Anjuli Jain Figuera

01-20-2025

**PS1**

You are a quantitative analyst studying the environmental factors that contribute to SUHI disparities in marginalized communities. You gather environmental and demographic data. Task: Perform data exploration and model development. Your data exploration includes descriptive statistics, histograms, scatter plots, and correlation analysis. Your model development focuses on developing a tree-based model and gaining insights through SHAP values. Summarize findings for your team

Summary

# Data Exploration

## Spearman Correlation Analysis

*How does race and poverty correlate with environmental factors and SUHII? Which variables are most correlated with SUHII? How does this compare to the research discussed in the introduction? Describe any other interesting correlations.*

A diagram of a number of people

Description automatically generated with medium confidenceVariables that are highlight correlated (spearman correlation) with SUHI include: Spatial Lag with a high positive correlation (r2=0.78), NDVI with a high negative correlation (r2=-0.73) and NDBI with a high positive correlation (r2=0.68). These relationships make sense since we know UHI is high in urban settings and high values of NDBI mean higher likelihood of built-up areas and low (near 0) values of NDVI can mean urbanized areas (hence the negative sign of the r2). In fact, we calculate that NDVI and NDBI are inversely highly correlated (r2=-0.8). For demographic factors, it’s interesting to note that Black vs. UHI has a correlation of 0.4 while White vs. UHI has a correlation of -0.32 (i.e. different signs) These relationships are not as strong. Note also that population density vs. UHI has correlation of 0.48. The demographic aspects are less correlated than the environmental/urbanization aspects.

## Histograms

*What do the distributions reveal about race, socioeconomic status, environmental factors, and SUHII? How does this compare to the research discussed in the introduction?*

The mean SUHI is 2.97 degrees and seems like a normal distribution around it’s mean. NDVI and NBI also seem normal with means of -0.08 and 0.03. The demographic variables are highly skewed left since Black and Hispanic are minority groups and the percent of people below the poverty line are also skewed left (all with means of approximately 15% - see descriptive statistics for exact values). There are only 40 tracts that are all Black and these have a mean UHI of 6.3 degrees and 39.8% below poverty and mean population density of 1453; while the 97 homogenous tracts that have 100% white populations have a mean UHI of 2.7 and 15% below poverty line and population density of 579. The NDBI’s are closer for the “all black” vs. “all white” with mean values of 0.1 and 0.07 respectively. These all Black vs. all White tracts have a discrepancy in population density, poverty and UHI.

The spatial lag has a more complex distribution.

A group of blue graphs

Description automatically generated

## Scatter Plots

*Create scatter plots to visualize the relationships between demographic and environmental factors, using urban heat island intensity as a heatmap.*

In the plot below we used a diverging color scale, here UHI values that are red represent tracts where urban temperatures are higher than similar rural ones since UHI is the difference in temperatures between urban and rural. NDVI and NBI show separation in the UHI color scale (i.e. blue and red values) in the vertical axis. These can be interpreted for example for NDVI as the NDVI values increase towards 1, the UHI color becomes blue meaning the urban areas become cooler. This essentially relates vegetation to a cooling effect for urban areas. With respect to demographics, Hispanic and White seem to have some blue across representation but NDVI vs. Black shows almost no blue dots as you move towards 100% black tracts. This plot also “tapers in like a sideways hour-glass suggesting tracts that become 50% black or more mostly have NDVI values around 0 meaning few green leaves, unhealthy vegetation or more urbanization.

A screenshot of a graph

Description automatically generated

# Model Development

## Model Performance

*RMSE is the standard deviation of residuals. Based on the RMSE, how did the model perform? Based on the R2, how well did the model fit the data? How does model performance impact your communication of the model to your team?*

The Root Mean Squared Error (RMSE) is the average distance between a model predicted value and actual value in the dataset. The RMSE of the model is of 1.3 degrees when predicting UHI which in the dataset has an average value of 2.9 in a range of -13 to 14 degrees. The R2 is 0.88 where values of 1 are a perfect fit. The model is seemingly capturing 88% of the UHI variance, giving reasonable confidence to communicate this model to the team.

## Predictive Factors of Urban Heat

*Using the SHAP values, discuss the most predictive factors of urban heat intensity. Why do you think these are the most predictive factors? How does this compare to the previous research discussed in the introduction?*

Spatial lag is the most important predictive factor followed by NDVI and population density. This aligns with research which touts the cooling impact of trees/vegetation and correlations population density with urban areas. In the bee swarm, the coloring by feature shows us patterns: for example NDVI and BSA both have a distinct RED/Blue color gradients (population density on the other hand has a lot of blue all across) -- this red/blue gradient means lower NDVI's increase UHI, while BSA is reversed, higher BSA, increase UHI. The gradient can also be appreciated with NDBI and to some extent poverty.

A screenshot of a computer

Description automatically generated

## Multi-Dimensional Analysis

*Using SHAP scatter plots, discuss how racial composition of a tract and environmental factors impact SUHII (i.e., use environmental factors as a heat map by race). Do the same for poverty status and environmental factors. Explain the disparities that these plots illuminate. How does this compare to the previous research discussed in the introduction?*

Shapely value comparison for NDVI with a colormap by Race – White (left) and Black (middle) and shapely value of NDBI vs White (right). The left and middle plot shows an approximately linear and negative trend between NDVI and UHI while the right plot shows a a non-linear relation between the NDBI contribution to the model predicted value. The coloring by race “White” shows areas thar are low white percentage and high NDBI values as a higher contributions but trending down.

A pink and blue dots

Description automatically generatedA blue and pink particles

Description automatically generated A diagram of a number of dots

Description automatically generated with medium confidence

Poverty status vs environmental factors – these plots show an approximately positive linear trend between poverty status and UHI model prediction meaning the presence of higher percentage of people living in poverty in the tract is a contributing factor to predicting higher UHI values. When these contributions to the model are color-coded by the environmental factors, NDVI and NDBI seem to have areas of reds or blues.

A screenshot of a graph

Description automatically generated

Higher NDVI’s (more vegetation) with higher proportion of white residents is related to the modeled UHI by reducing the temperature predictions.

## Summary

*Summarize the findings for your team. Based on your analysis, how do environmental disparities by race and socioeconomic status influence SUHII?*

**Key Findings**

Many match what is in the literature and previous research:

* Urban or built-up land area (measured by NDBI and inversely related to NDVI) is highly correlated (r2=0.8) to UHI.
* Tree cover (measured by NDVI) has a cooling effect reducing UHI temperature differentials for urban areas relative to rural areas.
* Black populations have access to less green/vegetated spaces and experience UHI across more tracts.
* We are able to fit a machine learning model using XGBoost to the data and have confidence in the model’s ability to predict UHI since the model has an R2 of 0.88.
* NDVI is an important feature (after Spatial Lag) where more vegetation means lower UHI and there is less green vegetation in neighborhoods that have higher proportion of Black residents.