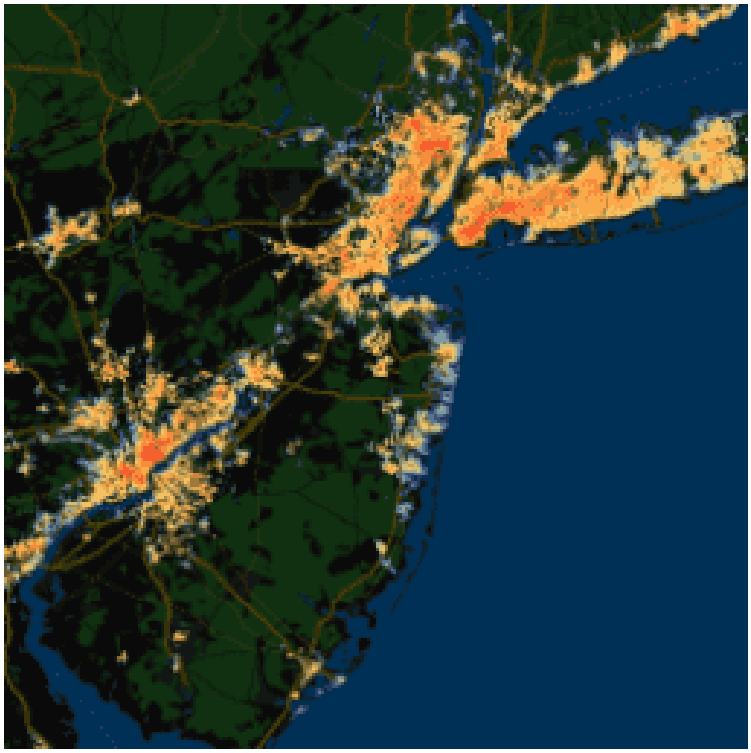


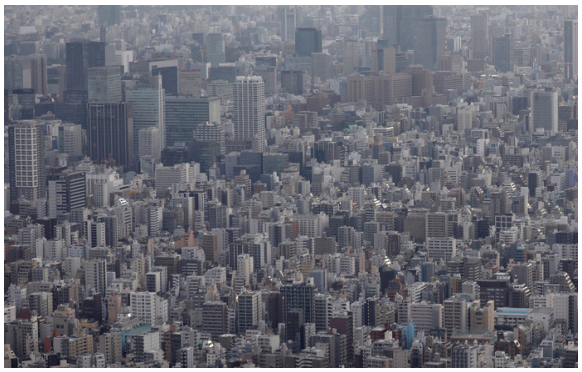
Prediction of Urban Heat Islands Using Machine Learning, Markov Chain Modeling, and Cellular Automation

Sadhil Mehta

Grade 12

Tippecanoe High School - Tipp City

Problem



Since the beginning of the 21st century, the urban growth dynamics of the world have shifted. Cities are expanding at exceptional rates worldwide as countries and, in general, as the world starts to urbanize and industrialize rapidly. However, this change has its resulting consequences. With countless environmental issues that are counterbalancing this rapid industrialization, the growth of urban areas can be stunted. One such obstacle is Urban Heat Islands. Urban Heat Islands (UHIs) are localized regions within metropolitan areas that experience significantly higher temperatures than surrounding rural areas. This phenomenon arises primarily due to human activities and the replacement of natural landscapes with impermeable surfaces such as asphalt, concrete, and buildings. Urban Heat Islands (UHIs) have devastating consequences which includes intensifying climate change, straining infrastructure, and endangering public health. Soaring temperatures lead to deadly heat waves, increased air pollution, and skyrocketing energy demands, disproportionately harming vulnerable populations. Worse, UHIs accelerate biodiversity loss, degrade water quality, and weaken city structures, making urban areas increasingly unlivable and economically unsustainable. These problems are especially present in cities such as Chandigarh, India, where my extended family resides. With the everflowing avalanche of problems UHIs can cause, it is paramount to detect and study them in detail.

Because of these implications, I decided to research Urban Heat Islands. While researching, I learned about the sheer absence of publically available methods to map out and predict Urban Heat Islands. While traditional observational methods provide insights into current UHI conditions, they are often time-consuming and resource-intensive. Predictive modeling offers a promising alternative, enabling researchers and policymakers to forecast future UHI trends and implement data-driven strategies to address them. Building on my work from past years, I hoped to incorporate machine learning and data-driven methods to study Urban Heat Islands worldwide.

Research

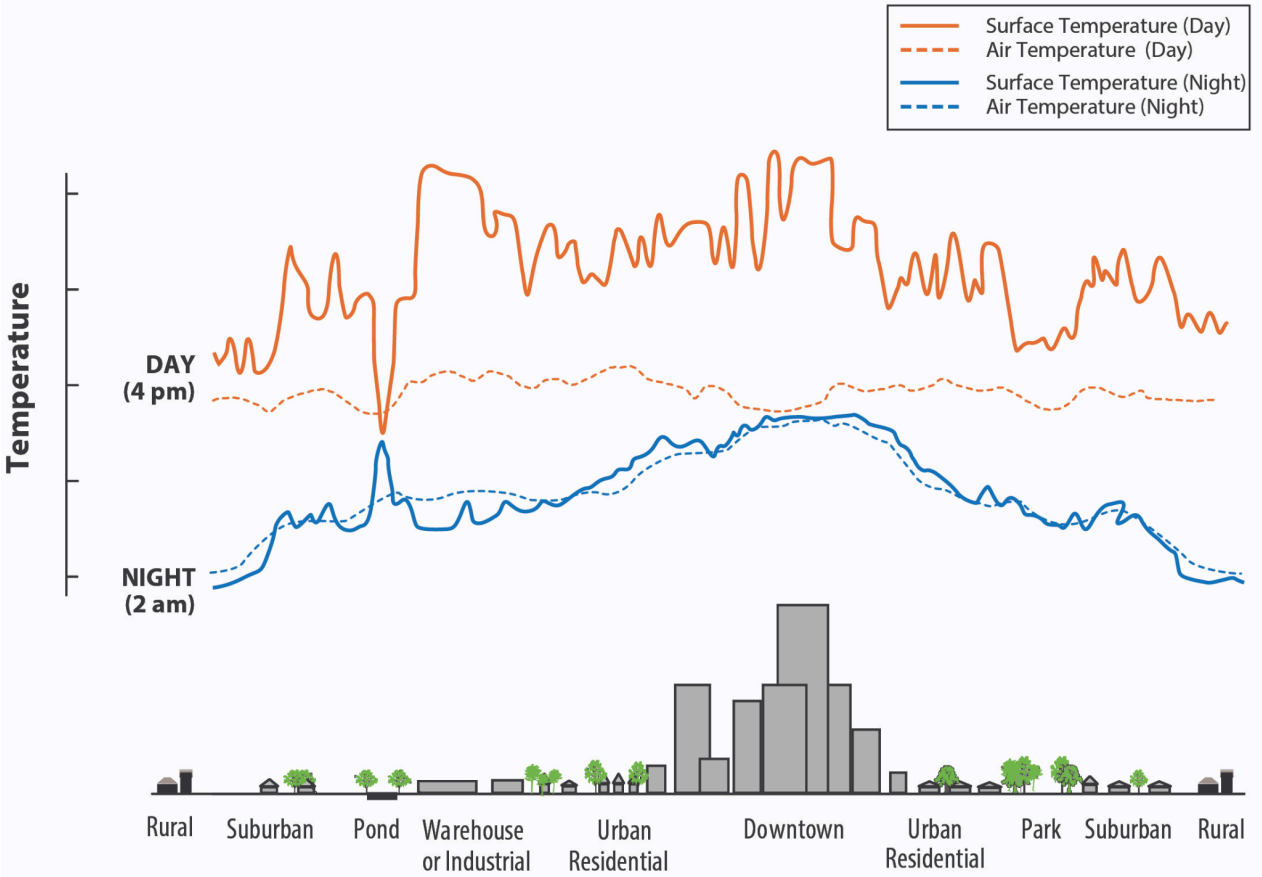
Before starting my analysis on Urban Heat Islands, I wanted to research current urban growth trends, Urban Heat Islands, and their applications in statistical and deep learning methods.

Urban Heat Islands

Causes

Urban Heat Islands (UHIs) are localized regions within urban areas that experience significantly higher temperatures than surrounding rural areas.[1] This phenomenon arises primarily due to human activities. As natural landscapes are uprooted, the urban island effect increases. Trees, vegetation, and water bodies tend to cool the air by providing shade, transpiring water from plant leaves, and evaporating surface water. Hard, dry surfaces in urban areas – such as roofs, sidewalks, roads, buildings, and parking lots – provide less shade and moisture than natural

landscapes, contributing to higher temperatures. These natural landscapes are then replaced with human-made materials such as asphalt and cement. These materials absorb and retain heat during the day and release it slowly at night, contributing to elevated temperatures.



In a scientific sense, UHIs are a direct consequence of the energy balance alterations within urban environments. The primary mechanisms driving this phenomenon include reduced evapotranspiration, surface albedo modification, heat emissions from anthropogenic activities, and modified wind patterns due to urban structures. Evapotranspiration, the process by which vegetation releases water vapor, is crucial for cooling natural landscapes. However, the replacement of vegetation with artificial materials reduces this cooling effect. Additionally,

urban surfaces tend to have lower albedo, meaning they absorb more solar radiation compared to natural environments. Waste heat from vehicles, industries, and air conditioning units further exacerbates the warming effect.

