Ayush Lahiri

Whitepaper: Predicting Urban Heat Island Effect based on Spatial Features

---

Research Question: What is the impact of the spatial distribution of buildings, street lines, and vegetative cover on the Urban Heat Island (UHI) intensity on a geographical unit and can the visual distribution of these urban characteristics gained through geospatial representations, be utilized to accurately predict UHI effects, using Neural Network based architectures?

Introduction

Urban Greenspaces have been extensively linked to numerous benefits for populations residing in urban spaces including human health and well-being by providing opportunities for active recreation and play, building restorative capacities, supporting social and community well-being, and mitigating the negative health effects of noise, air pollution, and heat exposure (Kajosaari et al., 2024). Further, they form a key component of sustainable urban planning and development, as increasing demands for urban spaces (and subsequently their economic benefits) come at odds with environmental preservation. The consequences of these trade-offs are particularly evident in two areas: local pollution levels and surface temperatures. Regarding surface temperatures, it is well-documented that densely urbanized areas, filled with impervious structures such as concrete roads and buildings, and a scarcity of vegetation, result in elevated surface temperatures compared to nearby regions with lower urban density (Chi, Guangqing, as cited in Bohn, 2023).

This effect is known as the Urban Heat Island Effect. It is best described by the United States Environmental Protection Agency (2014):

'*Heat islands are urbanized areas that experience higher temperatures than outlying areas. Structures such as buildings, roads, and other infrastructure absorb and re-emit the sun's heat more than natural landscapes such as forests and water bodies. Urban areas, where these structures are highly concentrated and greenery is limited, become "islands" of higher temperatures relative to outlying areas. Daytime temperatures in urban areas are about 1–7°F higher than temperatures in outlying areas and nighttime temperatures are about 2-5°F higher.*'

The increase in Urban Heat Island (UHI) intensity has significant repercussions, affecting health directly and exacerbating air and water quality issues, as well as impacting native species and their migration patterns (National Geographic, 2022). Moreover, the adverse effects of UHIs in the United States are distributed unevenly, with people of color more often residing in areas experiencing more intense UHI effects (Hsu et al., 2021), leading to disproportionately negative lifestyle outcomes.

Currently measurement of UHI effect is done through a deterministic algorithmic processing of temperature readings (Chakraborty et al., 2020). As a result, UHI can only be observed ex post, i.e. unless an urban area is planned, populated, and observed over multiple seasons and years, the extent of UHI effect in the area cannot be effectively known. The implication is, a populace must endure the effects for a certain number of years, before any information regarding UHI can be reliably known. Further, rectification through re-planning urban areas to reduce UHI effect is a time consuming, costly, and often unfeasible process that can also suffer from unequitable attention to disadvantaged communities.

Current prediction models utilize aggregated data for geographical units, employing features like average tree canopy coverage within a census tract (Assaf et al., 2023) and similarly infer the spatial distribution through aggregated metrics, such as vegetation types within a geographical area (Oh et al., 2020). However, these aggregated measures, despite their significance, fall short of capturing the true 'spatial' distribution of an urban unit, given the inherently visual nature of spatial data. While two urban areas of the same size, can have the same number of buildings, whether the buildings are clustered together or spread uniformly may impact surface temperatures (Wang & Xu, 2021). Capturing such details, requires a visual inspection of the urban area being assessed, a task, which on a large scale, is well suited to computer vision methods and machine learning.

To the best of my knowledge, there are currently no modelling efforts that aim to predict UHI effect using computer vision-based methods - an area this study seeks to address, by using vision-based machine learning methods like CNNs on spatial datasets about urban features such as buildings, streets, and vegetative cover to predict UHI effects.

A method to accurately estimate the UHI effect of an urban area has its primary utility in preemptive urban planning. At the planning stage, with theoretical layouts of future urban areas, being able to predict the potential UHI effects, allows for more effective city planning before investments are made. Further it would help reduce the negative effects faced by the area's populace during the 'UHI data collection' time period, as required by existing deterministic estimation processes for UHI effects. Further while extensive UHI estimates exist for various urban areas in the US, that may not be the case for many densely populated developing countries, or conflict regions where availability of recent estimates and accuracy may be lower.

Data

In line with the above outlined endeavor, my model aims to predict UHI effects by considering spatial distribution of urban characteristics, focusing on are buildings, street lines and vegetative cover and their relative distribution's and densities in a geographical unit. I experiment with two different datasets, both representing the spatial distribution of urban special features, albeit with significant differences in the display mode, overall information available in each sample, and observational unit. Both datasets covers three major cities in the United States, namely Chicago, New York City and Los Angeles.

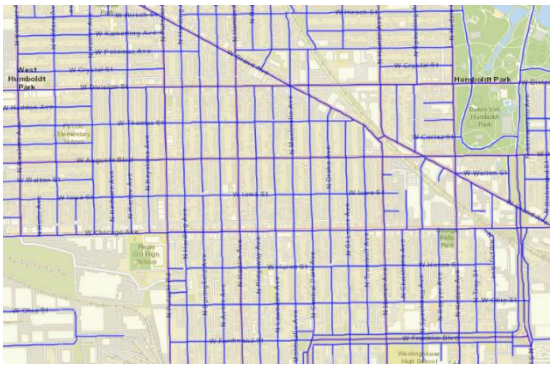
Dataset A: Segmented Urban Saptial Features

Input: The input data for analyzing buildings and streets sourced from three distinct geospatial datasets provided by the city and green coverage provided Multi-Resolution Land Characteristics (MRLC) consortium of federal agencies.

- **Buildings**: Firstly, we use the Building footprints in Chicago (City of Chicago, 2023), New York (Office of Technology and Innovation, 2024), Los Angeles (Los Angeles GeoHub, 2017) as unique shapefiles containing the exact geographical position, shape and size of each building in a city.
- **Streets:** Secondly, we source the street center lines data in Chicago (Levy, 2024), New York (Office of Technology and Innovation, 2024), Los Angeles (Los Angeles GeoHub, 2020) as unique shapefiles containing the exact geographical position and shape of major streets and roads in a city.
- **Greenspace cover**: To extract information regarding vegetative cover, or specifically in my case Tree canopy cover I use the 2021 CONUS Tree Canopy cover aerial imagery (Housman et al., 2023) in a Tiff format. This raster file contains pixels at a highly detailed 30-meter resolution, with each pixel value representing the density of green cover, along with accurate geographical position of each pixel.

Panel 1. (a) and (b) provide an example of what buildings and streets dataset in their raw form look like

(a) Segmented Building Vectors in Chicago



(b) Streets and Roads Vector in Chicago

Panel 1. Raw Input Data Examples

Processing Methodology: Using the boundaries provided by the US Census Bureau (2018) for census block-groups in each city, I iteratively subset each GIS dataset to the boundaries of a unique census block group and overlay the raster, buildings, and roads data into a composite visual output. Each input image thus corresponds to the spatial characteristics of a census block-group, across three cities, representing my observational unit. The resulting composite geospatial dataset offers a detailed visualization of Chicago's urban landscape, highlighting buildings, streets, and vegetative cover simultaneously while eliminating possible noise that may exist in composite aerial images as seen in figure 1.

Figure 1. 4-Band Aerial Image of Chicago

Census block-groups as a unit of analysis provide an appropriate middle ground in area covered between the highly localized census blocks and much larger census tracts, allowing us to focus on local UHI affects in the city while ensuring that the area under consideration is large enough for there to be variation amongst the data instances. Thus, samples can have different shapes and sizes and areas outside of the shape of each census block for a sample, is exported as a transparent layover png file.

The final input then fed into the model, would be an image as represented in panel 2. The segmented buildings are presented in light red, the green areas represent tree canopy coverage in the block with the darker hue's representing denser tree coverage, and the blue line segments represent streets and roads.

