NASA EEJ – Where the Green Grass Grows:

Geospatial Analysis Methodology

Introduction

This document outlines the steps taken to perform multiple geospatial analysis methods for monitoring trends of greenness and land surface temperature (LST) in the San Francisco Bay Area.

The first component of this analysis involves the classification of green space within the San Francisco Bay Area using a combination of geospatial analysis tools and machine learning techniques. The software tools implemented in this portion of the methodology include ArcGIS Pro and Google Earth Engine (GEE) for Javascript and Python. A geographic object-based image analysis (GeOBIA) approach was used to collect statistics from several remote sensing platforms and inform the training for a Random Forest (RF) Machine Learning (ML) classifier. To distinguish green space from other areas within Bay Area cities, two additional generalized classes were selected for classification: water and impervious/urban spaces, resulting in three classes with which the ML classifier was trained.

The second component of this analysis involves the determination of spatiotemporal trends with respect to greening and LST across the San Francisco Bay Area. The software tools implemented in this portion of the methodology include ArcGIS Pro, and GEE for Javascript and Python. In GEE, satellite observation time series were created for spectral greenness and LST from Landsat archival data between 1990 and 2020. After forming time series for greenness and LST in GEE, the time series data were exported to Google Drive and downloaded to a local machine for linear trend analysis in Python. Linear temporal trend statistics were calculated and aggregated to vector shapefile data tables for census tracts from the 2010 US Census.

Green Space Classification

Ground Truth, GeOBIA, and ML Training

Before training the ML algorithm, ground truth data was collected to enable image classification. The ground truth data was subset into two secondary datasets: training and validation. The ground truth dataset for this assessment was manually collected in ArcGIS Pro software using the "Create Point Feature" geoprocessing tool and high spatial resolution (0.6m) base map imagery from the National Agricultural Imagery Program (NAIP). The total number of ground truth training points was 2,764, of which 1,223 (~45%) were labeled as green space, 395 were labeled as water (~14%), and 1,146 (~41%) were labeled as urban/impervious. The ground truth data were divided into 85% training and 15% validation subsets. The training subset was used to collect statistical information and inform the ML algorithm, and the validation subset was used to assess the accuracy of the final ML classification results. After the collection and division of ground truth data, the subsets were uploaded to GEE as cloud assets for image classification.

Green space classification and batch processing

After collecting ground truth data, image segments (also known as geographic image objects) were created in GEE by implementing the simple non-iterative clustering (SNIC) algorithm (Tassi & Vizzari, 2020). To perform SNIC, a GEE image collection was formed by combining image composites from NAIP and Sentinel-2 over a grid of 385 10km^2 tiles in the San Francisco Bay Area. For each image collection, mean, standard deviation, and maximum composite values were calculated for all bands and spectral indices designated in Table 1. The image objects produced by SNIC, which intersected with the training data subset, served as the basis for image object statistic compilation. After compiling object statistics, the statistics were provided to the RF ML classifier algorithm for classification training.

Sensor	*Band/Index*
NAIP	Red
	Green
	Blue
	NIR
	Normalized Difference Vegetation Index
	Enhanced Vegetation Index
Sentinel-2	Red
	Green
	Blue
	VNIR (740nm)
	VNIR (842nm)
	SWIR (1610nm)
	Normalized Difference Vegetation Index
	Bare Soil Index

Table 1: Sensors and corresponding bands/spectral indices used in image collection composition.

After training the RF ML classifier, the classifier was applied to the entire San Francisco Bay Area for image classification. The resulting classification produced 8-bit raster datasets, where each pixel was classified as either (1) green space, (2) water, or (3) urban/impervious. Each raster dataset was exported to Google Drive in GeoTiff (GTIFF) format. To overcome memory limitations for exporting very-high spatial resolution rasters in GEE, batch processing was implemented using a gridded tile schema of 385 10km^2 tiles and the output classified images were resampled to 5m spatial resolution.

Classification Post-processing and ML Validation

After classification, all classified rasters were downloaded to a local machine and brought into ArcGIS Pro GIS software for raster mosaicking using the 'Mosaic to New Raster' geoprocessing tool. An overview layout of the mosaicked classification can be seen in Figure 1.