Impacts

The impact of UHIs is evident in several key areas. Urban Heat Islands can elevate electricity and energy demands, particularly for air conditioning, where demand increases from 1% to 9% for each 2°F change in temperature. This increased demand contributes to higher electricity expenses. Peak demand generally occurs on exceptionally hot afternoons, when offices and homes run air-conditioning systems, lights, and appliances. This increased demand can overload systems and require a utility to institute controlled brownouts or blackouts to avoid power outages. As a byproduct of this, companies start to rely on fossil fuel power plants as a power source to meet electricity needs. This increases pollution and greenhouse gas levels in the atmosphere. These pollutants are harmful to human health and also contribute to complex air quality problems such as the formation of ground-level ozone (smog), fine particulate matter, and acid rain. Ground-level ozone, in particular, is formed when nitrogen oxides and volatile organic compounds react in sunlight and hot weather. Along with this, heat-related death and illness are on the rise as these Urban Heat Islands pop up more and more. Heat is of greatest concern for groups such as older adults, young children, populations with low-income, people who work outdoors, and people with chronic health conditions, disabilities, mobility constraints, or taking certain medications. From 2004 to 2018 the Centers for Disease Control and Prevention recorded an average of 702 heat deaths per year.

Study and Mitigation

Scientifically, UHIs are often studied using satellite remote sensing, ground-based temperature measurements, and computational modeling. Satellite data from sources such as NASA's Landsat program and MODIS sensors provide valuable insights into land surface temperature variations. Ground-based meteorological stations complement this data by measuring atmospheric temperatures at different locations within urban and rural settings. Computational models, including mesoscale climate models and land surface energy balance models, help simulate the interactions between urban surfaces and atmospheric conditions.

Mitigating UHIs requires a multi-faceted approach that includes increasing urban greenery, implementing reflective roofing materials, improving urban design, and reducing anthropogenic heat emissions. Strategies such as green roofs, urban tree canopies, and permeable pavements help enhance cooling through increased evapotranspiration and shading. Policy interventions and urban planning initiatives integrate scientific findings to develop effective adaptation and mitigation strategies.[1]

Visualizing Urban Heat Islands

To effectively visualize and analyze UHIs, researchers employ remote sensing techniques and various indices that quantify vegetation cover, built-up areas, and surface temperatures.

A fundamental parameter in UHI studies is the Land Surface Temperature (LST), which represents the radiative skin temperature of the Earth's surface. LST is typically derived from thermal infrared (TIR) data obtained from satellite sensors such as Landsat's Thermal Infrared Sensor (TIRS). The process involves converting digital numbers to spectral radiance, followed by the application of the Planck function to estimate temperature values. Accurate LST retrieval requires considerations of surface emissivity and atmospheric corrections.

The Normalized Difference Vegetation Index (NDVI) is a widely used metric for assessing vegetation health and density. It is calculated using the reflectance values in the near-infrared (NIR) and red (Red) bands of satellite imagery:

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

In this formula, NIR represents the reflectance in the near-infrared band, and Red denotes the reflectance in the red band. NDVI values range from -1 to 1, where higher positive values indicate dense, healthy vegetation, values near zero suggest sparse or stressed vegetation, and negative values correspond to non-vegetated surfaces like water bodies. In UHI studies, areas

with low NDVI values often correlate with higher surface temperatures due to the lack of vegetation and the prevalence of impervious surfaces.

To quantify urbanization and its impact on surface temperatures, the Normalized Difference Built-up Index (NDBI) is utilized. NDBI is calculated using the reflectance values in the short-wave infrared (SWIR) and near-infrared (NIR) bands:

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR}$$

Here, SWIR represents the reflectance in the short-wave infrared band, and NIR denotes the reflectance in the near-infrared band. NDBI values range from -1 to 1, with higher positive values indicating built-up urban areas. Studies have shown a strong linear relationship between NDBI and LST, suggesting that areas with higher NDBI values tend to exhibit elevated surface temperatures, characteristic of UHIs.

Albedo, the measure of surface reflectivity, significantly influences surface temperatures. Surfaces with low albedo, such as asphalt and dark rooftops, absorb more solar radiation, leading to higher temperatures. Conversely, surfaces with high albedo reflect more solar energy, contributing to cooler temperatures. Quantifying albedo involves calculating the ratio of reflected radiation from the surface to the incoming solar radiation, which can be derived from satellite-based reflectance measurements across multiple spectral bands.

By integrating LST, NDVI, NDBI, and albedo data within Geographic Information Systems (GIS), researchers can create detailed spatial representations of UHIs. Overlaying these indices allows

for the identification of hotspots, assessment of the cooling effects of green spaces, and evaluation of the impact of urban materials on temperature distribution. Such comprehensive visualizations are crucial for urban planners and policymakers aiming to develop strategies to mitigate UHI effects, such as increasing urban greenery, implementing reflective building materials, and optimizing urban layouts to enhance natural ventilation.[2]

When I researched Urban Heat Islands, I found that I could use a variety of methods that could be used to predict their future. As cities come in multiple shapes and sizes, I wanted to select different methods that could tackle different types of cities and methods that could possibly work with both methods. I was going to rely on benchmark testing, however, it was nowhere to be found. I pondered if I could do this testing on my own and figure out which of these methods is truly the best for predicting the future Urban Heat Island. These are the ones I selected:

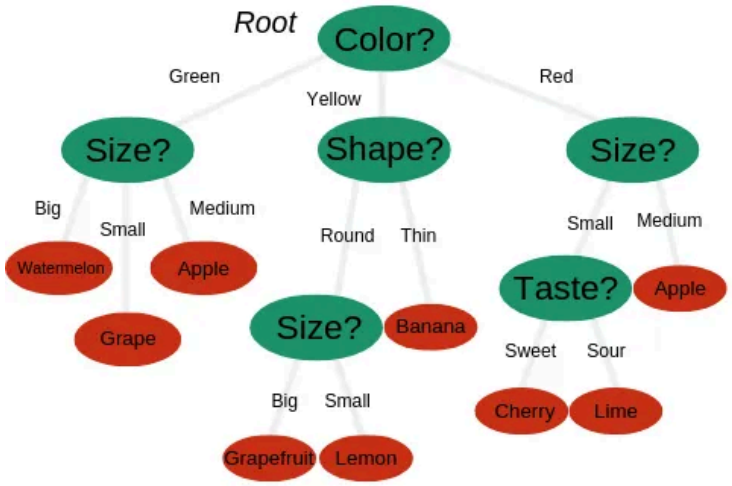
1. ***Random Forest Regression***
2. ***Cellular Automation***
3. ***Markov Chain Modeling***

I decided to research each of these methods.

Random Forest

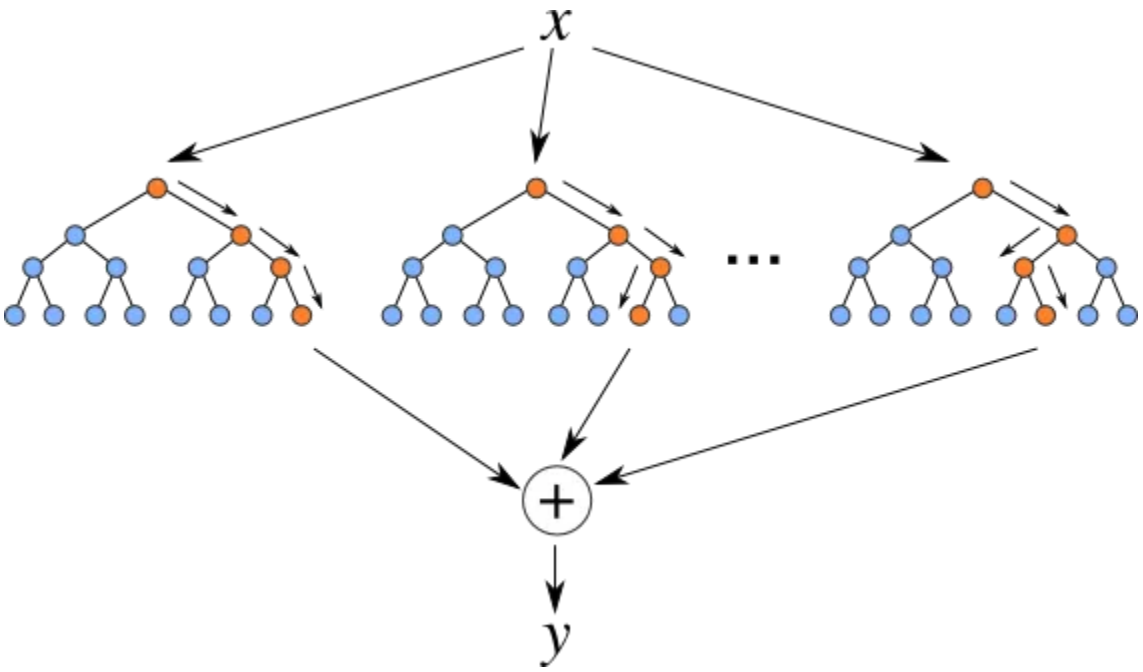
Introduction

The Random Forest Algorithm focuses on a collection of decision trees for its framework. The Random Forest Algorithm is composed of different decision trees, each with the same nodes, but using different data that leads to different leaves. It merges the decisions of multiple decision trees in order to find an answer, which represents the average of all these decision trees.



The random forest algorithm is a supervised learning model; it uses labeled data to “learn” how to classify unlabeled data. The Random Forest Algorithm is used to solve both regression and classification problems, making it a diverse model that is widely used by engineers. This allows it to be a very diverse model which can be applied in many settings and perform well. As it deals with trees, it can accommodate large amounts of data and prevents overfitting. Random Forest is considered ensemble learning, meaning it helps to create more accurate results by using multiple models to come to its conclusion. The algorithm uses the leaves, or final decisions, of

each node to come to a conclusion of its own. This increases the accuracy of the model since it's looking at the results of many different decision trees and finding an average. This is depicted in the diagram below.



