Haobing Liu
May 5, 2023

A Trial to Identify the Urban Heat Risk By Using Satellite Imagery in San Francisco

Introduction

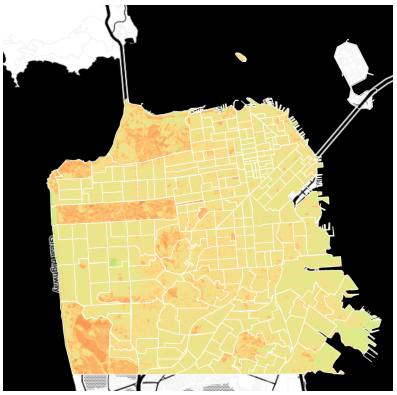
Urban heat risk is a growing concern in many cities around the world, including San Francisco. The rapid urbanization and industrialization of San Francisco. have led to the emergence of urban heat islands, characterized by higher temperatures in densely built-up areas compared to surrounding suburban areas. This phenomenon poses significant health risks to residents, especially during heat waves, and can exacerbate existing socioeconomic inequalities.

The main objective of this project is to assess the urban heat risk in San Francisco over 1 year using satellite imagery in 2020 and machine learning methods. The analysis will identify areas with high heat risk and help city planners and policymakers implement targeted interventions to reduce heat exposure, particularly for vulnerable populations.

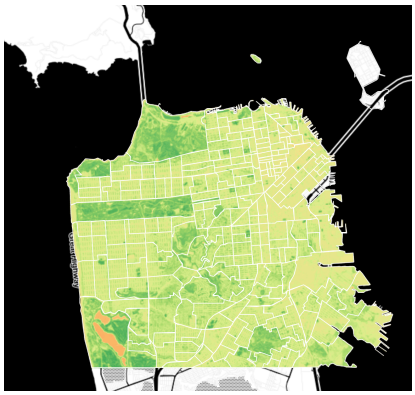
Data

For this project, I will use publicly available Landsat 8 satellite imagery through the Google Earth Engine of San Francisco in 2020. The Landsat 8 images provide high-resolution (30-meter) data in multiple spectral bands, including the thermal infrared band, which is essential for calculating land surface temperature. Then I use Python script that utilizes the Google Earth Engine (GEE) Python API to download Landsat 8 satellite imagery for the year 2020 over San Francisco with less than 5% cloud cover. The script loops through each image in the resulting image collection, exports them as GeoTIFF files, and uploads them to a Google Cloud Storage bucket named ‘sf_imagery’. Moreover, GeoTIFF images can be extracted from the specified Google Cloud Storage bucket.

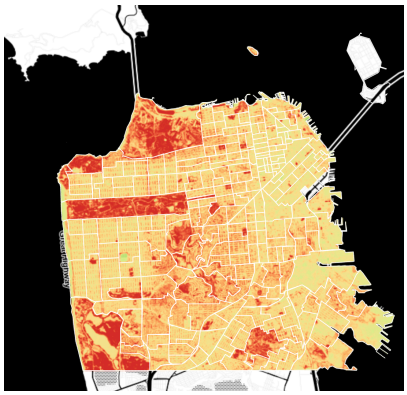
Preprocessing



NDBI



NDVI



BU

Urban heat risk is a complex phenomenon that can be influenced by a variety of factors, including:

The amount and type of vegetation: Vegetation can help cool an area through a process called evapotranspiration. The Normalized Difference Vegetation Index (NDVI) is a commonly used metric to measure vegetation, and it can be calculated from the red and near-infrared bands of a Landsat image.

The amount and type of built-up areas: Built-up areas can absorb and re-emit heat, leading to higher temperatures. This is known as the urban heat island effect. The Normalized Difference Built-up Index (NDBI) can be used to highlight built-up areas in a Landsat image.

The presence of water bodies: Water bodies can also influence local temperatures. They can be identified using the Normalized Difference Water Index (NDWI).

Land Surface Temperature(LST): The surface temperature can be calculated from the thermal infrared bands of a Landsat image. Note that this represents the "skin" temperature of the Earth's surface, not the air temperature.