Panel 2. Dataset A final input examples

Output: The output data consist of UHI values estimated using the Simplified Urban-Extent (SUE) algorithm by Chakraborty et al. (2020), released by the Yale Center for Earth Observation. These estimates represent the temperature difference between urban and surrounding rural areas, indicating the extent of the UHI effect. The data are derived from MODIS imagery and are available at a spatial resolution of 300 m (Chakraborty et al., 2020), available in GeoTiff format. This dataset provides various types of estimates such as diurnal, seasonal, monthly, yearly, by daytime and nighttime and long-term variability across various climate zones.

Processing Methodology: I find the mean of the annual average daytime UHI and night time UHI effects for each coordinate and discard coordinates where UHI value is missing or less than -800. The latter represents values for large bodies of water such as Lake Michigan. Since my selected features of buildings, roads, streets and tree canopy cover do not contain the information required to represent water bodies, for this dataset I remove such instances.

Subsequently, following a similar methodology as the input data, I use census block-group boundaries and extract the average pixel values representing the UHI effect in the census block-group. Since these average pixel values representing UHI correspond the boundaries of a unique census block group, I can use these coordinates to accurately match my inputs and outputs.

Table 1 discusses the details the descriptive statistics of the dataset A.

Parameter	New York	Los Angeles	Chicago	Overall
Total Block Groups	6,623	6,589	2,328	15,540
Total N : (Block Groups with UHI Data and complete features)	1,285	898	747	2,930
Tree Canopy Cover				
Mean	47.21	82.56	17.96	49.24
Max	255	269	254	269
Min	-38.38	-16.64	-20.55	-38.38
Urban Heat Island Effect				
Mean	-2.38	0.79	0.64	0.01
Max	2.80	3.49	2.93	3.49
Min	-51.90	-46.49	-32.34	-51.9
Number of Channels in Image		3		
Raw Image Size		1000×1000		

Table 1. Dataset A: Descriptive Statistics

Discussion

Complete data is available for 2930 census blocks groups (of the total 15,540 total available) constituting our full sample. This limitation arises from the range of UHI sensors in Chicago, which do not align perfectly with census

boundaries. Secondly, the data also excludes areas with a margin of error exceeding 3°C (Chakraborty et al., 2020). Finally, we also exclude data for census block groups where we have incomplete features in terms of our geospatial features such as buildings, roads and greenspace.

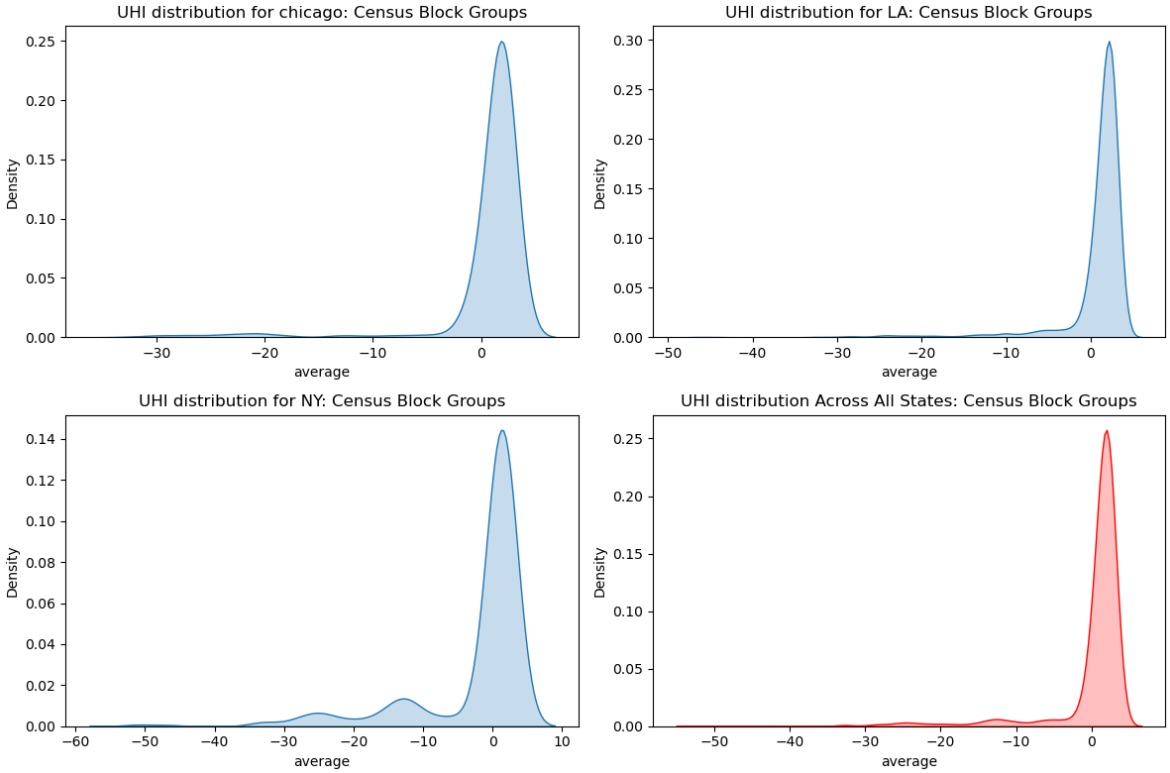


Figure 2. UHI distribution in dataset A: Observational Unit is census block group

As we see in figure 2., the data also exhibits significant right skewness, predominantly ranging between -1.5°C and 3°C — falling with the common distribution for UHI, typically extending from -4°C to 7°C based on geographical factors and in line for our cities in this case (Global Surface UHI Explorer, 2020). We find that New York has thicker tails in the negative UHI region. As we discuss later, this selection is deliberate and helps us ensure greater balance in the data.

Secondly, inspection of sample by state is necessary for Dataset A, due to the nature of the images and how census blocks are defined in each city. For example, I find through visual inspection, that census block groups in LA are generally much larger than NY and Chicago. This is equivalent to saying, that the images of block groups in LA are more likely to be 'zoomed out' thereby the shape and size of buildings appearing much smaller and roads appearing less pronounced.

This pattern of variation in the size/zoom level existing in all three states (more pronounced in LA), can thus indicate that our model may have a difficulty finding the average patterns present in an image that predict UHI and/or incorrectly associate the size of the buildings, roads or size of the census block group with UHI. This may be driven due to our overall training sample being small, and also insufficient from the perspective of having enough sample that would allow a discriminated learning of patterns at different zoom levels.

Dataset B: Composite Aerial Images

Input: The second dataset, we utilize for to find visual representation of urban spaces is composite aerial images of cities. Like Dataset A, this dataset also covers New York, Los Angeles and Chicago.

- Specifically, we use the city shape boundary and overlay a self-generated grid over the region, with each cell in the grid denoting a 500m X 500m region of the city.
- We then use the coordinates representing the limits of the cell, i.e. the corners of the cell to find the areas of the city contained within the boundaries of the latitude and longitude of each cell's corner.
- Next using Google Map's static API, we retrieve the aerial map image of the region contained within these boundaries
- Due to specification of all four corners of the cell in our request to the GMS API, with all four corners of each cell designed to represent the same area, all images in the dataset follow the same zoom level.

The final input then fed into the model, would be an image as represented in panel 3

(a) **(b)**

Panel 3. Dataset B final input examples

Output: Similar to Dataset A, we continue using the same source for our output UHI effects and use the bounding box coordinates of each cell to extract the respective average UHI effect in the bounding box.

Unlike the approach taken for dataset A, where observational units with average UHI less than -800 were removed, I refrain from doing so for this dataset. This decision is informed by the understanding that the aerial images contain all possible information available within the urban environment, including areas with extreme UHI values.

I do not apply additional preprocessing to the retrieved images, in order to highlight the contrasting levels of signal present within them. Specifically, my segmented images dataset is curated to represent an urban area with all extraneous noise removed, focusing solely on relevant features. In contrast, dataset B comprises aerial images that encapsulate the full spectrum of information available within an urban setting

Table 2 below provides relevant descriptive statistics for dataset B (aerial images).