Mathematics Behind Random Forest

The mathematical approaches of random forest depend on the application of it. When using the Random Forest Algorithm to solve regression problems, you are using the mean squared error (MSE) to show how your data branches from each node.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2$$

This is where N is the number of Data points, fi is the value returned by the model, and yi is the actual value of the data point i. This formula calculates the distance of each node from the predicted actual value, helping to decide which branch is the better decision for your forest. Here, yi is the value of the data point you are testing at a particular node and fi is the value returned by the decision tree.

The GINI index is used when performing classification. This formula uses the class and probability to determine the Gini of each branch on a node, determining which of the branches is more likely to occur. Here, pi represents the relative frequency of the class you are observing in the dataset, and c represents the number of classes.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

The entropy of the model can also be used to determine how nodes branch in a decision tree.

Entropy uses the probability of a certain outcome in order to make a decision on how the node should branch. Unlike the Gini index, it is more mathematical intensive due to the logarithmic function used in calculating it.[4]

Cellular Automation

Cellular Automata (CA) are discrete, abstract computational systems that provide representations of non-linear dynamics in various fields. CA are discrete as they are composed of individual units. They are composed of a finite or denumerable set of homogeneous, simple units, the atoms or cells. At each time unit, the cells instantiate one of a finite set of states. They evolve in parallel at discrete time steps, following state update functions or dynamical transition rules. These update each cell state based upon the states of other cells in the cell's local neighborhood. They are also abstract in the sense that purely mathematical terms and physical structures define them. Thirdly, CAs are computational systems: they can compute functions and solve algorithmic problems.[3]

Markov Chain Model

Markov chains are mathematical systems that “hop” from one state to another based on sets of probabilities. For example, if you made a Markov chain model of a baby's behavior, you might include "playing," "eating", "sleeping," and "crying" as states, which together with other behaviors could form a 'state space': a list of all possible states. In addition, on top of the state space, a Markov chain tells you the probability of hopping, or "transitioning," from one state to any other state---e.g., the chance that a baby currently playing will fall asleep in the next five minutes without crying first. A Markov chain is a stochastic process, but it differs from a general stochastic process in that a Markov chain must be "memory-less." That is, (the probability of) future actions are not dependent upon the steps that led up to the present state. The simplest

example of the weather model. It shows the probability of transition given the current state. If current weather is cloudy then there is a 50% chance of rain and 40% chance of being sunny. So the probability of being sunny given that it's cloudy is only dependent on the previous state which is cloudy. It does not take into account, which state was before that.

The basic property of a Markov chain is that only the most recent point in the trajectory affects what happens next. This is called the Markov Property. It means that Xt+1 depends upon Xt , but it does not depend upon Xt−1, . . . , X1, X0. That is formulated into the Markov Property below:

$$\mathbb{P}(X_{t+1} = s \,|\, X_t = s_t, X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) \;=\; \mathbb{P}(X_{t+1} = s \,|\, X_t = s_t),$$

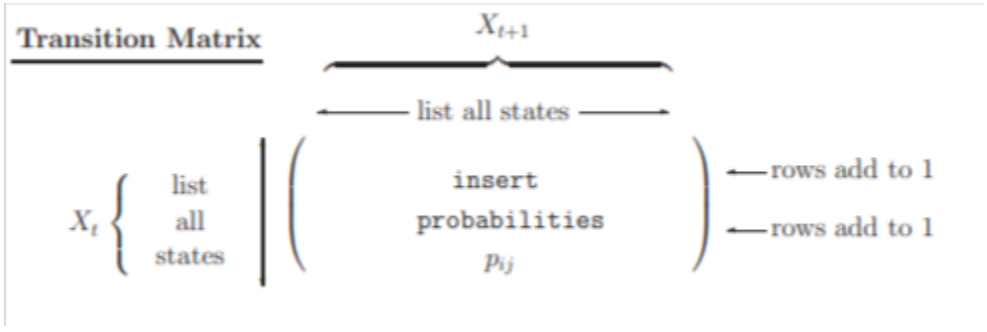
for all $t = 1, 2, 3, \ldots$ and for all states $s_0, s_1, \ldots, s_t, s$.

Explanation:

$\mathbb{P}(X_{t+1} = s \;|\; X_t = s_t,\; X_{t-1} = s_{t-1}, X_{t-2} = s_{t-2}, \ldots, X_1 = s_1, X_0 = s_0)$

- $\mathbb{P}(X_{t+1} = s)$ — *distribution of X_{t+1}*
- $X_t = s_t$ — *depends on X_t*
- $X_{t-1} = s_{t-1}, X_{t-2} = s_{t-2}, \ldots, X_1 = s_1, X_0 = s_0$ — *but whatever happened before time t doesn't matter.*

These probabilities for the state changes are based on transition matrices. These are numeric ways to show the transition models.[5]



After researching my methods, I wanted to select various testing cities to test each of the three methodologies. These cities should have varied on multiple factors including, Urbanization levels, Green Space, Temperature and Climate Zone, Current UHI effect, population, etc. They should also be major cities. After research, these are the cities I selected.

Testing Cities:

New York City, USA

- **Why? High population density, extreme urbanization, and a strong UHI effect.**
- **What does it test? Model accuracy in highly developed urban cores.**

2. Singapore

- **Why? A tropical, high-density city with extensive greenery and sustainable urban planning.**
- **What does it test? The impact of green infrastructure on UHI mitigation.**

3. Dubai, UAE

- **Why? Rapid urban development in a desert climate with extreme temperatures.**
- **What does it test? How UHI manifests in arid regions with high reflectivity and artificial cooling.**

4. São Paulo, Brazil

- **Why? High urban sprawl, moderate temperature fluctuations, and a mix of green and built-up areas.**
- **What does it test? Effects of sprawl vs. vertical development on UHI.**

5. Mumbai, India

- **Why? High humidity, dense informal settlements, and rapid urbanization.**
- **What does it test? How informal settlements and industrialization influence UHI.**

6. London, UK

- **Why? A temperate city with a well-documented UHI and extensive historical climate data.**
- **What does it test? UHI evolution in a temperate climate and policy-driven mitigation strategies.**

7. Lagos, Nigeria

- **Why? A fast-growing African megacity with high temperatures and large informal housing sectors.**
- **What does it test? How urban expansion in developing regions impacts UHI.**

8. Tokyo, Japan

- **Why? High-tech urbanization with frequent temperature fluctuations.**
- **What does it test? The effect of seasonal variations and high-rise density on UHI.**

9. Toronto, Canada

- **Why? A northern city with cold winters but noticeable UHI in summer.**
- **What does it test? Seasonal UHI patterns and mitigation efforts in cold climates.**

10. Istanbul, Turkey

- **Why? A historically significant, rapidly urbanizing city with varied terrain.**
- **What does it test? The influence of topography and historical development on UHI.**

Urban Development Testing Cities

New York City, USA

New York City is one of the most developed and densely populated urban areas in the world, characterized by its extensive high-rise infrastructure, mixed land-use zoning, and complex transportation network. The city's urban heat island effect is intensified by its high concentration of impervious surfaces such as asphalt roads, concrete buildings, and minimal green spaces in some areas. NYC has undergone significant redevelopment, particularly in areas like Hudson Yards and Lower Manhattan, where new high-rise structures have replaced older buildings. Efforts to mitigate UHI include initiatives such as green roofs, reflective surfaces, and expanded urban parks like the High Line and Brooklyn Bridge Park. However, disparities in heat distribution persist, with lower-income neighborhoods experiencing more intense heat stress due to limited green spaces and higher densities of industrial or commercial zones.[7]

Singapore

Singapore is a unique case of urban development, known for its carefully planned, high-density urban landscape that integrates sustainability at its core. Despite rapid development, the city-state has incorporated greenery into its urban fabric through vertical gardens, green roofs, and the preservation of natural areas such as the Central Catchment Nature Reserve. Singapore's strict urban planning ensures that buildings are designed with ventilation corridors to minimize heat retention, and initiatives such as the "City in a Garden" strategy have successfully helped mitigate urban heat island effects. The government also mandates sustainable architecture practices, including water-absorbing building materials and extensive tree planting, making it a global model for mitigating heat stress in a tropical urban environment.[8]

Dubai, UAE

Dubai has experienced rapid urbanization since the 1990s, transforming from a small desert settlement into a modern metropolis dominated by high-rise skyscrapers, artificial islands, and vast commercial districts. The city's urban development relies heavily on air conditioning and artificial cooling, exacerbating its urban heat island effect in an already extreme desert climate. The widespread use of glass and concrete, combined with a lack of natural vegetation, contributes to significant heat retention. While there have been efforts to introduce green spaces, such as in the Dubai Creek area and artificial lakes, the city's heavy reliance on artificial water resources for cooling poses long-term sustainability concerns. The rapid pace of

expansion, particularly in areas like Business Bay and the Marina, has made heat management a growing challenge.[10]

São Paulo, Brazil

São Paulo is a sprawling megacity with a mix of high-rise commercial districts and low-rise residential neighborhoods. The city has undergone rapid expansion, particularly in its outskirts, where informal settlements (favelas) continue to grow due to high demand for housing. This unregulated expansion has led to increased deforestation in surrounding areas, reducing the presence of green spaces and exacerbating urban heat island effects. The city center, with its dense high-rises and limited tree cover, experiences higher temperatures than suburban and rural regions. However, São Paulo has also made efforts to reclaim urban spaces through projects such as the Minhocão elevated park and initiatives to restore green corridors along major streets and avenues.[7]