To develop a model for urban heat risk, I need to compute these and possibly other features from the Landsat image and use them as inputs to the model:

- NDVI is related to vegetation cover. The formula for NDVI is (NIR - Red) / (NIR + Red). NIR is Band 5, Red is Band 4.

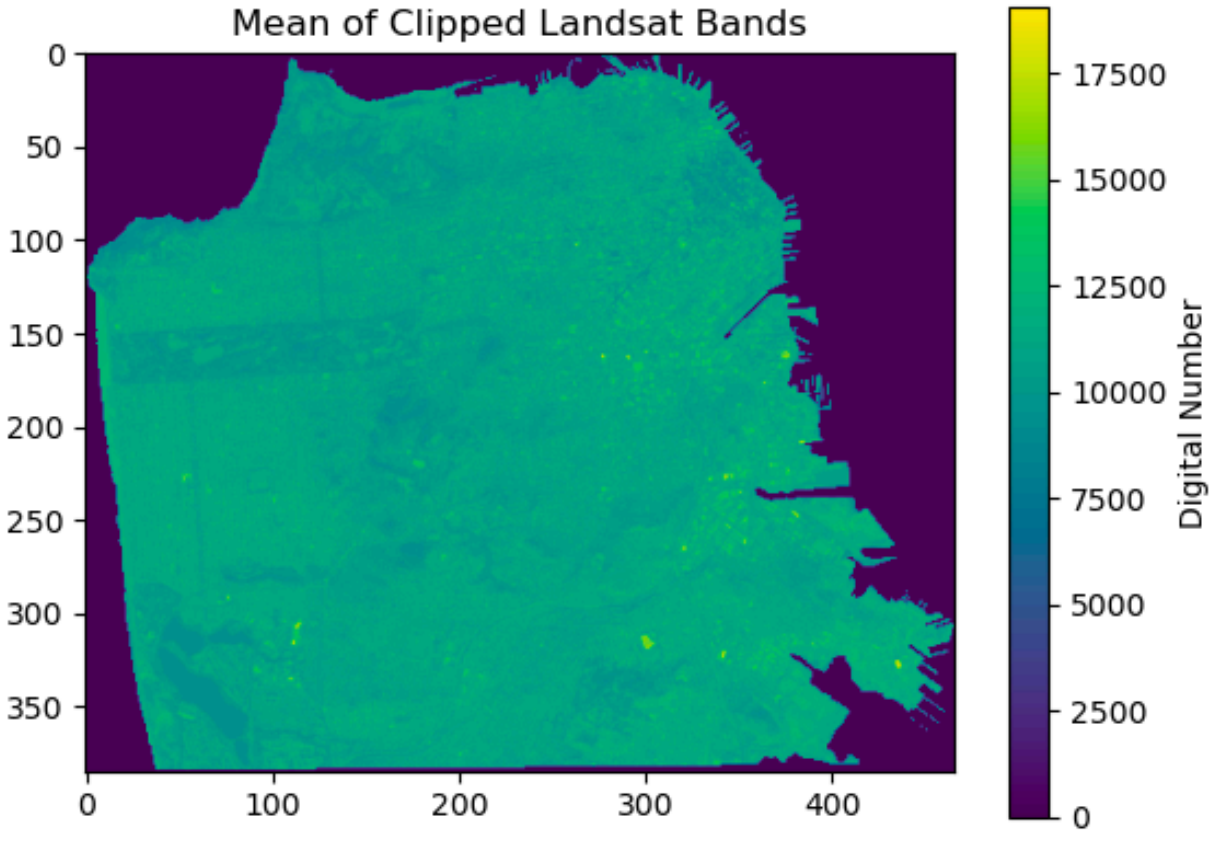
• NDBI is related to built-up (urban) areas. The formula for NDBI is (SWIR - NIR) / (SWIR + NIR), SWIR is Band 6.

• BU (Built-Up index) is a measure of urbanization, , The formula is NDBI - NDVI in this case.

