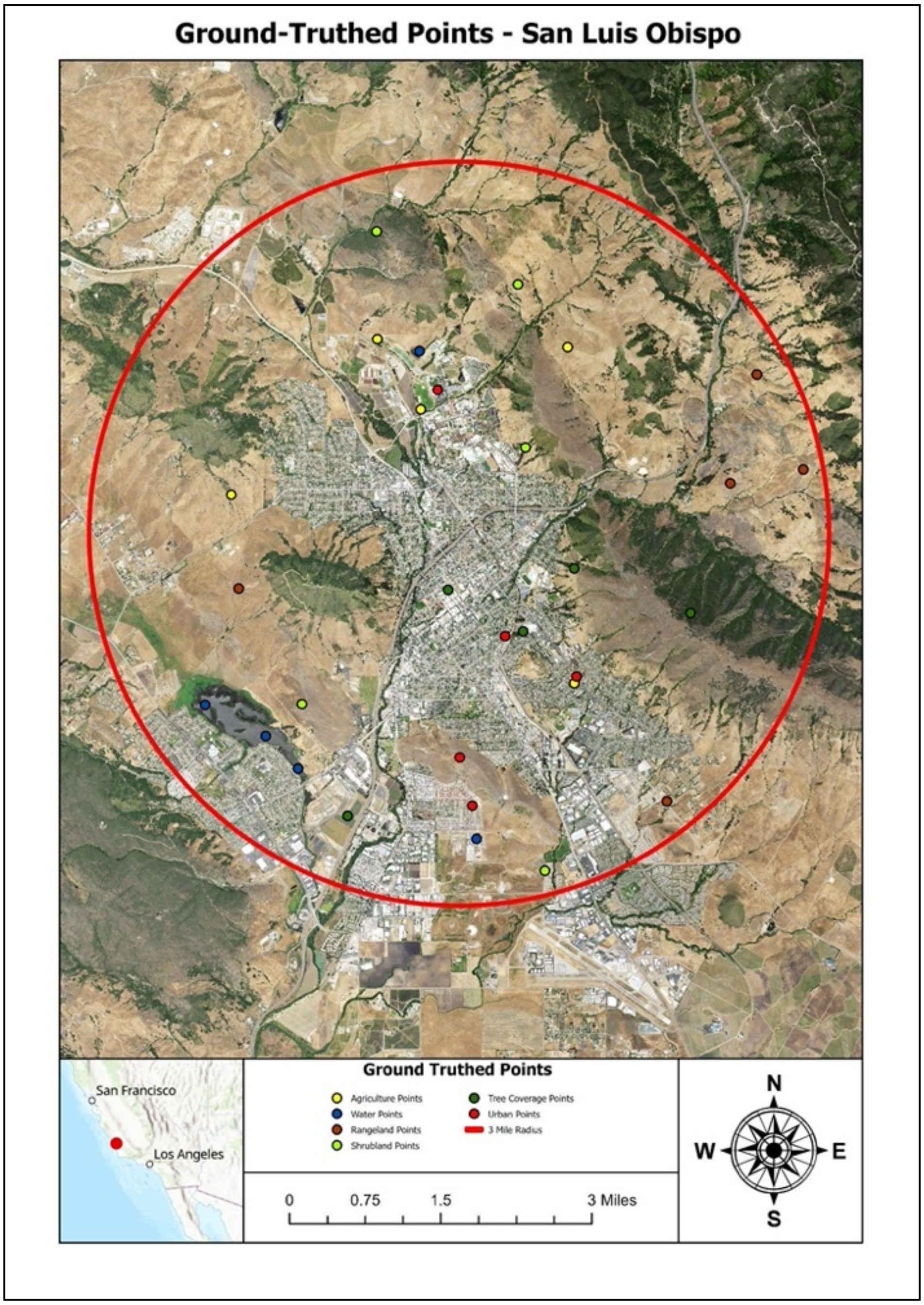
Statistics resulting from the six-class supervised classification. The agriculture class is the largest class, comprising 26.89% of the 3x3 NAIP mosaic and 154.91 square miles of land coverage. Shrubland was classified as the next largest class consisting of 24.96% of the satellite imagery and 143.80 square miles. Tree Coverage follows with 23.30% of the satellite imagery and 134.18 square miles of land coverage, followed by rangeland consisting of 15.87% of the satellite imagery and 91.44 square miles of land coverage. Urban is the second smallest class, with 8.63% of the satellite imagery and 49.73 square miles of land coverage. Lastly, water, the smallest land class, comprises 0.34% of the satellite image and 1.96 square miles of land coverage.