Parameter	Value
Total 500m X 500m cells	6,184
Total N : (Block Groups with UHI Data)	5704
Urban Heat Island Effect	
Mean	-82.53
Max	3.53
Min	-898.49
Number of Channels in Image	3
Raw Image Size	600×300

Table 2. Dataset B: Descriptive Statistics

Discussion

We have data for 5,704 unique 500m x 500m cells across all three cities, out of a total of 6,184 cells, indicating a higher number of samples in dataset B compared to dataset A. This difference is due to two main factors. Firstly, each observational unit in dataset B is a uniform 500m x 500m area, which is generally smaller than the varying sizes of census block groups in dataset A, particularly in cities like Los Angeles where block sizes can be much larger. Secondly, dataset B includes all data, regardless of extreme UHI values, because it uses aerial images that provide a complete view of urban features. Conversely, dataset A uses segmented observations where specific features are isolated and noise is removed, and hence required omissions of samples without the needed features to explain extreme UHI values such as bodies of water.

Examining the distribution of the UHI effect in dataset B in figure 3, we observe a general right-skewed distribution with most data points concentrated between -1.5 and 3, peaking at 0, similar to dataset A which is expected since our data source remains the same. The inclusion of samples with large negative UHI values contributes to more negative average and minimum values in this dataset.

Since, the observational units in this dataset cover a smaller area, the average UHI effect for each cell is computed over fewer coordinates. This reduction in the number of coordinates per cell means that the averaging process is less diluted by a large volume of UHI values, which could explain the higher maximum values observed in this dataset compared to dataset A , likely accounting for the higher maximum values observed in dataset B. Figure 2 illustrates this with fatter tails in the distribution, particularly noticeable for New York.

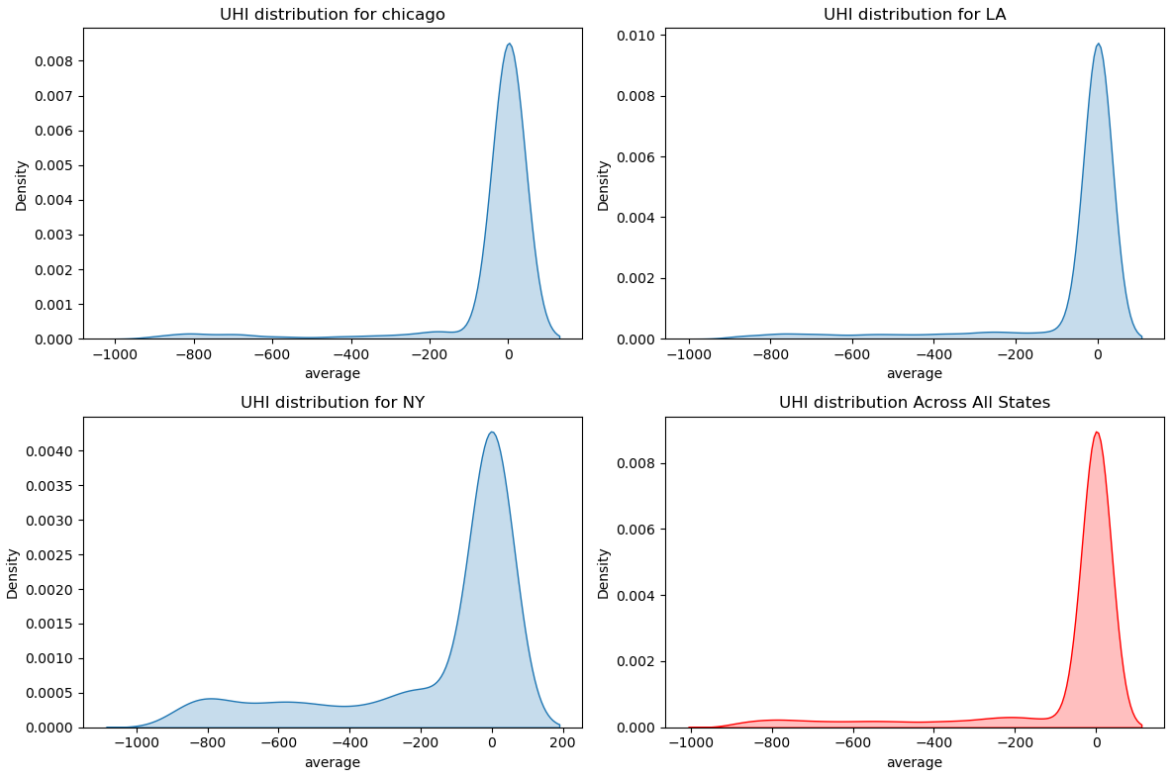


Figure 3. UHI distribution in dataset A: Observational Unit is census block group

Methodology

(A) Data Processing

A.1 Rationale for classification over regression

1) To effectively process the Urban Heat Island (UHI) effect data for classification rather than regression, we consider the nuances of UHI values as continuous data. In its raw form, the average UHI of each census block group in dataset A and 500m X 500m cell in dataset B is a continuous value. The significance of variations between continuous UHI values varies greatly across the scale. For instance, **a transition from a UHI of -800 to -700 often does not imply a significant change in urban heat**, and similarly, UHI values such as -700 or even -200 generally indicate a low UHI, which might not prompt urgent policy decisions or design changes. Further **these values are influenced by more global features such as proximity to large water bodies and geographical positioning of the region - factors out of control for planners**. In contrast, the **variations between UHI values closer to zero, such as from -1.5 to 0 to 1.5 to 3, are much more critical**, with UHI values generally not exceeding 7, and in our dataset capped at 4.5 (here we are discussing the UHI value of each coordinate in our raw UHI data and not the average UHI value we find for an observational unit). These differences can markedly affect how an area is perceived and managed in terms of urban heat, with **an area having a UHI of 1.5 being significantly worse off than an area with UHI of -1.5 and in turn are a product urban planning decisions**.

   The minor shifts in UHI values, such as from 1.5 to 1.1 are thus not trivial and can indicate significant differences in environmental and urban characteristics, such as lower population density, increased green space, material of buildings, type of vegetation etc. To capture these critical, yet subtle, variations accurately on a continuous scale, it is necessary to utilize multiple high-resolution data sources, such as 3D LiDAR for detailed mapping of building structures, along with supplementary multi-modal data that includes population density, types of green spaces, and building materials, are essential.

   My **objective in this study is to identify major trends in UHI that will assist urban planners in making informed decisions during the initial stages of the planning process, during which emphasis will be identifying the most significant differences**, i.e. whether the UHI effect falls within the -1.5 to 0 range or the 0 to 1.5 range, for example. Identifying the numerically minor but substantively major trends, such as the difference between 1.5 and 1.1 on a continuous scale, typically pertain to more detailed aspects of urban planning that become relevant in later stages. Further limited data availability, for example 3D LiDAR of buildings and computational resources such data necessitates, also informs my decisions to approach this as a classification task.

By categorizing the continuous UHI values into discrete buckets, I concentrate on the broader, more immediate trends that are affected by broader spatial characteristics of an urban area such as building density and distribution (over more detailed characteristics such as building material and height required for the regression task).

A.2 Creating Classes

1) I first undertake categorization by creating **six buckets** namely: less than -1.5, between -1.5 and 0, between 0 and 1.5, between 1.5 and 3, more than 3. However, **this created significant imbalance**

between the classes, dominated by the between 1.5 and 3 class being nearly 4 times the smallest class of more than 3. Combined with a small dataset of 2930 images, there simply isn't enough training samples for each class to learn generalizable patterns. As I found later **during model experimentation, this led my model to predict all cases as the dominant classes for which enough samples were available**, i.e. between 0 and 1.5 and between 1.5 and 3. I thus decided to **merge classes into 3 broad classes, namely – less than 0, between 0 and 1.5 and more than 1.5, to create balance and greater generalizability**.