• EVI and SAVI are other vegetation indices that are sometimes used in place of or alongside NDVI;

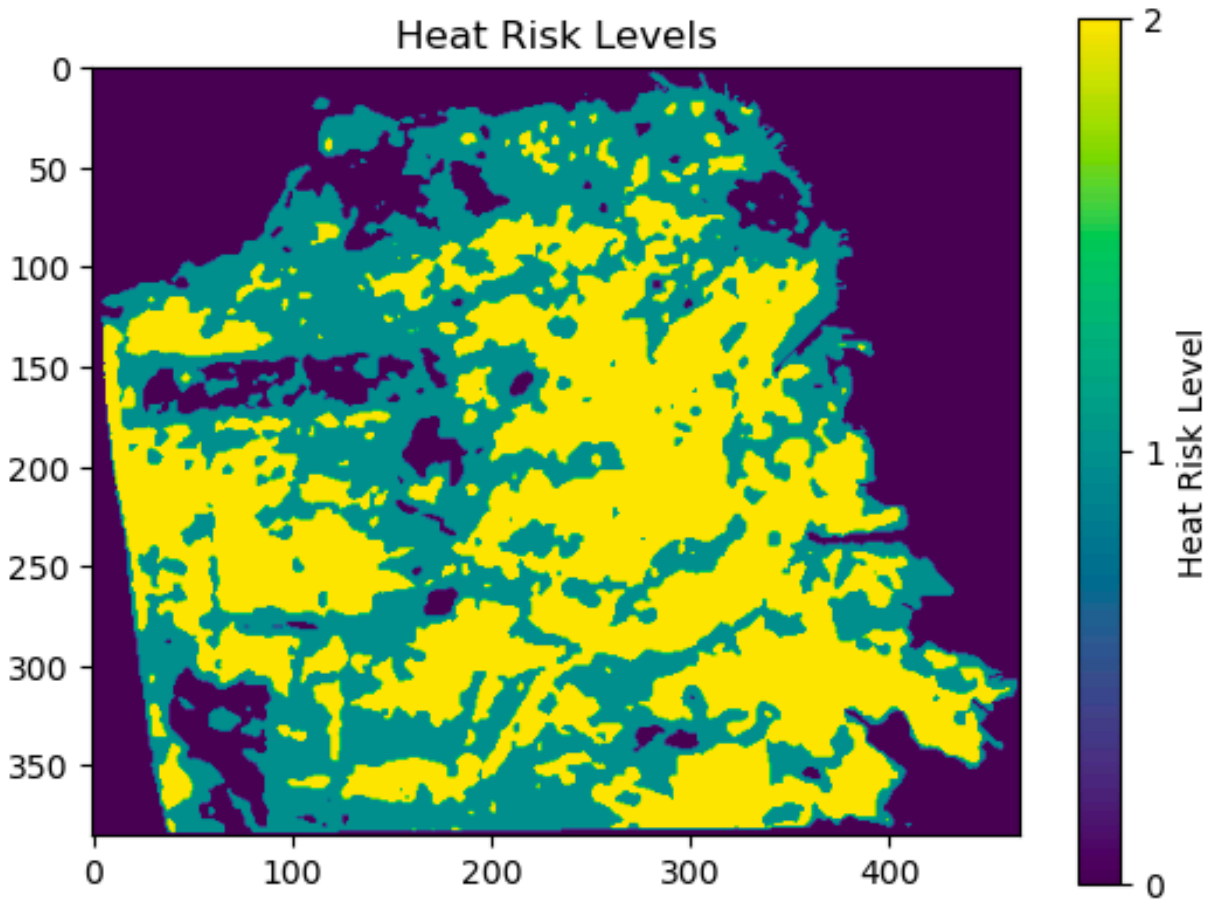
• LST is the Land Surface Temperature, using the Landsat Thermal Infrared band (usually Band 10 for Landsat 8), and then converting the digital numbers to brightness temperature and then apply an empirical equation to estimate the LST.

• Heat Risk Index = LST + NDBI - NDVI, also a definition indicates that areas with high LST, high NDBI, and low NDVI are at high risk.