Mumbai, India

Mumbai is one of the most densely populated cities in the world, with a highly compact urban core and an expanding metropolitan periphery. The city's development is marked by a stark contrast between modern high-rise buildings and widespread informal housing settlements. Due to limited space, much of Mumbai's urban expansion occurs through vertical growth, leading to the construction of commercial and residential skyscrapers. The high proportion of asphalt roads and concrete structures, combined with the city's humid climate, intensifies the urban heat island effect. The presence of the Arabian Sea moderates temperatures slightly, but

poorly ventilated areas such as Dharavi, one of the largest slums in Asia, experience extreme heat stress due to overcrowding and minimal greenery. Urban planning efforts have aimed to introduce more open spaces, such as the Coastal Road project and green initiatives in Bandra-Kurla Complex, but challenges persist in balancing development with environmental sustainability.[9]

London, UK

London has a well-documented urban heat island effect, partly due to its historical development pattern of low-rise, densely packed buildings with limited green roofs. The city has expanded through both infill development and the conversion of industrial sites into residential and commercial areas. Despite its significant green spaces—such as Hyde Park, Richmond Park, and Hampstead Heath—dense areas like the City of London and Canary Wharf experience higher temperatures due to the concentration of office buildings and infrastructure. London's temperate climate helps moderate UHI effects, but its increasing reliance on glass-and-steel structures, especially in newly developed areas like Stratford and Nine Elms, contributes to localized heat retention. Efforts such as urban greening initiatives, river restoration projects, and increased use of permeable surfaces aim to mitigate these effects.[6]

Lagos, Nigeria

Lagos is one of the fastest-growing cities in the world, with rapid urbanization driven by population growth and economic expansion. The city's development is characterized by unplanned urban sprawl, particularly in informal settlements where poor infrastructure and

lack of vegetation exacerbate heat stress. The central business districts, such as Victoria Island and Ikeja, feature modern high-rise buildings, but much of the urban area consists of low-rise, densely packed housing. Lagos faces significant challenges in managing its urban heat island effect due to its tropical climate, increasing energy consumption, and limited urban planning regulations. However, efforts to introduce green spaces, such as the Eko Atlantic project and coastal restoration initiatives, highlight ongoing attempts to improve the city's resilience to rising temperatures.[8]

Tokyo, Japan

Tokyo is a highly developed, densely populated metropolis with a strong urban heat island effect due to its concentration of high-rise buildings, extensive concrete surfaces, and high energy consumption. The city has undergone extensive redevelopment, particularly in areas like Shibuya, Shinjuku, and Marunouchi, where older buildings are being replaced with modern skyscrapers. Tokyo's public transportation infrastructure reduces vehicular heat emissions, but the widespread use of air conditioning contributes to nighttime temperature increases. The city has taken active steps to counteract UHI, including increasing tree coverage, developing rooftop gardens, and implementing heat-reflective building materials. Seasonal variations in temperature also play a role, with Tokyo experiencing hot summers and relatively mild winters.[9]

Toronto, Canada

Toronto experiences a seasonal urban heat island effect, with significant warming during summer months due to its dense urban core, road networks, and industrial zones. The city's

rapid urban development has led to increased high-rise construction, particularly in downtown areas such as the Financial District and along the waterfront. Toronto's extensive suburban sprawl, characterized by single-family homes and green lawns, helps moderate temperature variations, but commercial and industrial zones contribute to heat accumulation. In response, Toronto has implemented various UHI mitigation strategies, including increasing urban canopy cover, requiring green roofs on new buildings, and expanding public park spaces like the Bentway and Don Valley greenbelt.[6]

Istanbul, Turkey

Istanbul is a transcontinental city with a complex urban structure that blends historical architecture with modern development. The city's expansion has been marked by rapid urbanization, particularly in suburban areas such as Başakşehir and Ataşehir, where new residential and commercial complexes have emerged. Istanbul's historical districts, such as Sultanahmet and Beyoğlu, are characterized by narrow streets and older buildings that retain heat, while newly developed high-rise areas contribute to increased heat accumulation. The Bosphorus Strait helps regulate temperatures, but the city's growing vehicular traffic and high-density construction have intensified its urban heat island effect. Green space preservation, particularly in parks like Emirgan and Belgrade Forest, plays a critical role in counteracting these trends.[8]

Procedure

Materials

Software:

1. **Google Earth Engine/Google Colab**
2. **LandSat Data Set**

Code

1. **Define the Area of Interest and dates for analysis**

   The area of interest is the city we are studying. This is usually the central city and the surrounding areas. The dates are going to be used as training data.

2. **Apply Scaling Factors**

   Landsat 8 images are stored in Google Earth Engine as raw digital numbers (DNs). These need to be converted to meaningful values (like reflectance and temperature). The Surface Reflectance Correction converts DN values into reflectance values using scale factors. The thermal band correction converts the DN values into brightness temperature.

3. **Apply a Cloud Mask**

Clouds can obstruct ground-level observations and sometimes affect the NDVI and other values as they interfere with the color bands. The code selects all the clouds by the QA_PIXEL band and sets the clouds and their shadows to zero.

4. **Filter and Process the Image Collection**

   The Landsat 8 dataset is loaded in, and we apply the scaling and masking upon it based on the dates we declared earlier. We take the median across the image to remove outliers that can affect our study later on.

5. **Compute the NDVI**

   We compute the Normalized Difference Vegetation Index to assess vegetation health and occurrence. This is done by the following formula and code below.

   Band 5 (Near Infrared, NIR) is absorbed by the green vegetation whereas Band 4 (Red) is reflected by the vegetation.

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

6. **Compute the Fraction of Vegetation and the Emissivity**

   We use the following code to calculate the minimum and maximum values for NDVI and use those to calculate our Fraction of Vegetation using the formula and code below:

$$Pv = \left(\frac{NDVI - NDVI_{min}}{NDVI_{max} - NDVI_{min}} \right)^2$$

This fractional vegetation is then used to calculate the Emissivity. Emissivity is going to be used to calculate our base Land Surface Temperature (LST) We use the following formula to calculate Emissivity.

$$\varepsilon = 0.004Pv + 0.986$$

7. **Compute the Base Land Surface Temperature**

   The Land Surface Temperature (LST) will be the ultimate measure used to calculate the urban heat island effect and the urban thermal field Index. We use the following formula and code to calculate the LST and visualize it.

$$S_T = \frac{T_B}{1 + \left(\frac{\lambda \times T_B}{\rho}\right) \times \ln \varepsilon} - 273.15$$

8. **Sample the data for the Random Forest Regressor to predict the LST**

   Divide the 2017 LST, NDVI, and EM data into 70% for training and 30% for testing.

9. **Train the Random Forest Regressor**

   Train the Random Forest Regressor to predict the 2022 LST by inputting in the 2017 NDVI ,EM, and LST . Also, apply the regression to the area of interest and visualize it.

10. **Compute the Urban Heat Island (UHI) Effect and Urban Thermal Field Variance Index (UTFVI)**

    Compute the UHI and UTFVI for the predicted 2022 LST of your area of interest using the following formulas:

$$UHI_N = \frac{T_s - T_m}{T_{Std}} \qquad UTFVI = \frac{T_s - T_m}{T_s}$$

11. **Repeat steps 1-7, 10, but change your dates to 2022 and use the actual 2022 data to find your UHI and UTFVI.**

    This will serve as the base we compare our methods to.

12. **Get the Urbanized Layer**

    Get the urban layer by using the NDVI and setting the benchmark that anything under 0.3 NDVI is an urban area.

13. **Define the Neighborhood Kernel and Simulate the Cellular Automata**

    Define the neighborhood kernel, which is the area that determines whether the non-urban area turns into an urban area. Then, create a function that iterates for 5 times (5 years) that can predict the Urban Area in 2022. From that urban area prediction in 2022, set a rule where each urban area growth increases the 2017 LST by 2 degrees celsius.

14. **Repeat step 12 to calculate the UHI and UTFVI for the cellular automated results**
15. **Declare the Transition Probability Matrix**

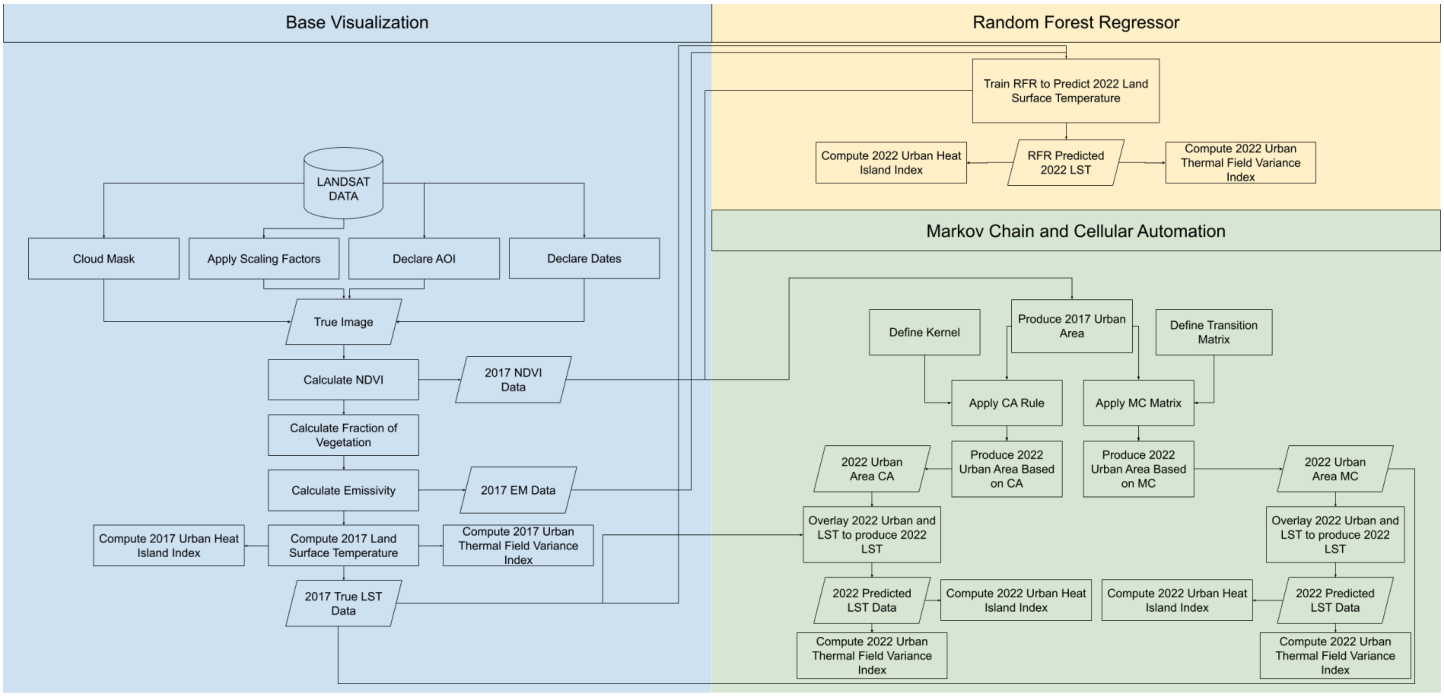
    Declare the transition probability matrix that will be used for the Markov Chain Model. The Matrix should include the chance for an urban area to stay urban and a non-urban area changing into an urban area.