Figure 4: Confusion Matrix for Aerial Visual Accuracy Assessment of Supervised Classification

OID_	ClassValue	C_10	C_20	C_40	C_50	C_70	C_80	Total	U_Accuracy	Kappa
1	C_10	6	2	0	1	0	1	10	0.6	0
2	C_20	0	6	0	1	2	1	10	0.6	0
3	C_40	0	0	21	2	0	0	23	0.913043478	0
4	C_50	0	0	2	19	1	3	25	0.76	0
5	C_70	0	0	0	0	14	2	16	0.875	0
6	C_80	0	0	0	1	17	9	27	0.333333333	0
7	Total	6	8	23	24	34	16	111	0	0
8	P_Accuracy	1	0.75	0.913043	0.791667	0.411765	0.5625	0	0.675675676	0
9	Kappa	0	0	0	0	0	0	0	0	0.603414

The accuracy assessment using 111 randomly generated points resulted in a 72.9% accuracy rate. 81 of the 111 points were correctly classified into one of the six categories. The least accurately classified class during the supervised classification was “C_80,” or agriculture. “C_70” or rangeland would often be classified as agriculture as they share similar spectral properties. The supervised classification yielded a Kappa coefficient of .603414, meaning the classification is 60.3% better than would have occurred strictly by chance.

Figure 5: Ground-Truthed Points- San Luis Obispo



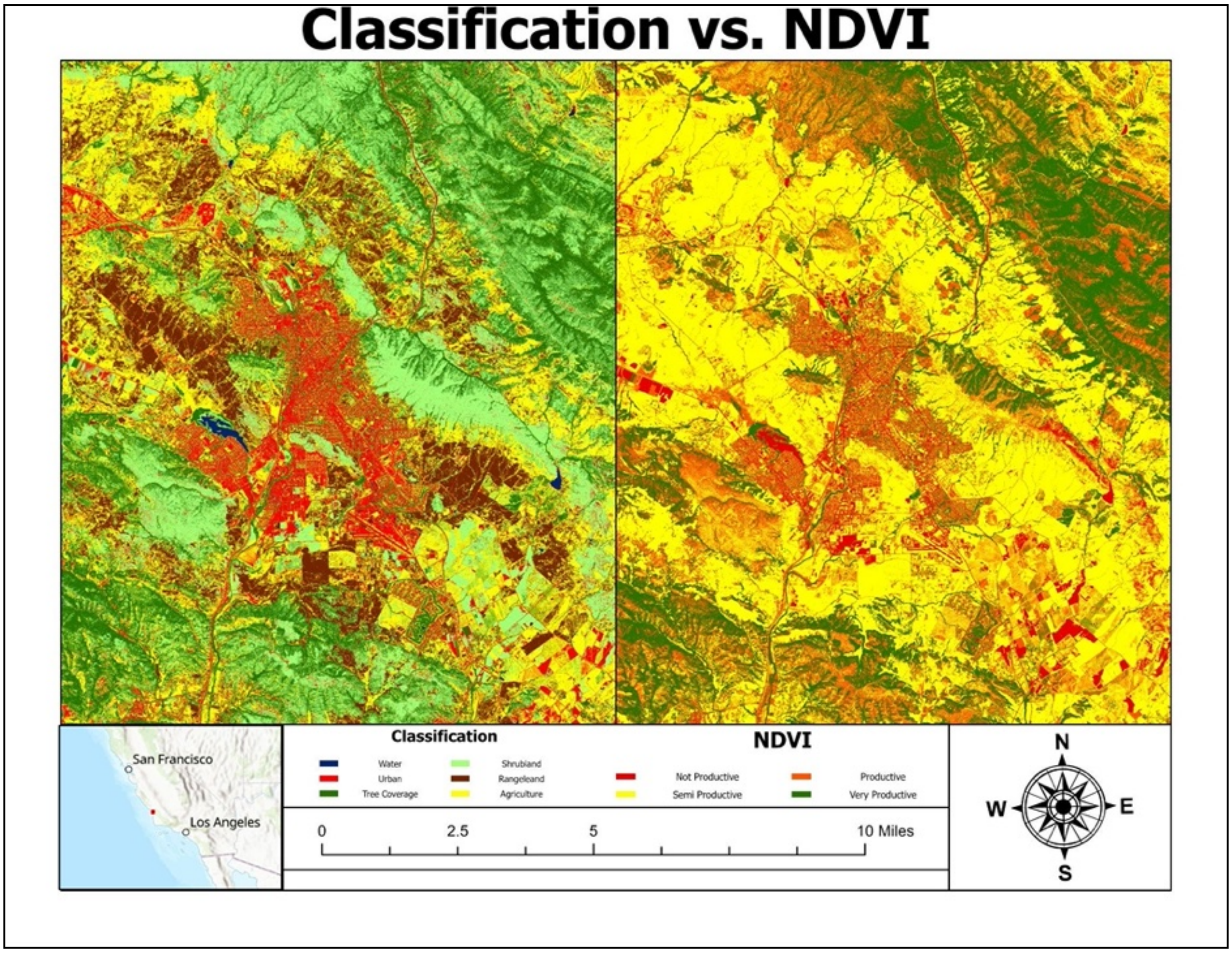
After carrying out the aerial visual accuracy assessment of the supervised classification, I generated 30 random points within a three-mile buffer of my campus residence to ground truth. This map displays the 30 locations I truthed.