2) For the **experimentation with dataset A, I chose to exclude all cases from Los Angeles**. This decision stemmed from the observation that the **census block groups in Los Angeles are significantly larger than those in New York and Chicago**, which adversely affected the model's generalization capabilities. In samples of LA, features such as buildings, appear disproportionately small relative to the total area of the blocks, with green spaces dominating these images, compared to the relatively low frequency of such large blocks in Chicago and New York. This **leads to high variation within and across classes and the model has difficulty learning generalizable features**.

3) This **not an issue for dataset B consisting of aerial images, due to all images being at the same zoom level** and hence one feature is unlikely to varyingly dominate the other due to issues in spatial representation. Thus in dataset B I do not drop any images, but undertake similar merging of classes due to the imbalance existing in dataset B as well.

Table 3, highlights the features of the datasets used as final inputs to my models and the key differences between dataset A and dataset B.

Dataset A	Dataset B
Final N: 2032	Final N: 5704
Less than 0: 331 Between 0 and 1.5: 478 More than 1.5: 816	Less than 0: 1,543 Between 0 and 1.5: 1,384 More than 1.5: 3,156
Cities Covered: New York, Chicago	Cities Covered: New York, Los Angeles, Chicago
Each image only contains information about tree canopy cover, roads and buildings.	Each image contains all information captured in a raw aerial image of a city.
Each image shape is different, following the shape of a census block group with different zoom levels.	Each image shape is square representing an area of 500m X 500m of equal size, corresponding to a consistent grid over the city with the same zoom level.
Census block group sizes are highly variable equivalent to varying zoom levels, i.e., bigger the census block, lower the zoom level.	Grid sizes are a constant, smaller observational unit at 500m X 500m, than the census block group, leading to larger N.
Source: Various GIS files of tree canopy cover, roads and buildings available from each county's data repository.	Source: Google Maps Static API.

Table 3. Final composition of and differences between dataset A and dataset B

Note on Dataset Size: One important takeaway, that emerges in our model results as well, is that these **datasets are simply too small, especially dataset A**. While I would like to include more cities, **the retrieval and processing of the final segmented images, that involves joining multiple spatial datasets and subsetting them in R are highly computationally expensive processes that took 2 days+** dependent on the city on the personal computational resources available to me. Google Maps Static API is paid service and hence I could cover more cities than those available, due to resource constraints. As I discuss on my final section, I recognize this as a major limitation for my project.

A.3 Data Splits

I undertake an 80/10/10 training, validation, and testing split. Given the imbalance of my classes I undertake stratified sampling to ensure that each subset maintains a consistent proportion of classes as found in the original dataset, thereby preserving the distribution, and ensuring representative samples during model training, validation, and testing phases. This stratified approach helps mitigate the risk of bias and improves the model's ability to generalize across the true range of classes in the dataset.

- For dataset A, my models are trained on a training set of size 1625 images and validated and tested on sets of 203 and 204 images each.
- For dataset B, my models are trained on a training set of size 4494 images and validated and tested on sets of 606 and 604 images each.

(B) Modelling

I use a Convolutional Neural Net architecture to predict the average UHI for each observational unit. Given my inputs are images, identification of the edges separating impervious surfaces like roads and buildings from green spaces and the areas contained within these images are the fundamental features that need to identified to predict UHI. A CNN is the best suited architecture for pattern extraction from images and finding relationships with the target features. This is due to the architecture's ability to convolve over the entire image and extract topological features in a hierarchical manner, progressing from low-level to high-level level details. Over multiple iterations, the dense representations of full images in the intermediate layers of the network become more representative of the features in the image which are most useful in accurately predicting the outcome through tuning of weights and biases.

For both dataset A and B, I train identical sets of models with respect to their architectures, only varying the input size of images (180x180 for dataset A and 300X300 for dataset B) and hyperparameters such as learning rate.

Specifically, the models I experiment with are as follows:

1) **Unstacked Custom CNN 1**: For dataset A it is designed to process 180x180 RGB images, and for dataset B it is designed to process 300X300 RGB images. The architecture consists of a sequence of five convolutional layers with increasing filter sizes—32, 64, 128, 256, and 256—and corresponding max pooling steps to reduce dimensionality. This is followed by flattening and a final dense layer with three output neurons using a softmax activation to classify the images into three categories.

2) **Stacked Custom CNN 2**: With similar input configurations as CNN 1, this architecture involves a sequence of 7 convolutional layers, stacked in pairs aimed at extracting greater detail. We start with two convolutional layers each with 32 and 64, followed by another set of convolutional layers with increased filter counts: 128 and 256. After each pair of convolutional layers, a max pooling layer is used to reduce

the dimensionality. We continue with another pair of convolutional layers of 256 filters each, followed by a layer with 512 filters. After pooling step and flattening step, a final a dense layer with three output neurons and a softmax activation function is used.

3) **Custom CNN with image augmentation**: I also undertake image augmentation in one of my custom CNNs. Depending on the performance of the models without augmentation, I select the higher performing model for each of the two datasets I experiment with, and introduce an additional layer of augmentation. This includes random horizontal flipping, rotation by up to 10%, and zooming by up to 20%. This process does not increase the size of the dataset; instead, it augments images randomly during training. The model thus sees different variations of the same images, which helps in preventing overfitting by ensuring that the model does not merely memorize the training data but learns to generalize from varied representations of data. In implementing these augmentations, I maintain the original architecture of the CNN, only integrating these augmentation layers.

For all three custom CNNs, the convolutional layers use a ReLu activation function, a 3X3 kernel size and input values ranging from 0 to 255 in their raw form, are re-scaled to be between 0 and 1.

I also use two pre-trained architectures, namely:

4) **VGG-16 (Simonyan & Zisserman, 2015)**: I utilize the VGG-16 architecture, a widely recognized CNN developed by the Visual Graphics Group at Oxford known for its deep yet straightforward architectural design. I employ the Keras implementation of VGG-16 pre-trained on ImageNet. The primary architecture of VGG-16 consists of 16 layers, including thirteen convolutional layers with 3x3 kernels, followed by five max pooling layers interspersed after certain convolutional blocks to reduce spatial dimensions progressively. It is particularly advantageous for a wide range of image recognition tasks because it has been extensively trained on the ImageNet dataset, which comprises over a million labeled images across a thousand categories.

   The model is configured to retain the model's base pre-trained weights, and I replace the top layers with a configuration suitable for my purpose. After the inputs with random augmentation as in model 3, are passed through the fixed base, it is followed by a flattening layer. This vector feeds into a dense layer of 256 neurons, followed by an output layer consisting of three output neurons using a softmax activation.

5) **Google Inception V3 (Szegedy et al., 2015)**: I also experiment with Google's Inception V3 CNN that stems from Google's earlier Inception models. It advances the original architecture by incorporating 'modules' that allow it to efficiently learn representations with fewer parameters. Each module in Inception V3 uses a combination of filters of different sizes operating at the same level, which enhances the network's ability to capture information at various scales. Furthermore, Inception V3 introduces factorized convolutions and expansions of the filter bank outputs which reduce the computational burden while increasing the network's depth and width. I employ the Keras implementation of Inception V3, pre-trained on the extensive and diverse ImageNet dataset. The architecture includes symmetrical and asymmetrical building blocks, including convolutions, average pooling, max pooling, concats, dropouts, and fully connected layers, all optimized for performance and lower computation costs.

   For my specific application, I keep the base model of Inception V3 with its pre-trained weights untrainable to capitalize on the pre-learned features, while replacing the top portion of the network with a configuration suitable for my purpose. Following the input processed through the unchanged base, the

architecture leads into a flattening layer, followed by a dense layer of 512 neurons with ReLu activation, culminating in a softmax output layer with three neurons.

- During, I employ two key Keras callbacks to optimize performance and prevent overfitting. First, the *ModelCheckpoint* callback is used to save the model only when its performance on the validation set improves, specifically monitoring the validation loss to determine the "best" model.