16. **Apply the Markov Chain Model**

Apply the Markov Chain Model to urban expansion. From that urban area prediction in 2022, set a rule where each urban area growth increases the 2017 LST by 2 degrees celsius.

17. **Repeat step 12 to calculate the UHI and UTFVI for the Markov Chain results**

Flow Chart



Measures

1. **Mean Land Surface Temperature (LST)-** Represents the average thermal emission from a given area, often derived from satellite imagery such as Landsat, providing insights into heat distribution across urban and rural landscapes.
2. **Standard Deviation of LST-** Measures the variability in land surface temperatures within a region, highlighting areas with significant temperature fluctuations, which may indicate heterogeneous land cover, Urban Heat Islands, or localized heat sources.

3. **Urban Heat Island Index**- The Urban Heat Island Index (UHII) is a quantitative measure used to assess the intensity of the urban heat island (UHI) effect by comparing temperatures in urban areas to those in surrounding rural or less-developed regions. It is typically calculated using land surface temperatures (LST) or air temperatures from satellite imagery, weather stations, or climate models. The UHII helps researchers and policymakers identify hotspots of excessive heat retention due to factors such as high impervious surface cover, low vegetation, and anthropogenic heat emissions.
4. **Urban Thermal Field Variance Index (UTFVI)**- The Urban Thermal Field Variance Index (UTFVI) is a metric used to evaluate the ecological and environmental impact of urban heat islands by analyzing the spatial variability of land surface temperatures. Unlike the UHII, which focuses on the temperature difference between urban and rural areas, the UTFVI classifies urban regions based on temperature gradients and their potential stress on ecosystems and human health. It is often derived from remote sensing data and categorized into different levels, ranging from healthy thermal environments to extreme thermal stress conditions. The UTFVI provides valuable insights for urban planners to assess thermal comfort, identify areas requiring cooling interventions, and enhance the overall livability of cities.

Urban thermal field variation index	Urban heat island phenomenon	Ecological evaluation index
<0	None	Excellent
0–0.005	Weak	Good
0.005–0.010	Middle	Normal
0.010–0.015	Strong	Bad
0.015–0.020	Stonger	Worse
>0.020	Strongest	Worst

Hypothesis

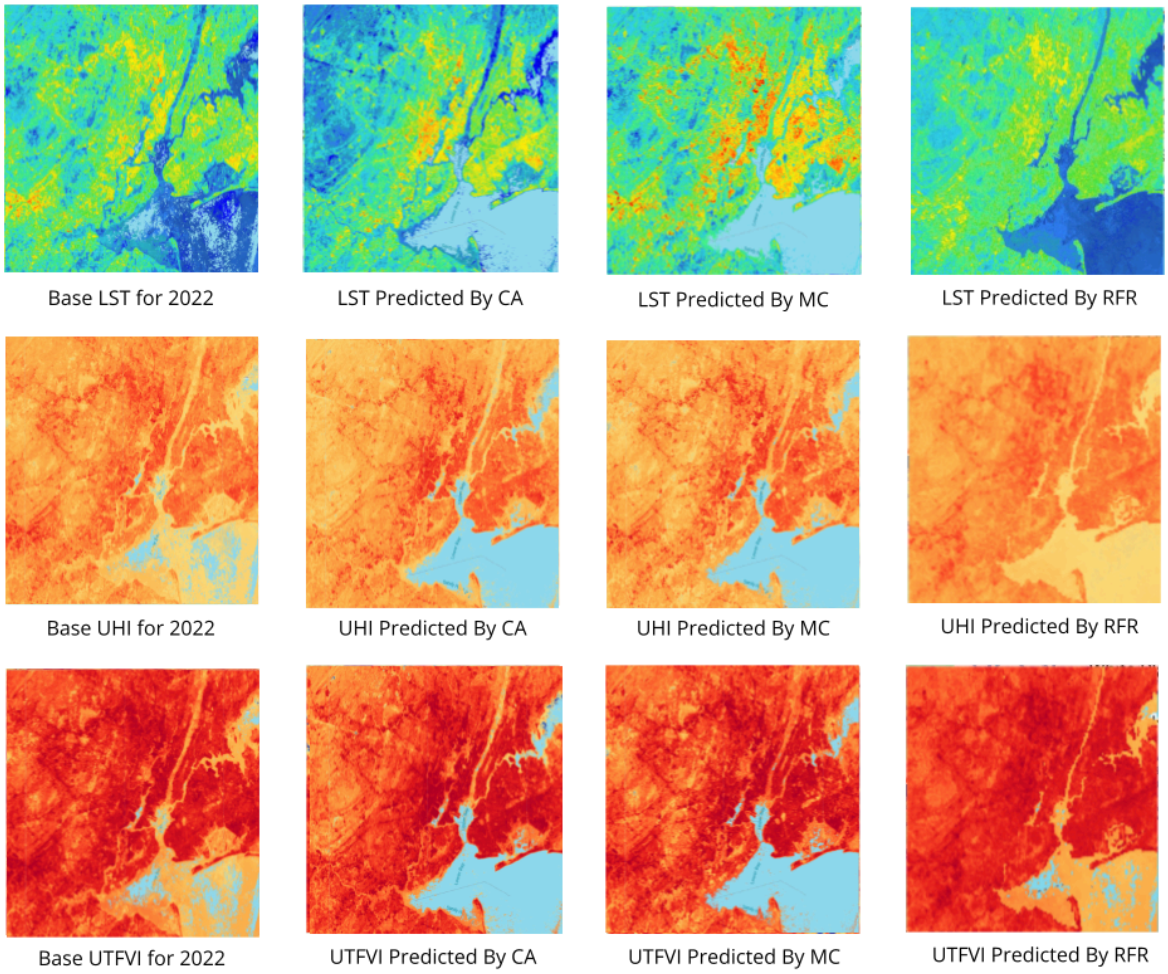
From my research on every test city, Markov Chain methods, and Cellular Automation, I shall see varying results per method and model. However, I believe that my Random Forest Regression will predict all of the Land Surface Temperatures with great accuracy no matter the city. This is because, instead of relying on a single decision tree (which can easily overfit), Random Forest builds multiple trees and takes the majority vote (classification) or average (regression). This smooths out noise and prevents over-reliance on specific patterns. It is also very accurate as each tree is trained on a random subset of data (with replacement). This increases diversity among trees and ensures that the model is less sensitive to specific training examples. For the Markov Chain Model and the Cellular Automated Model, they will all perform worse than the Random Forest Regression. For the Markov Chain and Cellular Automata, I believe that each city will have varying results for accuracy. This is because the Markov Chain model is more based upon probabilities, whereas the Cellular Automaton model is based on the individual “states” of the neighborhood of pixels that surround a certain pixel. This means that the Markov Chain model will be overall more accurate for cities with lower urban development rates. This includes the testing cities of New York City, London, Singapore, Toronto, and Tokyo. The Markov Chain Model is based on probabilities and does not guarantee urban development. These cities have already peaked with their urban growth. In New York, there have been more and more efforts to create green spaces, and with low population growth, the Markov Chain model will do a better job in predicting the urban heat island. This is

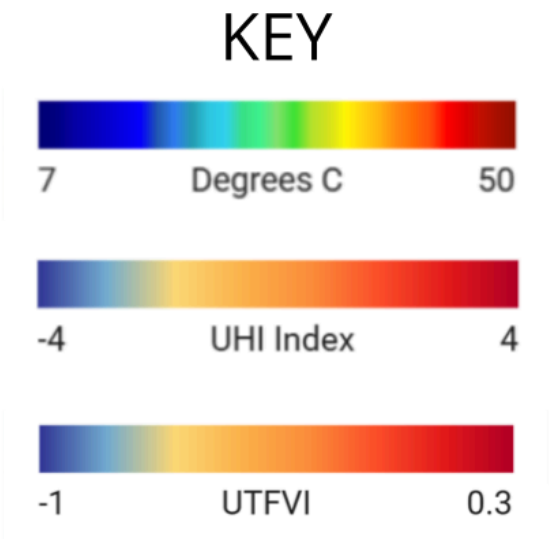
very similar to London as the city has a relatively slow, large-scale urban transformation rather than neighborhood-scale growth. Singapore has also had limited urban growth, thus making it characteristic of a Markov Chain Model. This principle also holds true for Tokyo and Toronto. For the cities of Mumbai, Lagos, Istanbul, Dubai, and Sao Paulo, the Cellular Automation model will perform better. This is because these cities have had rapid urban growth, and the cellular automation model is good for dense, urban environments as it is based on neighborhood relationships. A lot of these cities also have unstructured growth and thus the cellular automation model, which has a more chaotic behavior, will be more characteristic to predict the future Urban Heat Islands in each city.