Figure 6: Confusion Matrix for Ground-Truthed Accuracy Assessment of Supervised Classification

OID_	ClassValue	C_10	C_20	C_40	C_50	C_70	C_80	Total	U_Accuracy	Kappa
1	C_10	4	1	0	0	0	0	5	0.8	0
2	C_20	0	4	0	0	1	0	5	0.8	0
3	C_40	0	0	4	0	0	1	5	0.8	0
4	C_50	0	1	2	1	1	0	5	0.2	0
5	C_70	0	0	0	0	5	0	5	1	0
6	C_80	0	1	0	1	2	1	5	0.2	0
7	Total	4	7	6	2	9	2	30	0	0
8	P_Accuracy	1	0.571429	0.666667	0.5	0.555556	0.5	0	0.633333	0
9	Kappa	0	0	0	0	0	0	0	0	0.559999

The ground-truthed accuracy assessment using 30 randomly generated points around San Luis Obispo resulted in a 63.3% accuracy rate. 19 out of the 30 points were correctly classified into one of the six categories. The classes that were least accurately classified during the supervised classification were “C_50,” or shrubland, and “C_80,” or agriculture. Multiple times shrubland was misclassified as urban, tree coverage, and rangeland. Many times “C_70” or rangeland would be classified as agriculture as they do share similar spectral properties. The supervised classification yielded a Kappa coefficient of 0.55999, meaning the classification is 55.99% better than would have occurred strictly by chance.

Figure 7: Supervised Classification vs. NDVI



A supplementary map displaying the comparison between the Supervised Classification and the NDVI results.

Discussion and Conclusion:

My results suggest the supervised classification of the Greater San Luis Obispo Area can be more accurate. According to the Kappa coefficient, accuracy of the accuracy assessments, and confusion matrices, the aerial visual accuracy assessment was more accurate than the GPS ground-truthed accuracy assessment. This could have come about for a multitude of reasons, one of them being the lack of consistency in accuracy assessments between the ground-truthed and aerial visual accuracy assessments. In the ground-truthed accuracy assessment, I generated 30 random points, leading to less precise results as there were fewer samples to collect data from, compared to the aerial visual accuracy assessment, where 111 random points were generated, leading to more precise results. Ideally, at least 50 random points should be generated per class to ground truth, but due to a lack of time and economic constraints, 5 points was a more realistic number for me to ground truth in San Luis Obispo.

In a future study, I would add various methods of “ground-truthing” to see which method is most accurate and field check 50 random sites per class. Theoretically, ground-truthing with a GPS is the most precise way to verify if one’s supervised classification is accurate; however, in this case, it is not.