  The choice to **prioritize the lowest validation loss over the highest accuracy as the criterion for the best model is driven by a focus on ensuring confident classification capabilities**, i.e whether the probabilities of the dominant class are substantially higher compared to a more uniform distribution across classes, the latter suggesting uncertainty in the model's predictions. This is quantified through the cross-entropy loss, utilized across all models. We **value this over accuracy percentages, which can sometimes reflect random fluctuations within an epoch, especially in unstable training**.

- Second, the *EarlyStopping* callback monitors the validation accuracy, ceasing training if there is no improvement for six consecutive epochs. This callback not only halts training to save computational resources and helps in discouraging overfitting, but also restores the weights of the best-performing iteration once the stopping condition is met.

- Finally for **Inception V3 on dataset B (aerial images), I undertake L2 regularization** in the top dense layer and adaptive learning rate decay to dampen issues of overfitting.

Results and Model Evaluation

Table 4, below reports the test accuracies of all the models I experimented with and their test accuracies across both datasets

Model	Accuracy (Dataset A)	Accuracy (Dataset B)
Custom CNN (5 conv layers, filters: 32-256)	52%	68%
Custom CNN Stacked (7 conv layers (3 pairs stacked), filters: 32-512)	49%	67%
Custom CNN unstacked with Data Augmentation (5 conv layers, filters: 32-256, input augmented)	59%	68%
VGG-16 (Base layers frozen, input augmentation)	47%	66%
Google Inception V3 (Base layers frozen)	54%	**70%**

Table 4. Test Accuracy across dataset A (segmented images) and B (aerial images)

After extensive experimentation with hyperparameters and various architectures, it appears that none of the models are particularly effective at extracting meaningful information from Dataset A. The highest accuracy recorded was 55% with the 5-layer unstacked CNN that included data augmentation. This performance, only slightly better than a random coin flip, underscores the limited predictive power of the models for this dataset. In fact, two of the models exhibited accuracies below 50%, suggesting that they performed worse than random guessing, which indicates that these models are unable to find any signal in the data and are likely fitting noise.

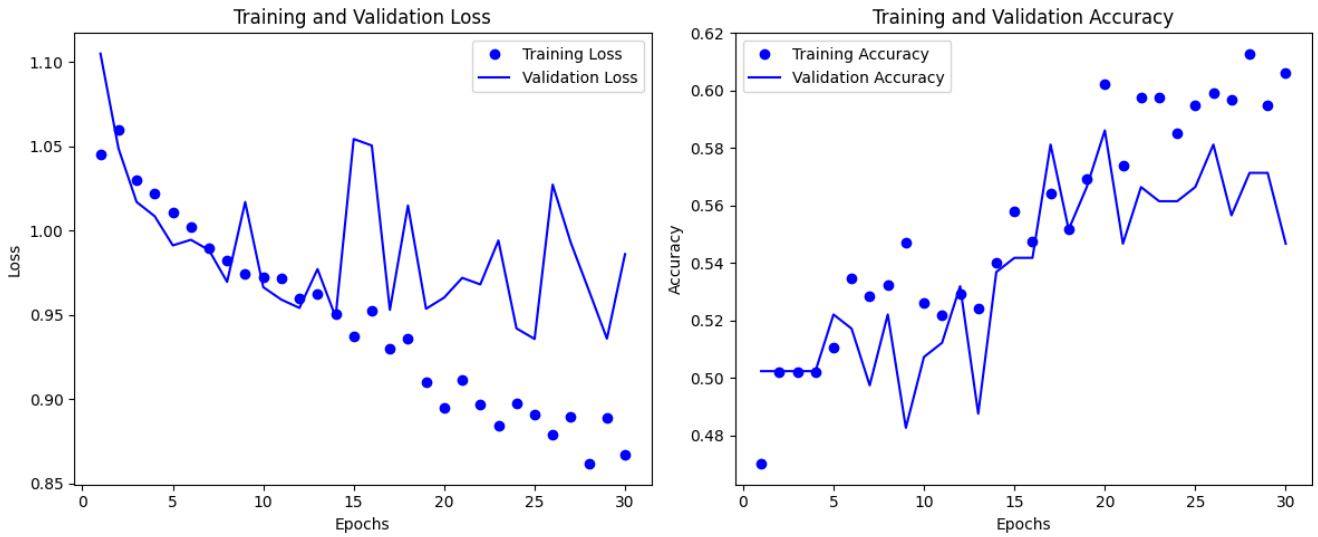
For Dataset B, there is a significant improvement in model performance, with accuracies consistently above 67%; notably, the Inception V3 model achieved 70% accuracy. While these results are not optimal, they are

markedly better than those for Dataset A. The consistent performance across different architectures for both datasets suggests a fundamental disparity in the signal-to-noise ratio between them. This difference likely accounts for the varying model efficacies, emphasizing that the primary variable affecting performance is the inherent quality and characteristics of the datasets themselves.

Note: We next discuss the training and validation accuracy and loss patterns only for the best performing models on both the datasets, i.e. custom conv 1 with 5 unstacked layers and augmentation for dataset A and inception V3 for dataset B. For conciseness, I focus on the training patterns of only the best-performing models for both datasets, although other models exhibit very similar trends in terms of the deficiencies we see for the best performing ones, only much more pronounced; overfitting and training instability being the primary issues, I discuss in more depth next.

Note: Number of epochs may vary due to early stopping condition of no improvement for 6 epochs.

<u>Discussion: Evaluation of best performing model with dataset A – spatially segmented images</u>



(b) Training and Validation Loss **(a) Training and Validation Accuracy**

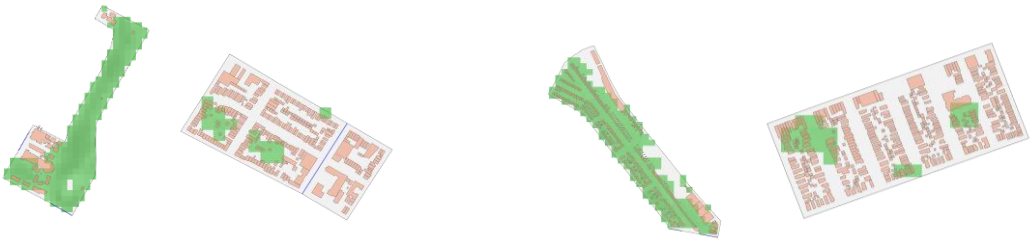
Panel 4. Training Custom 5 conv with augmentation: Dataset A (segmented images)

Observing the performance of the 5-layer CNN with augmentation on Dataset A in panel 4, two primary trends are notable. First, the model begins to overfit around the 20th epoch. However, this overfitting is not substantial, as the training accuracy only reaches a maximum of about 63%, while the validation accuracy remains in the 50-58% range. This suggests that the model is unable to effectively capture the signal in the data, indicated by the relatively low training accuracy and corroborated by the increasing loss. As we saw in Table 4, more sophisticated architectures and powerful pre-trained models also fail to find signal in the data, suggesting that the simplicity of the architecture is not the primary issue. It is more plausible that the data itself is noisy.

Secondly, the training is highly unstable. This instability is likely an artefact of both a very training sample of around 1500 images and primarily indicates that the model is contending with noisy data, and the features it learns are not generalizable across different samples within the same class. Specifically, the model may encounter significant variability in the features of samples from the same class, or it may find that features of different classes closely resemble each other. Panel x, represents an example of many such instances. On the left, two samples with a UHI effect of less than 0 feature distinctly different landscapes—one is dominated by green spaces

with few buildings, as expected for a low UHI effect, while the other is dominated by buildings with minimal green spaces. Conversely, on the right, samples representing a UHI effect greater than 1.5 show nearly identical images to those on the left, despite belonging to the extreme opposite class.