RESULTS

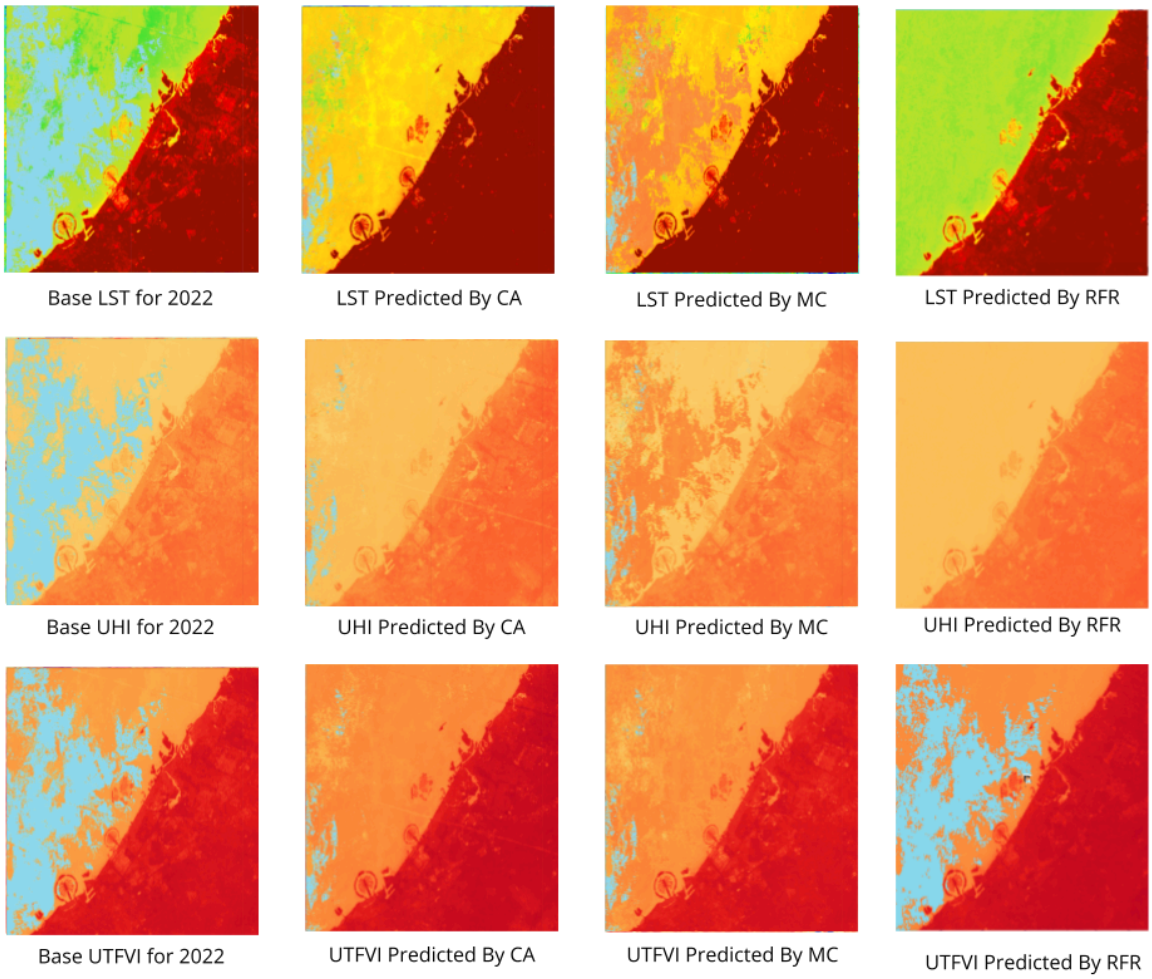
Map Visualizations

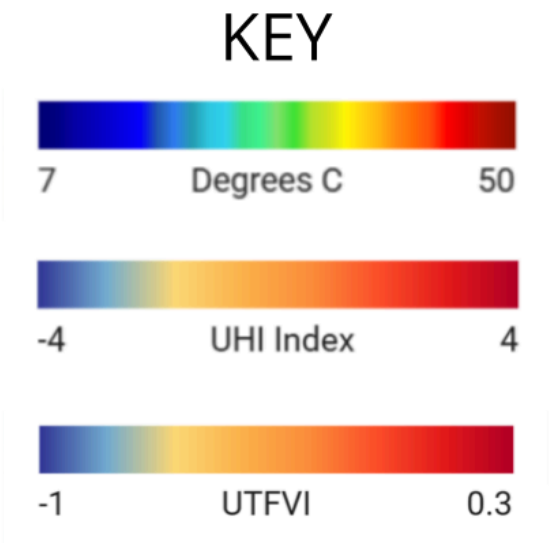
New York City, USA



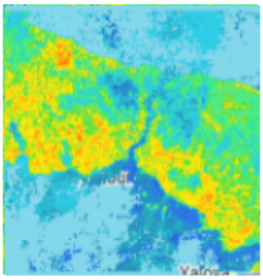


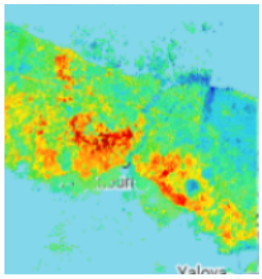
Dubai, UAE


Base LST for 2022
LST Predicted By CA
LST Predicted By MC
LST Predicted By RFR
Base UHI for 2022
UHI Predicted By CA
UHI Predicted By MC
UHI Predicted By RFR
Base UTFVI for 2022
UTFVI Predicted By CA
UTFVI Predicted By MC
UTFVI Predicted By RFR