Works Cited

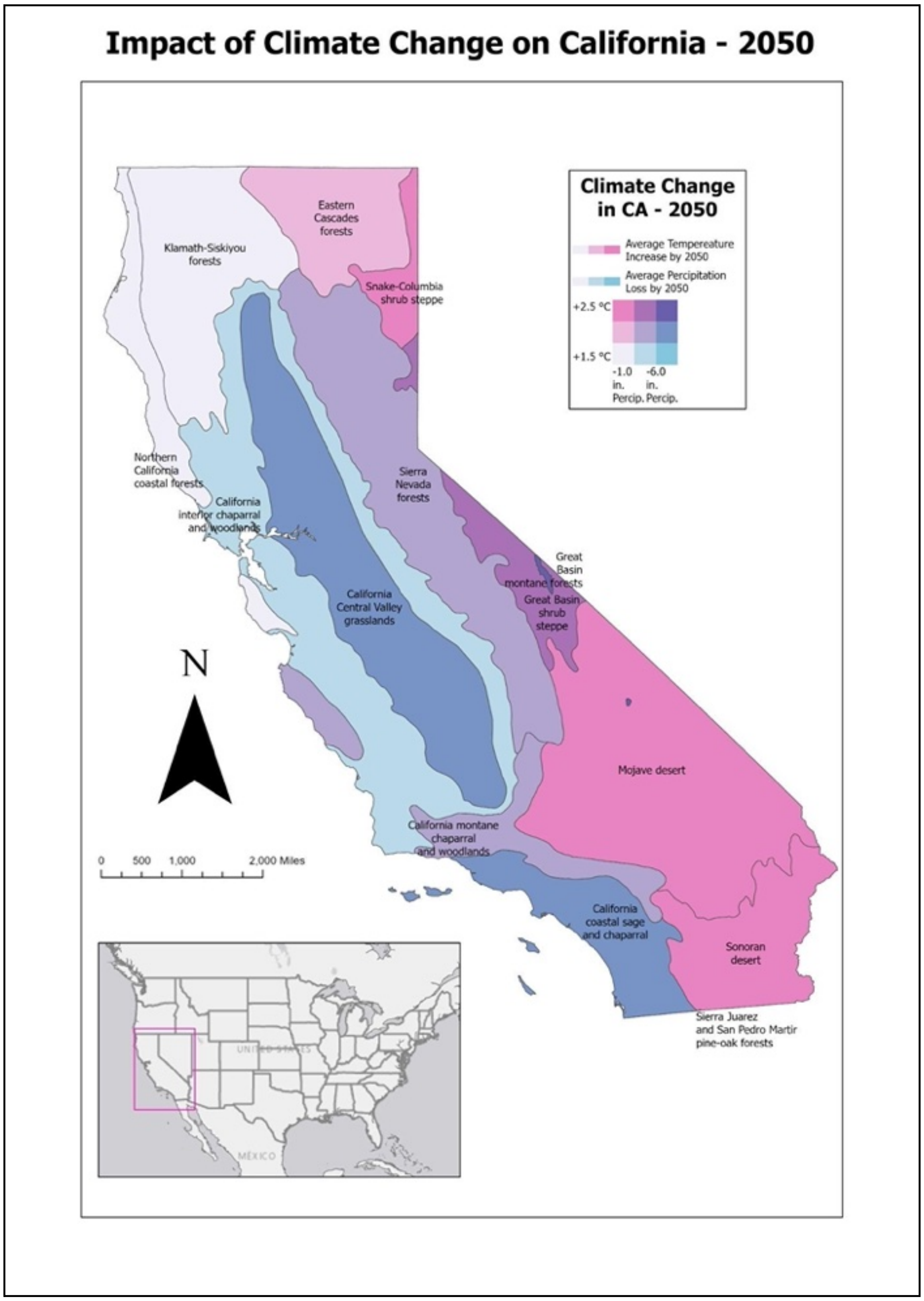
Congalton, Russell G., and Kass Green. *Assessing the accuracy of remotely sensed data: principles and practices*. CRC press, 2019.

Foody, Giles M. "Status of land cover classification accuracy assessment." *Remote sensing of environment* 80.1 (2002): 185-201.