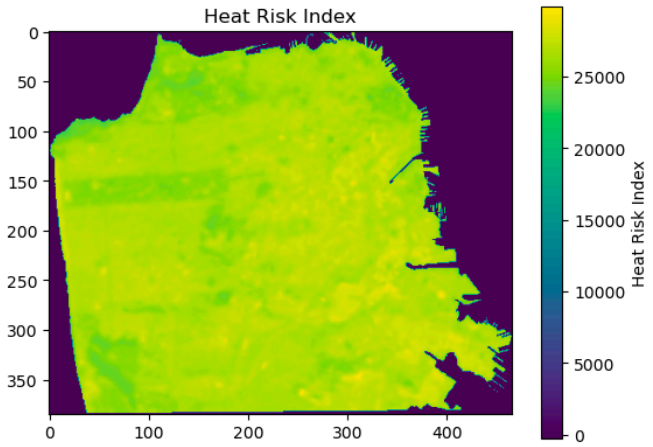
The generated plot reveals the spatial variation of the mean pixel values across the clipped Landsat-8 bands for the San Francisco region of interest. The color scheme used in the plot

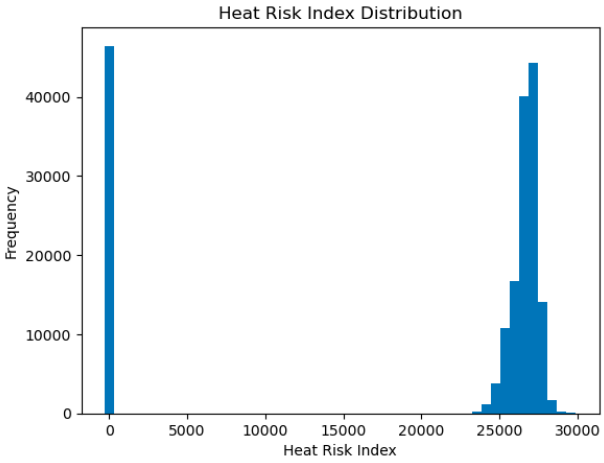
indicates the digital number values for each pixel, which can be used to infer information about the characteristics of the land cover, such as vegetation, water, and urban areas.



Heat Risk Levels

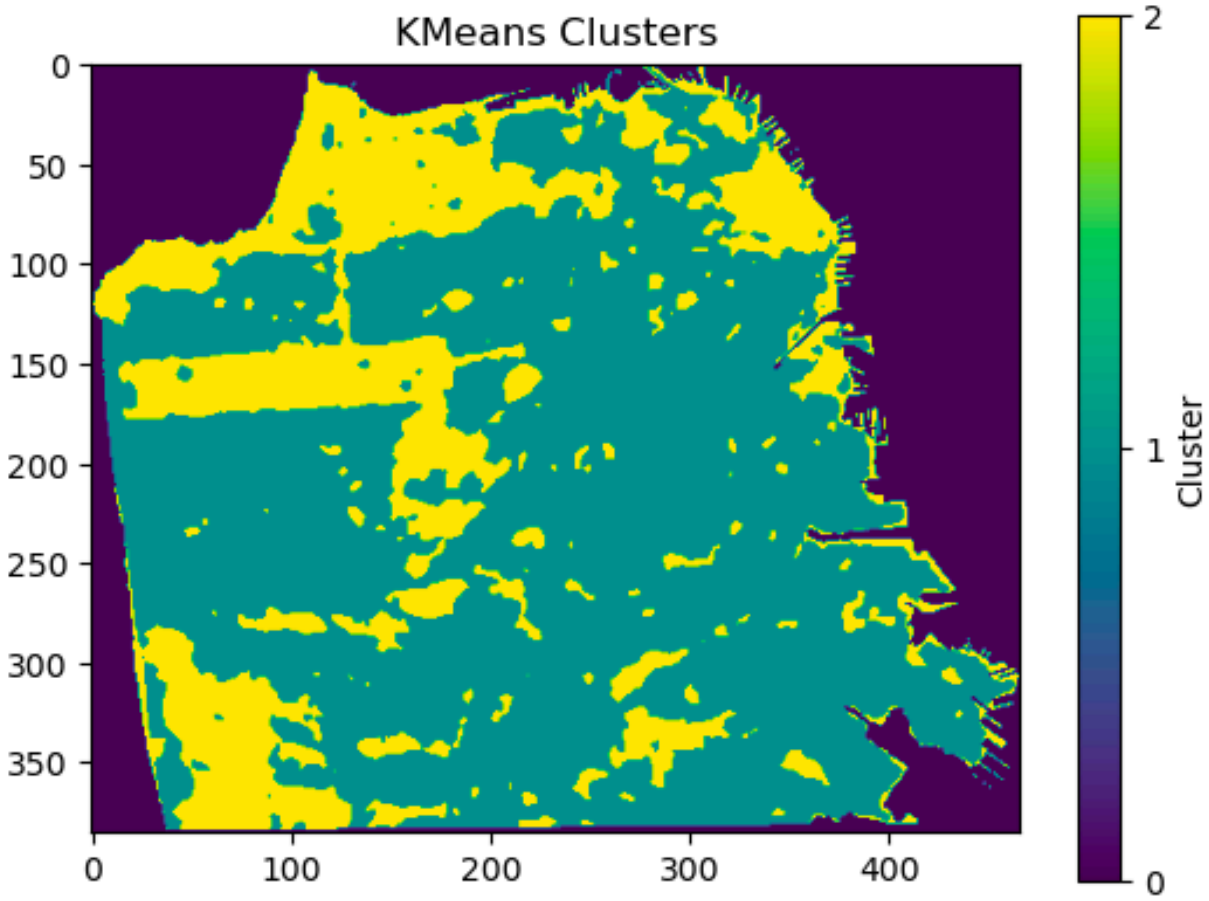
The vegetation indices (NDVI, NDBI, etc.) are calculated from the clipped raster and a custom mask is created based on the raster nodata values. The Heat Risk Index is calculated using the clipped vegetation indices and the assigned labels based on heat risk levels. The labels are assigned based on the heat risk index and plotted using Matplotlib. The colors in the plot represent different levels of heat risk, with blue indicating low risk, green indicating medium risk, and yellow indicating high risk. The mean value of the heat risk index is 19698.53, with a minimum value of -270.80 and a maximum value of 29885.79.