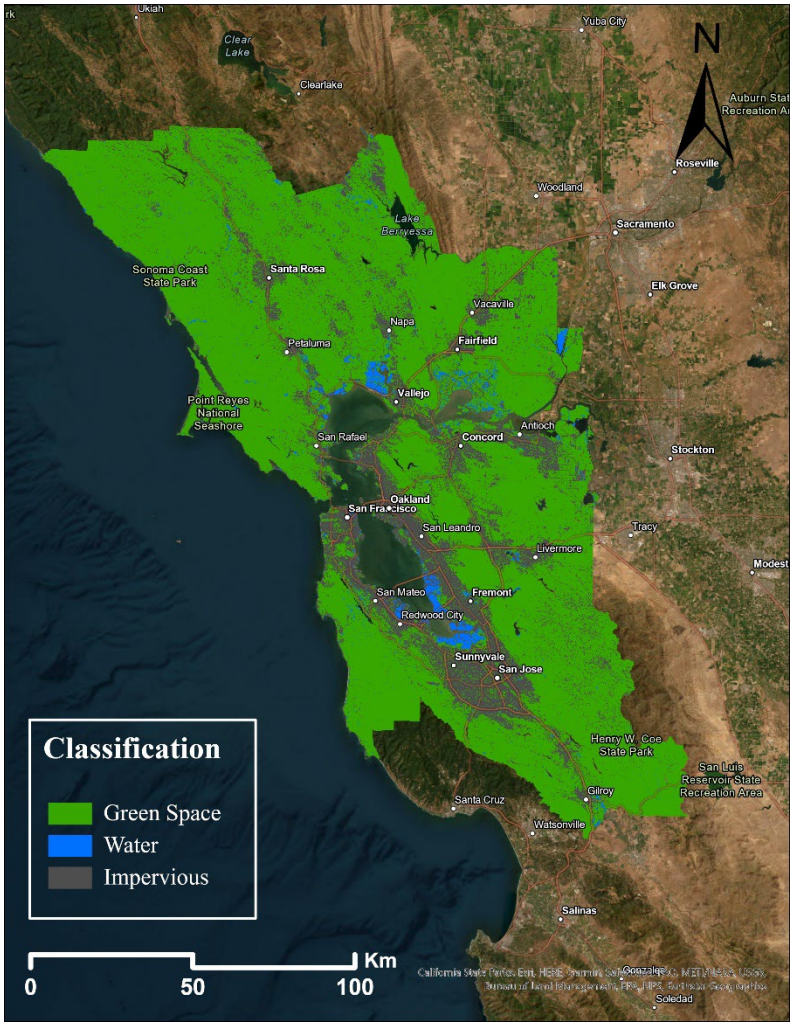

Figure 1: Overview of GeOBIA green space classification over the San Francisco Bay Area

To validate the classification results from the ML classifier, we employed two methods, the first being a classical pixel-wise validation and the second being a polygon-based classification assessment. For both accuracy assessments, we followed the methodology outlined by Congalton and Green (2019).

Spatiotemporal trends of greenness and LST

Time series of median annual and triennial greenness and LST over the San Francisco Bay Area were created in GEE using Landsat archives from Landsat 5 and 8. Time series were created using data acquisitions beginning in 1990 and ending in 2020. Landsat greenness was calculated using the Normalized Difference Vegetation Index from near infrared and red surface reflectance bands. LST measurements were provided by algorithmically derived surface temperature bands from Landsat 5 and 8 Level 2 Collection 2 Tier 1 products. After forming the time series dataset at annual and triennial medians, the time series were subset into the months between July and October to represent the hottest portion of the year in the San Francisco Bay Area. Linear trend analysis was then performed on a pixel-wise basis for both the annual and triennial time series and July-October subsets. The linear trend analysis examined the relationship between the Normalized Difference Vegetation Index (NDVI) or LST (dependent) and time (independent) since the start of the analysis period. Calculated trend statistics include covariance, correlation, slope, intercept, t-statistic, standard error, and p-value.

Socioeconomic, greenness, and LST analysis

To compare the trends of observed physical remote sensing variables with socioeconomic data, the spatial statistics of NDVI, LST, and classified green spaces were collected and aggregated to vector shapefiles representing US Census Tracts. To aggregate these trend statistics at the tract level, median values of trend stats were collected by using tract extents as spatial filters in Python. Additionally, the total area of classified green, urban/impervious, and water were added as attributes to the same US census tract shapefile over the San Francisco Bay Area. This shapefile was then utilized to perform bivariate analysis and geographically weighted regression between physical observations (NDVI, LST, classified green space) and socioeconomic variables in RStudio programming language. Socioeconomic variables included US Census low income move-out/high income move-in rates and the US Center for Disease Control's (CDC) social vulnerability index (SVI) (Lehnert et al., 2020).

Web tool development

The results of all analysis and methodologies described in this document have been aggregated into a web-based tool for visualization and interaction through the GEE WebApp platform. The final web-tool can be publicly accessed through the following link:

NASA EEJ - Where the Grass Grows Greener