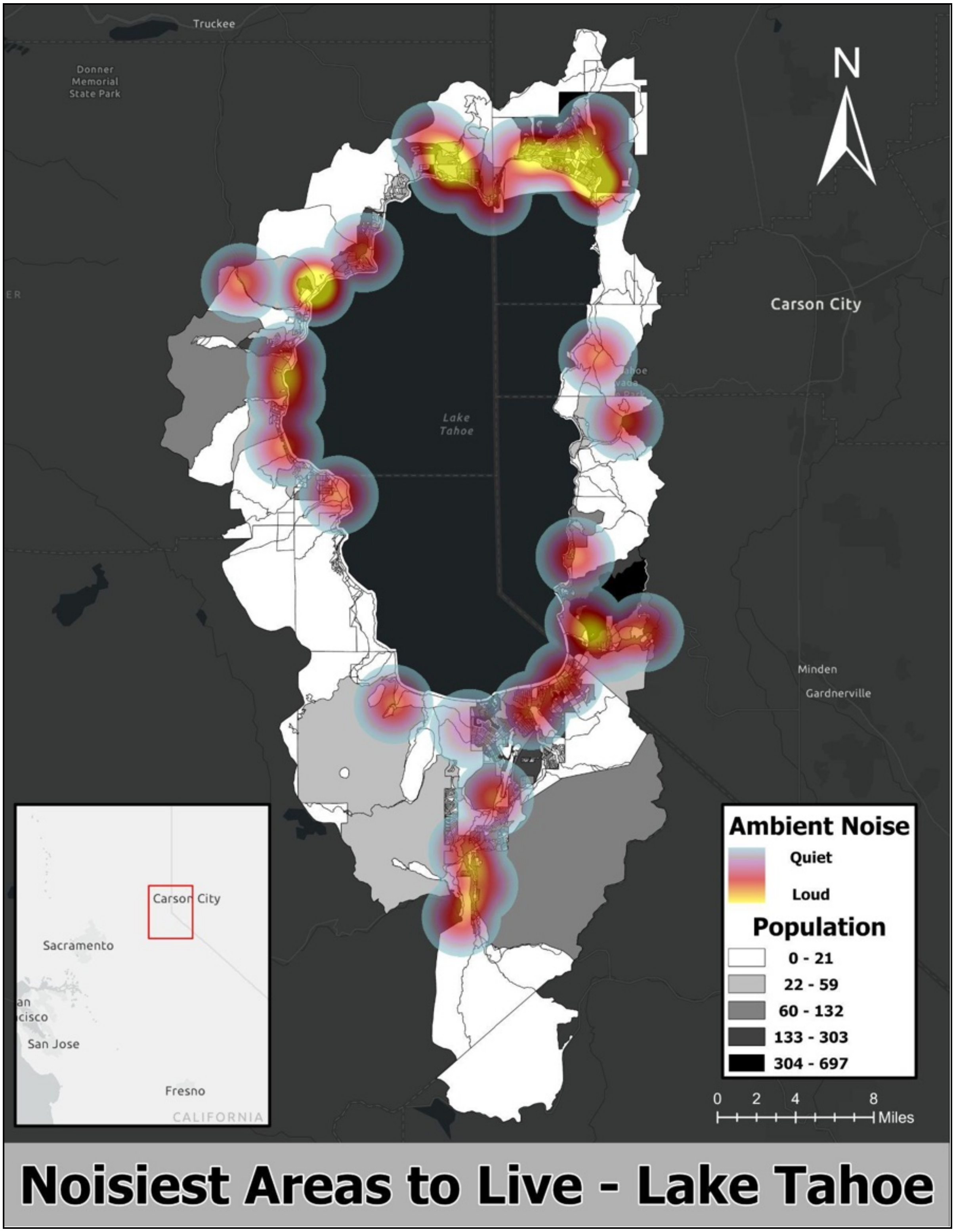
Lillesand, Thomas, Ralph W. Kiefer, and Jonathan Chipman. Remote sensing and image interpretation. John Wiley & Sons, 2015.

Rwanga, Sophia S., and Julius M. Ndambuki. "Accuracy assessment of land use/land cover classification using remote sensing and GIS." *International Journal of Geosciences* 8.04 (2017): 611.