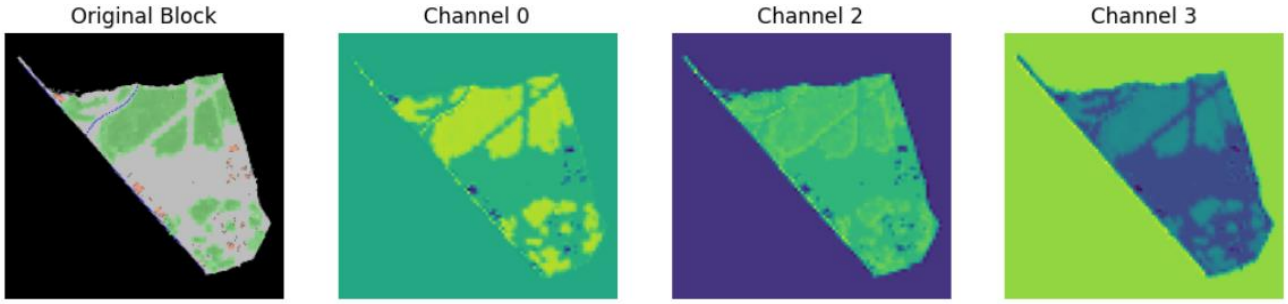
As noted previously, dataset A is a small sample overall, with less than 1000 training samples for each class, the lowest being 300. With not enough training samples to learn the variability within a sample and the possible similarity across samples, the model gets conflicting signals, leading to difficulties in accurately discriminating between classes. The model adjusts its parameters based on one set of data, only to find these adjustments counterproductive in the next epoch.



(a) UHI less than 0 **(b) UHI more than 1.5**

Panel 4. Within class variation and across class similarity: Dataset A (segmented images)

Further, this issue is most likely exacerbated by the varied shapes of each census block group when used as observational units. Features should ideally be derived only from within the area within a census block group. The dataset in its raw form, has images with the background area around the shape of the census block as transparent. When fed as input to the model, the shape of these transparent images gets standardized into rectangular images by adding black spaces around them. This is likely misleading the model as it could interpret the varying areas and distributions of these black regions as relevant features. As observed in the feature map in panel 4, the model is considering these non-informative black areas as part of the feature space, and is thus most likely inadvertently fitting to noise, which is not representative of the underlying data characteristics.



Panel 5: 5- Conv CNN with augmentation: Dataset A - Feature Map (layer 2)

Conclusion and Limitations of CNN with dataset A

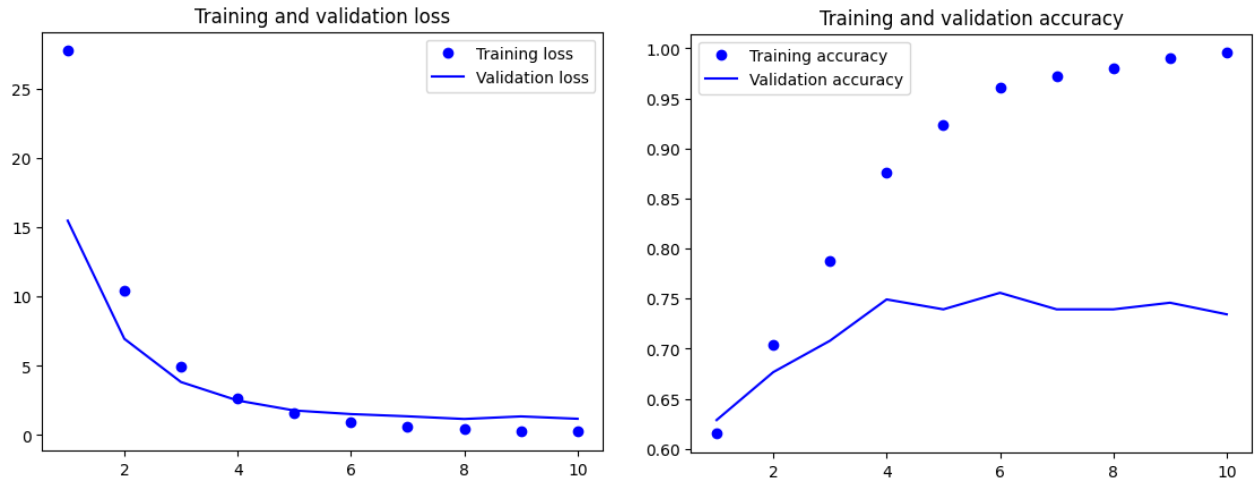
Overall, from the evaluation of the best-performing model on Dataset A, we conclude that the data is not well-suited for predicting UHI. This issue stems firstly from a very small dataset of less than 3000 images, unable to provide the model with information regarding each class.

Secondly and primarily, the definition of the observation unit as a census block, which results in non-standard shapes and sizes of each sample. This variability adds noise through the image background and introduces non-

generalizable features. By removing all spatial features from an urban space except for three—buildings, green spaces, and roads—we likely inadvertently eliminate crucial visual features that are important for identifying UHI.

Additionally, representing green spaces through pixels dilutes the impact of edges, such as a tree line that separates it from an urban space or lines along rows of urban spaces. In our images, due to the pixel shape, what should appear as discrete edges often merges into a continuous green region, frequently overlapping with buildings or roads, making it difficult to distinguish between green spaces based on edges alone. Lastly, it is probable that these four features alone do not capture all the information required to explain UHI. For instance, in panel XX, samples within a class exhibit extreme variations in features while being highly similar to samples from other classes. This indicates that there are additional factors influencing UHI, consistent with findings in the literature, which we will discuss in more detail later

Discussion: Evaluation of best performing model with dataset B– aerial images



(a) Training and Validation Loss **(b) Training and Validation Accuracy**

Panel 6. Training Inception V3: Dataset B (segmented images)

Observing the performance of Inception V3 on Dataset B in panel 6 two primary trends are evident. First, the model begins to overfit around the 5th epoch, after which its performance plateaus. Training accuracy climbs above 95%, suggesting that the model is struggling to extract any additional meaningful information from the images beyond the information required to achieve 75% accuracy. Despite multiple fine-tuning steps, such as reducing the complexity of the additional dense top layer from 512 to 256 neurons, applying L2 regularization with values ranging from 0.01 to 0.05, and experimenting with an exponential decay learning rate with decay rates from 0.5 to 0.96, the pattern of overfitting persists. This is likely due to the highly complex base architecture of Inception V3, which may be too intricate for the nuances of Dataset B, leading to excessive memorization rather than learning generalizable patterns.

On the other hand, we observe that validation loss continues to decrease despite fluctuations in accuracy, including minor dips. This indicates that the model is becoming more confident in its predictions of class probabilities for the samples, suggesting a more calibrated model even as it struggles with overfitting. On the test dataset, the model achieves 70% accuracy, which might have been slightly higher had we defined our best model based on validation loss rather than validation accuracy. However, as discussed previously, we prioritize confident predictions over accidental accuracies.

From Table 3, we can see that our errors are primarily stemming from class 0 (heat effect between 0 and 1.5) being overpredicted as class 2 (more than 1.5) relative to the other two classes. Intuitively this may be expected as the model would in essence, be trying to identify the ratio and distribution of green spaces versus impervious (concrete) structures, as those are the main signals in our images to predict urban heat island effect. It is more likely that images are dominated by either concrete structures or greenspaces following identifiable edges and the density and distribution of either of the dominant feature clearly indicates UHI as less than 0 or more than 1.5. However, for cases between 0 and 1.5, samples may similarly be following a pattern of domination by either of our two primary features like in the other two classes and hence the difference in the UHI outcome is then likely driven by other information not available in the images such as population density and geographic location. This issue in discussed in the next section.

Further the issue of a small dataset may also be contributing to this issue. With about 4400 training samples and some class imbalance, the model may simply not have enough data to learn features. The high success rate of class 2, the class with the highest number of samples also indicates the same.

[0 = between 0 and 1.5, 1 = less than 0, 2 = more than 1.5]

Predicted Label	0	1	2
True Label			
0	50	36	51
1	16	118	20
2	30	28	255

Table 5: Confusion Matrix (Inception V3) – Dataset B

Given the disproportionate errors in class 0, i.e. UHI between o and 1.5 I inspect the feature maps of samples belong to this class, both incorrectly predicted in panel 7 and 8 and correctly predicted in panel 9 and 10.