Methods

First I trained one unsupervised machine learning to identify clustering patterns. I used the KMeans algorithm to cluster the Heat Risk Index values into three clusters, which are represented by different colors in the resulting plot. The color bar on the right side of the plot



shows the cluster values (0, 1, or 2) and the title of the plot indicates that it shows the KMeans clusters.

In this section, a high-risk threshold is defined based on the calculated heat risk index, and binary labels are created to indicate whether an area is considered high-risk or not. The Heat Risk Index array is then reshaped along with the binary labels array to prepare for model training.

The Random Forest Classifier has an overall accuracy of 0.91, indicating that it correctly classified 91% of the high-risk and non-high-risk areas. The precision for class 1 (high-risk) is 0.86, which means that 86% of the areas predicted as high-risk are actually high-risk. The recall for class 1 is 0.42, which means that only 42% of the actual high-risk areas were correctly identified by the model.

```
Random Forest Classifier:
              precision    recall  f1-score   support

           0       0.91      0.99      0.95     48363
           1       0.86      0.42      0.56     5460

    accuracy                           0.91     53823
   macro avg       0.88      0.71      0.75     53823
weighted avg       0.91      0.91      0.90     53823
```

In Logistic Regression model, the precision for class 0 is 0.86, meaning that out of all the predicted class 0 instances, 86% are true class 0 instances. The recall for class 0 is 0.71, meaning that out of all the true class 0 instances, the model correctly identifies 71% of them. The F1-score is 0.78, which is the harmonic mean of the precision and recall. Overall, the accuracy of the model is 0.64, which means that it correctly predicts the class for 64% of the instances.

```
Logistic Regression Report:
              precision    recall  f1-score   support

           0       0.86      0.71      0.78     48363
           1       0.00      0.00      0.00      5460

    accuracy                           0.64     53823
   macro avg       0.43      0.36      0.39     53823
weighted avg       0.78      0.64      0.70     53823
```

In the SVM Classifier report, we see that the model has an overall accuracy of 1.00, indicating that it is able to correctly predict the class labels for both the high-risk and non-high-risk areas. The precision and recall values for both classes are also very high, indicating that the model has a very low rate of false positives and false negatives. This model is overfitted.

```
SVM Classifier:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00     48363
           1       1.00      0.95      0.98      5460

    accuracy                           1.00     53823
   macro avg       1.00      0.98      0.99     53823
weighted avg       1.00      1.00      1.00     53823
```