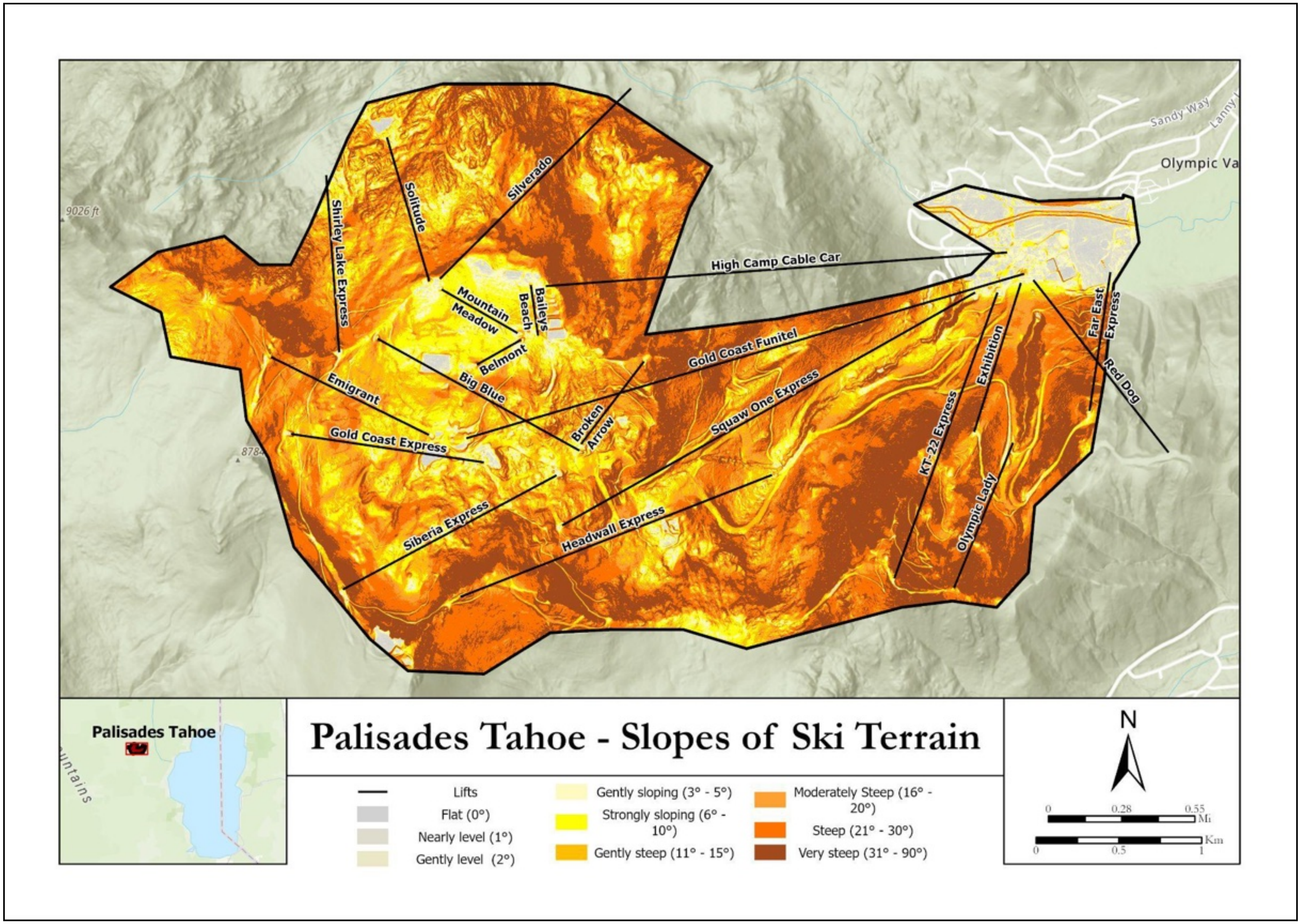
Miscellaneous Maps



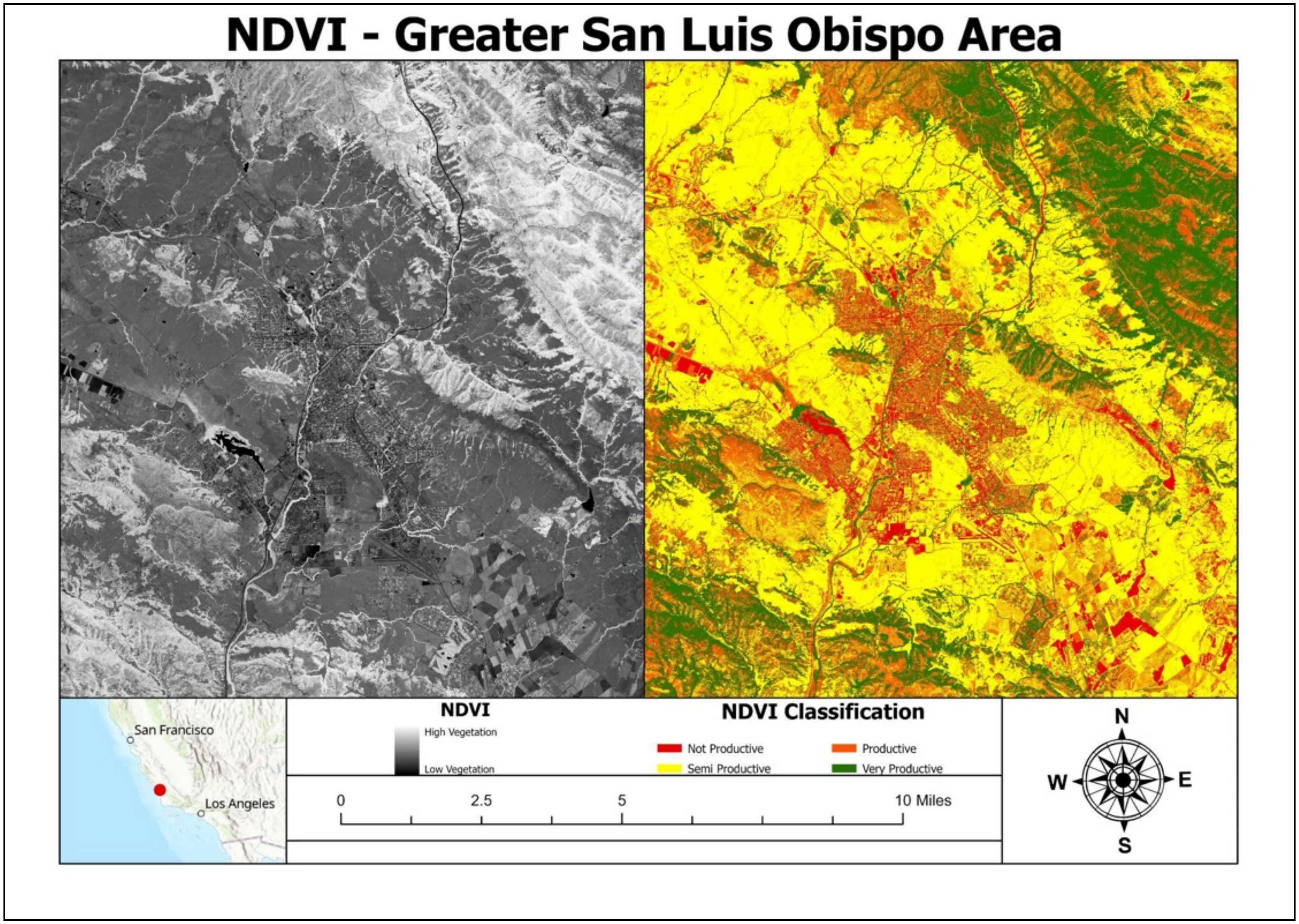
Climate change is here. This map visually represents the impact climate change will have on the rates of precipitation and temperature in California by the year 2050. Areas that are pink will experience at least a 2.5 °C temperature increase, and areas that are bluer will experience a loss of over 6 inches of precipitation. Areas that are purple or some combination of pink and blue will experience both increased temperatures and decreased precipitation.



Lake Tahoe is a very popular destination for a second home, an escape from reality and into nature. As Lake Tahoe grows in popularity, so do the levels of noise. This map depicts the noisiest areas in Lake Tahoe due to transportation and recreation, alongside the respective population of that area, displaying optimal areas to reside.



Palisades Tahoe is a world-renowned ski resort in Lake Tahoe, California. This map displays the different slopes of the different ski runs on the mountain. The darker the color, the steeper the slope of the terrain.



Prior to supervised classification, I ran an NDVI to help facilitate the creation of classes and the overall classification of land in the area. The map portrays the NDVI symbolized in two different manners, allowing for a better understanding of the vegetation in the area of interest. A majority of the productive and dense vegetation in the area is in the northeast and southwest quadrant of the NAIP mosaic.

Land Use and Land Cover Supervised Classification of Santa Catalina Island off the coast of Southern California. After running an accuracy assessment and producing a confusion matrix, the overall accuracy resulted in 81.43% and a Kappa coefficient of .72639.