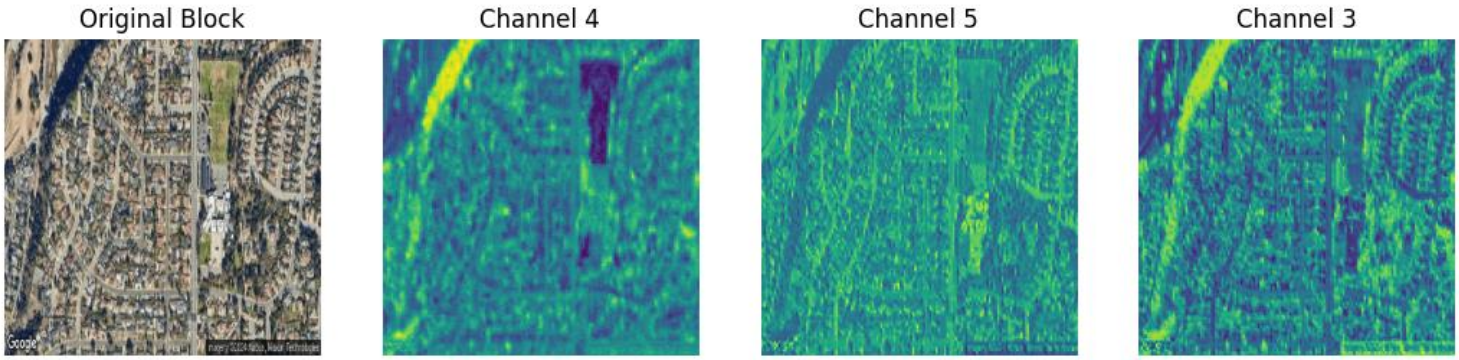
I find that the model Inception V3 generally correctly identifies different sets of edges pertaining to greenspaces and concrete structures but upon careful inspection we can see small regions where it does indeed confuse certain regions of greenspace, concrete and water as the same. It also seems to be able to discriminate between non-green and non-concrete spaces, i.e. clearances.

Observing the model's feature map for samples which have been incorrectly identified in Panel 7, we can better understand why the model is predicting the image as more than 1.5. The tree canopy region is very finely dispersed amongst a dense housing region and the model has difficulty separating the edges for trees from the buildings, interpreting the green regions as as part of the concrete structures. It thus receives the signal that this is dense urban region and hence predicts it to have a high UHI.

Conversely, in Panel 8 the image is predicted incorrectly as less than 0 and we see the opposite issue in edge identification occurring. While it generally correctly identifies the clearance areas, the edges at the point where the dense greenspace region in the right, meets the urban areas on the left seem to be interpreted as a greenspace region. Further roadways with trees along them, are also getting considered as greenspace. Thus with a signal of a lot of greenspace, it incorrectly predicts the region as having low UHI.

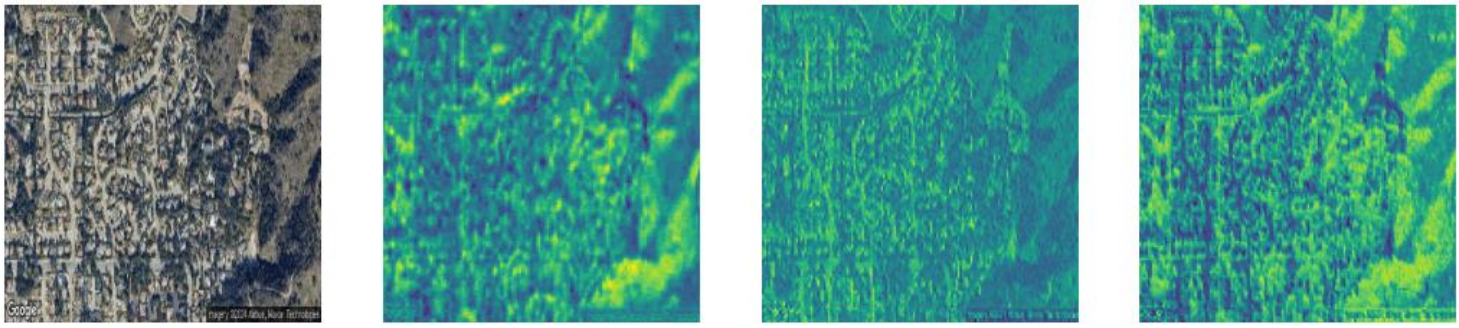
Finally, we can discern why the last two images in panel 9 and 10 have been predicted correctly. Both images have a nearly equal split with good distribution of greenspace around concrete regions. The edges separating greenspaces and buildings are found with high accuracy. While in image 2 some blurring of edges can be seen, this is not substantial enough to throw of the model's conclusion that neither feature - greenspace or concrete areas are dominating in density and distribution, leading it to conclude a UHI of between 0 and 1.5.

Predicted Class = UHI more than 1.5, Original Class = UHI between 0 and 1.5

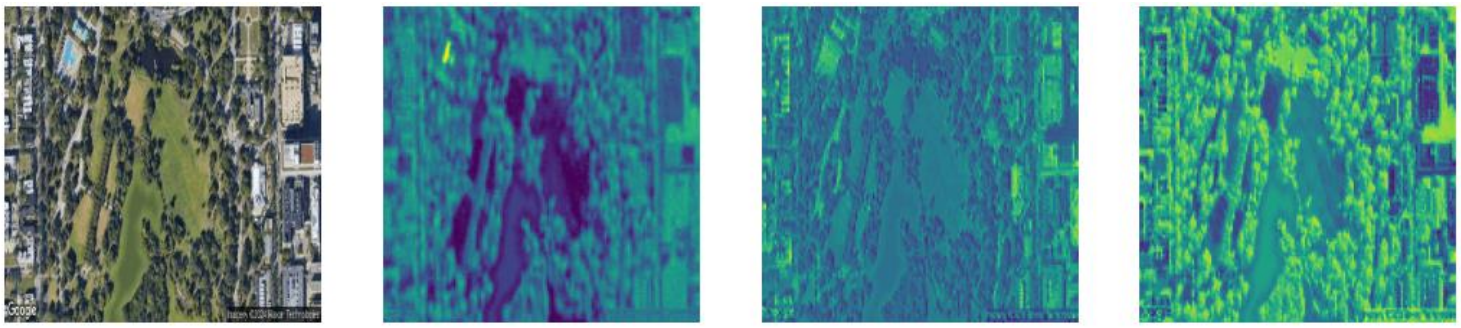


Panel 7: Feature Map of <u>Incorrect Predictions</u>:

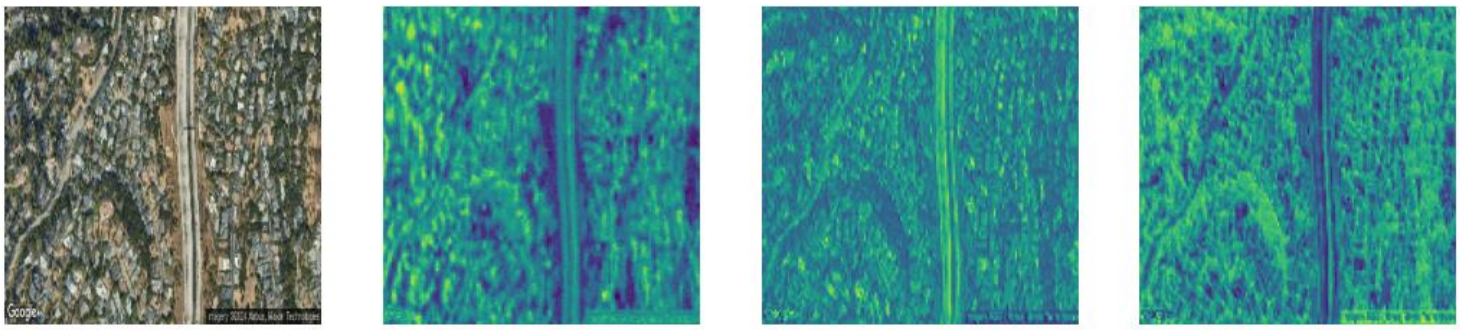
Predicted Class = UHI less than 0, Original Class = UHI between 0 and 1.5



Panel 8: Feature Map of <u>Incorrect Predictions</u>:



Panel 9: Feature Map of <u>Correct Predictions</u>: Original Class = UHI between 0 and 1.5



Panel 10: Feature Map of <u>Correct Predictions</u>: Original Class = UHI between 0 and 1.5

Conclusion and Reflection on Overall Performance

Urban Heat Island effect is influenced by a myriad of factors such as population density of a region, type of green cover (certain types of vegetation having a more prominent effect), the baseline temperature of a place due to its geographic location, the material of buildings etc. This was further evident from analyzing misclassified samples, like the ones shown above, many of which if I were to naively classify based solely on the images myself, I might misclassify as well.