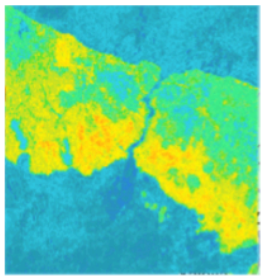

KEY
7
Degrees C
50
-4
UHI Index
4
-1
UTFVI
0.3

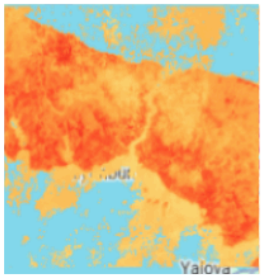
Istanbul, Turkey

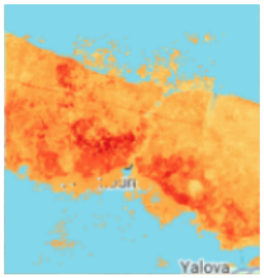

Base LST for 2022

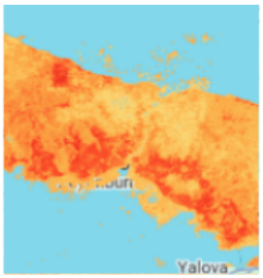

LST Predicted By CA

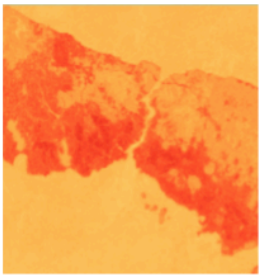

LST Predicted By MC

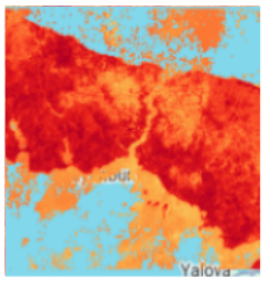

LST Predicted By RFR

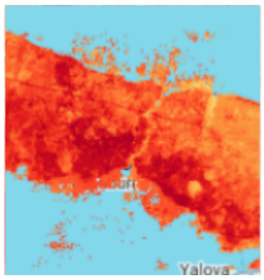

Base UHI for 2022

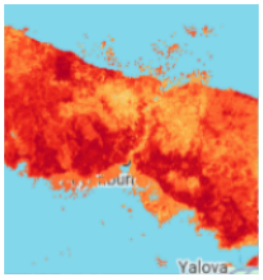

UHI Predicted By CA

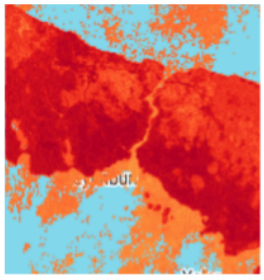

UHI Predicted By MC

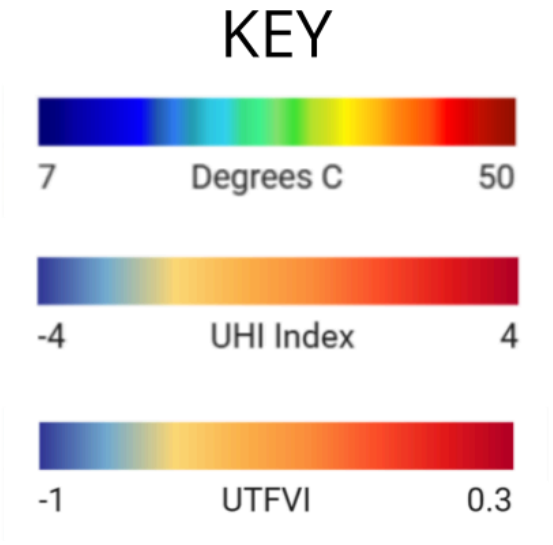

UHI Predicted By RFR


Base UTFVI for 2022

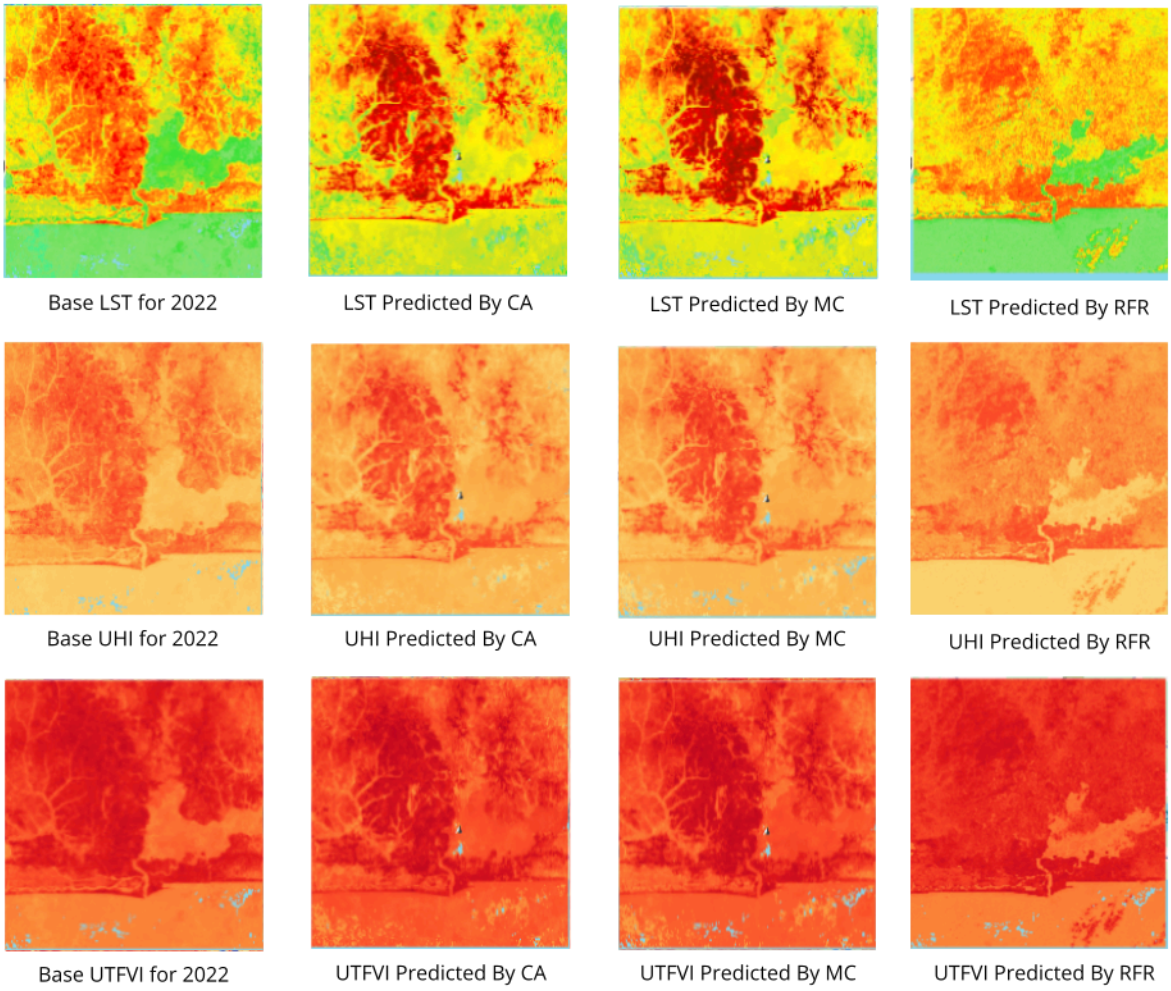

UTFVI Predicted By CA

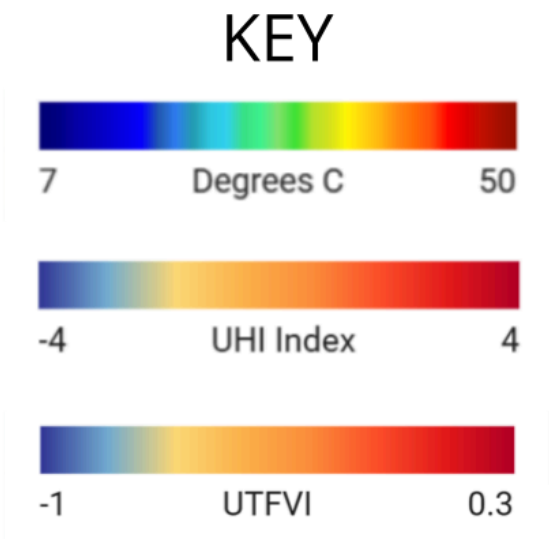

UTFVI Predicted By MC


UTFVI Predicted By RFR

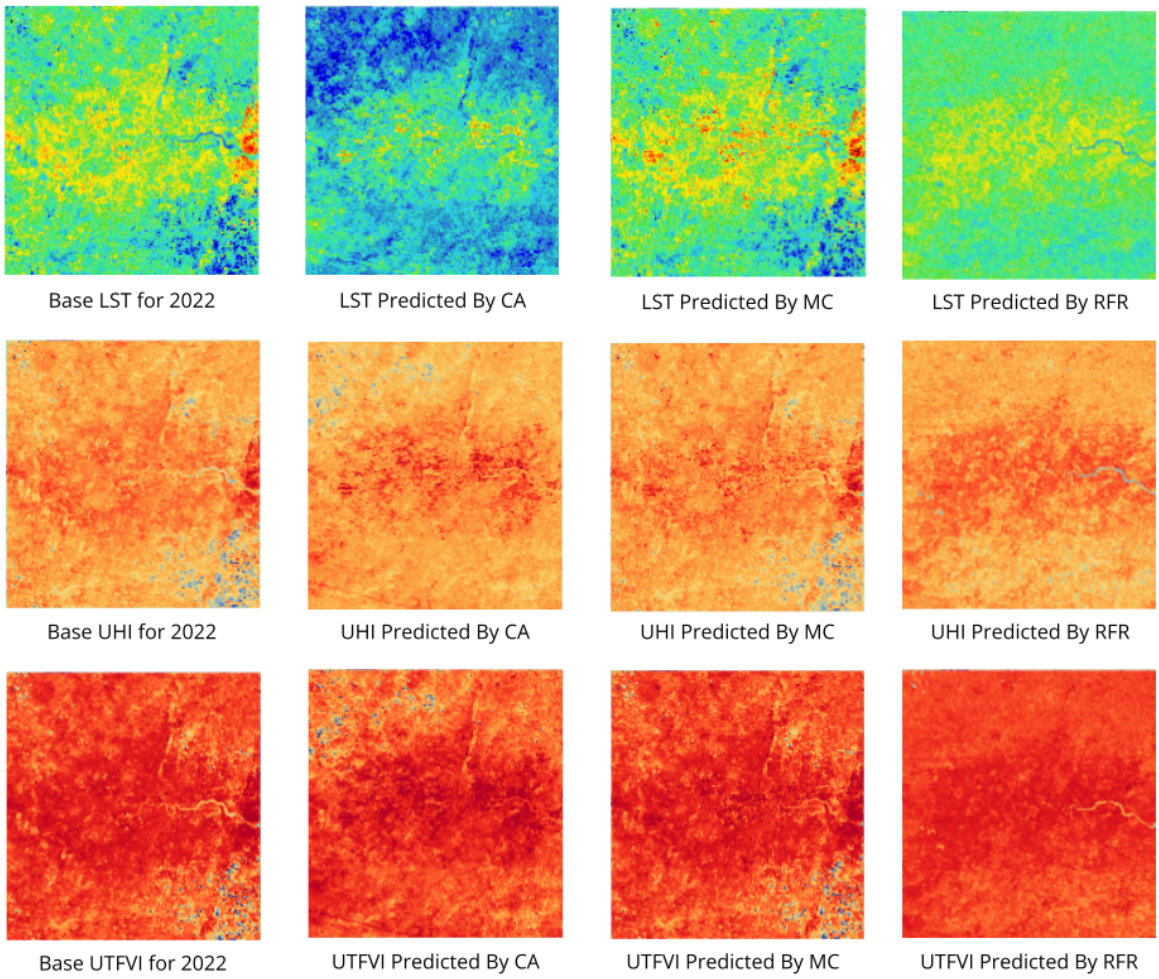


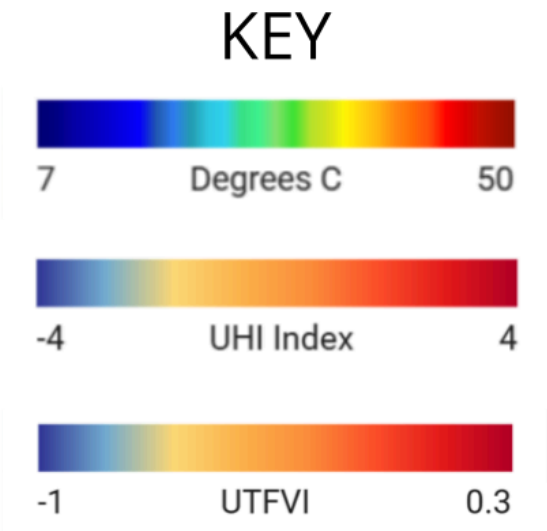
Lagos, Nigeria



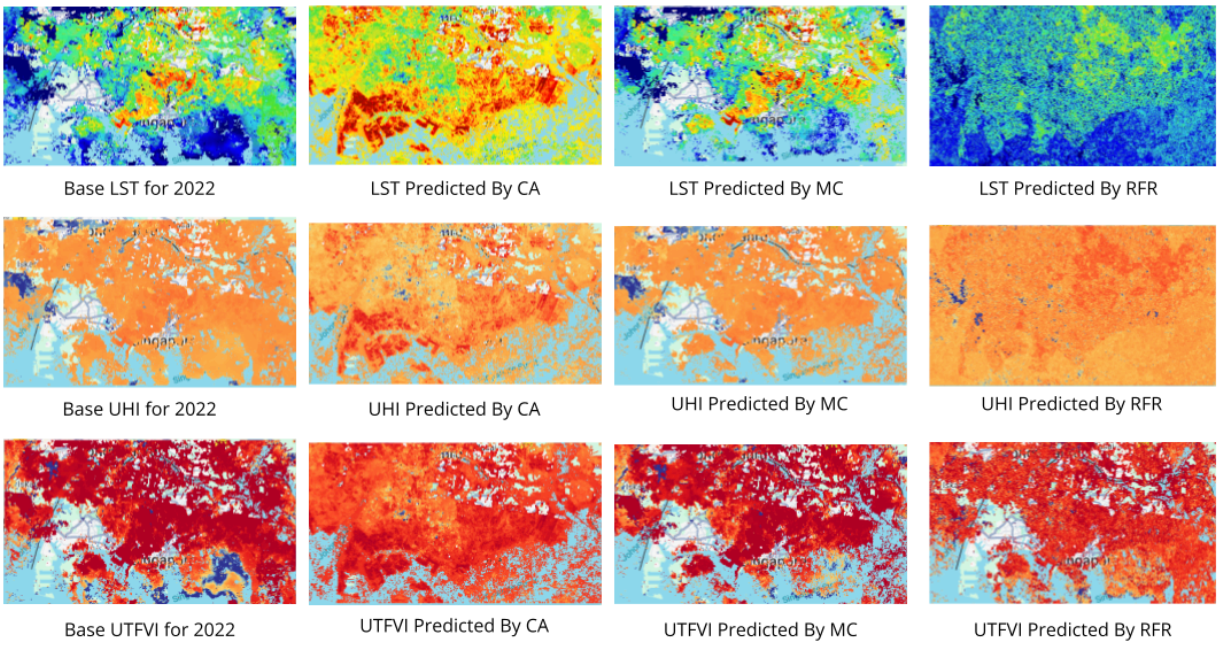


London, UK


Base LST for 2022
LST Predicted By CA
LST Predicted By MC
LST Predicted By RFR
Base UHI for 2022