As the literature suggests and from my results, I therefore expected that the impervious surfaces' and greenspaces' density and distribution in an image alone, is unlikely to explain 100% of the variation of UHI. Ideally, I would define performance success nearer to 80% test accuracy given the information advantage in images of impervious surfaces and greenspaces density and distribution which would be difficult to capture in tabular data. Despite experimenting with two different datasets, experimenting with different number of output classes, 5 different models – both self-developed and poweful pre-trained architectures, hyperparameter tuning such as adjusting learning rates, undertaking adaptive learning rates and changing the depth and width of the model, I was at best able to achieve 70% test accuracy on Dataset B. This too occurred with overfitting despite regularization and decreasing model complexity, likely an artefact of the highly complex base architecture.

Overall, I was happy with my vision-based approach to UHI given the influence of spatial patterns, which is difficult to capture in tabular data. Given the small dataset of about 5000 images, I find 70% test accuracy somewhat satisfactory. Nonetheless, my results also highlight the importance of undertaking a multi-modal big data approach, by supplementing with much more image data across more cities and with tabular information that affects UHI not available in images, discussed prior - a future avenue for extending this prediction exercise.

However, I find that the model's ineffectiveness with dataset A prevents me from achieving my original goal of identifying Urban Heat Islands (UHI) during the pre-urban planning phase. Dataset A's images are more akin to early-stage urban planning schematics, unlike the four-band aerial images in dataset B that depict fully developed urban areas. Modifying these established urban structures is typically slow, expensive, and often impractical and hence benefits much less from predictions of UHI.

I primarily attribute the failure of dataset A, to the very small dataset and varying shapes and zoom levels of each sample due to the observational unit being at the census block group level. Due to the nature of each of all the GIS datasets used - shape of a county, shape of census block groups, vectors for buildings and roads and the raster for greenspace - which I have access to, are each configured with unique Coordinate Reference Systems (CRS).

These CRS specifications are determined for specific compatibility among the different datasets, but do not align with a self-created standardized grid layout. Consequently, with the available data it is not possible to apply a uniform grid overlay across the county to create equally spaced observational units (instead of using the census block groups shape) for dataset A, which would help standardize shape and scale across observational units. Given that on dataset B of aerial images we saw that the noise in the images was confusing the model in identifying edges appropriately, I still do believe that with more high-resolution data sources, pre-processing and data volume combined with multi-modal data can be more beneficial. This also provides a future avenue for using more flexible GIS datasets available through proprietary sources such as ArcGIS.

Bibliography

Assaf, G., Hu, X., & Assaad, R. H. (2023). Predicting Urban Heat Island severity on the census-tract level using Bayesian networks. Sustainable Cities and Society, 97, 104756–104756. https://doi.org/10.1016/j.scs.2023.104756

Bohn, K. (2023, May 23). Strategic city planning can help reduce urban heat island effect | Penn State University. Www.psu.edu. https://www.psu.edu/news/agricultural-sciences/story/strategic-city-planning-can-help-reduce-urban-heat-island-effect/

Chakraborty, T., Hsu, A., Manya, D., & Sheriff, G. (2020). A spatially explicit surface urban heat island database for the United States: Characterization, uncertainties, and possible applications. ISPRS Journal of Photogrammetry and Remote Sensing, 168, 74–88.

City of Chicago. (2023). Building footprints in Chicago [Data set]. City of Chicago. https://data.cityofchicago.org/Buildings/Building-Footprints-current-/hz9b-7nh8

Cook_County_GIS. (2023). Aerial imagery reference tiles [Illinois--Cook County] [Data set]. Cook Central. https://hub-cookcountyil.opendata.arcgis.com/datasets/cookcountyil::cook-county-aerial-imagery-2023/about

Hsu, A., Sheriff, G., Chakraborty, T., & Manya, D. (2021). Disproportionate exposure to urban heat island intensity across major US cities. Nature Communications, 12(1). https://doi.org/10.1038/s41467-021-22799-5

Kajosaari, A., Kamyar Hasanzadeh, Fagerholm, N., Pilvi Nummi, Kuusisto-Hjort, P., & Kyttä, M. (2024). Predicting context-sensitive urban green space quality to support urban green infrastructure planning. Landscape and Urban Planning, 242, 104952–104952. https://doi.org/10.1016/j.landurbplan.2023.104952

Los Angeles GeoHub. (2017, August 3). Building Footprints. Geohub.lacity.org. https://geohub.lacity.org/datasets/813fcefde1f64b209103107b26a8909f/explore

Los Angeles GeoHub. (2020, June 26). Streets (Centerline). Geohub.lacity.org. https://geohub.lacity.org/datasets/d3cd48afaacd4913b923fd98c6591276

Levy, J. (2024). Street center lines in Chicago [Data set]. City of Chicago. https://data.cityofchicago.org/Transportation/Street-Center-Lines/6imu-meau

National Geographic. (2022, May 20). Urban Heat Island | National Geographic Society. Education.nationalgeographic.org. https://education.nationalgeographic.org/resource/urban-heat-island/

Oh, J. W., Ngarambe, J., Duhirwe, P. N., Yun, G. Y., & Santamouris, M. (2020). Using deep-learning to forecast the magnitude and characteristics of urban heat island in Seoul Korea. Scientific Reports, 10(1), 3559. https://doi.org/10.1038/s41598-020-60632-z

Simonyan, K., & Zisserman, A. (2015). Published as a conference paper at ICLR 2015 VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION. https://arxiv.org/pdf/1409.1556

Szegedy, C., Vanhoucke, V., Ioffe, S., & Shlens, J. (2015). Rethinking the Inception Architecture for Computer Vision. https://arxiv.org/pdf/1512.00567v3

United States Census Bureau. (2022). Cartographic boundary files. Retrieved from https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html

United States Environmental Protection Agency. (2014, February 28). Heat Island Effect. US EPA. https://www.epa.gov/heatislands

Wang, M., & Xu, H. (2021). The impact of building height on urban thermal environment in summer: A case study of Chinese megacities. PLoS ONE, 16(4), e0247786. https://doi.org/10.1371/journal.pone.0247786United States Environmental Protection Agency. (2014, February 28). *Heat Island Effect*. US EPA. https://www.epa.gov/heatislands

National Geographic. (2022, May 20). *urban Heat Island | National Geographic Society*. Education.nationalgeographic.org. https://education.nationalgeographic.org/resource/urban-heat-island/

Office of Technology and Innovation, City of N.Y. (2024, March 8). Building Footprints. NYC Open Data. https://data.cityofnewyork.us/Housing-Development/Building-Footprints/nqwf-w8eh

Office of Technology and Innovation , City of N.Y. (2024, March 8). NYC Street Centerline (CSCL). NYC Open Data. https://data.cityofnewyork.us/City-Government/NYC-Street-Centerline-CSCL-/exjm-f27b

Wang, M., & Xu, H. (2021). The impact of building height on urban thermal environment in summer: A case study of Chinese megacities. *PLoS ONE*, *16*(4), e0247786. https://doi.org/10.1371/journal.pone.0247786