UHI Predicted By CA
UHI Predicted By MC
UHI Predicted By RFR
Base UTFVI for 2022
UTFVI Predicted By CA
UTFVI Predicted By MC
UTFVI Predicted By RFR


KEY
7
Degrees C
50
-4
UHI Index
4
-1
UTFVI
0.3

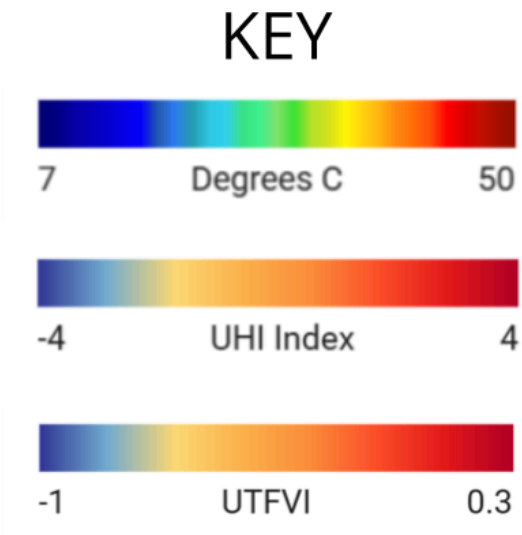
Singapore



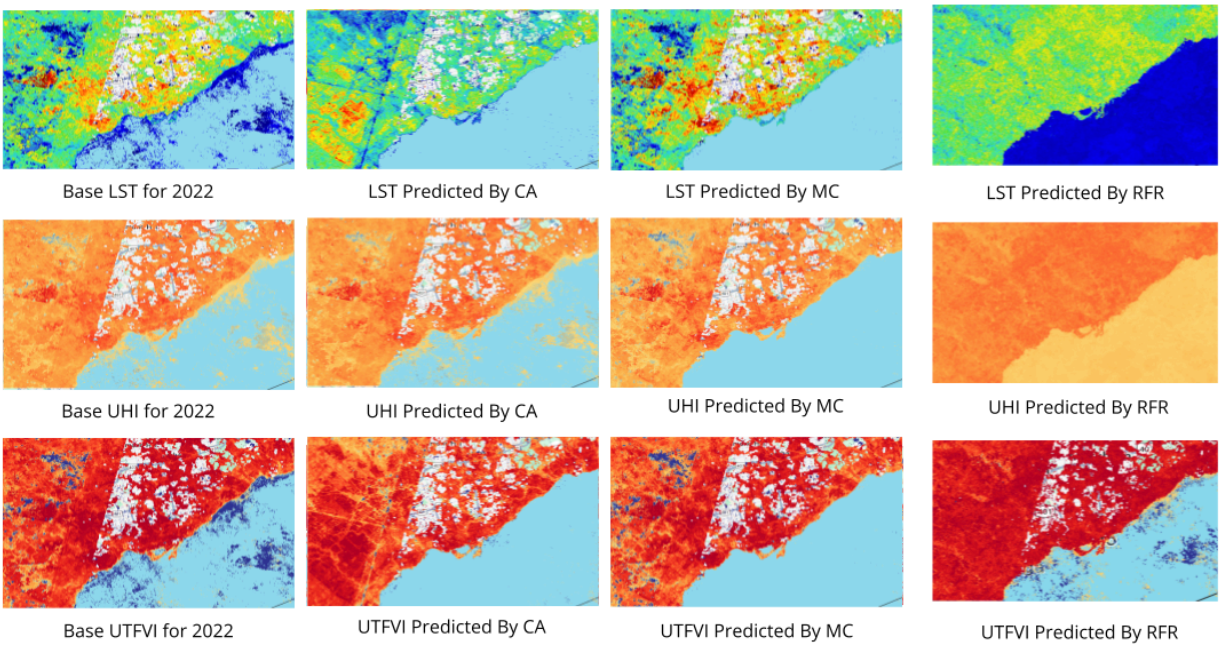
Base LST for 2022 | LST Predicted By CA | LST Predicted By MC | LST Predicted By RFR

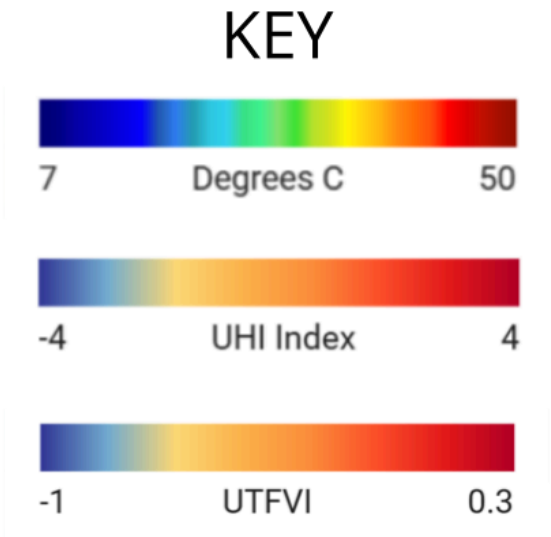
Base UHI for 2022 | UHI Predicted By CA | UHI Predicted By MC | UHI Predicted By RFR

Base UTFVI for 2022 | UTFVI Predicted By CA | UTFVI Predicted By MC | UTFVI Predicted By RFR

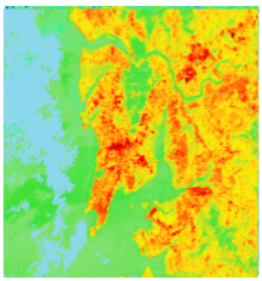


Toronto, Canada

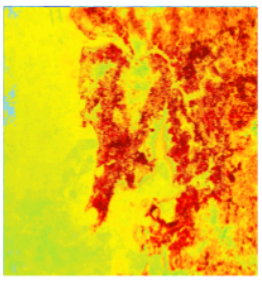




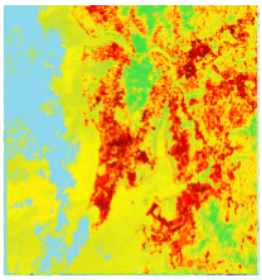
Mumbai, India



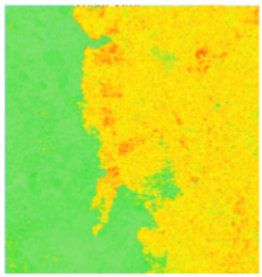
Base LST for 2022



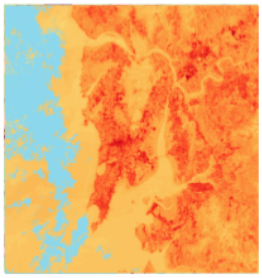
LST Predicted By CA



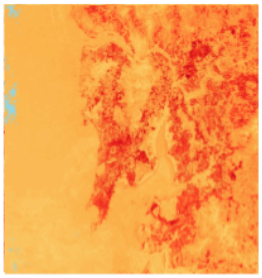
LST Predicted By MC



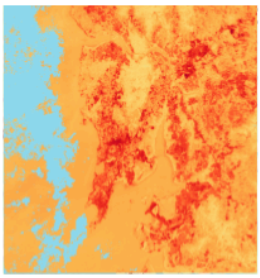
LST Predicted By RFR



Base UHI for 2022

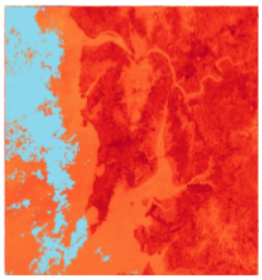


UHI Predicted By CA

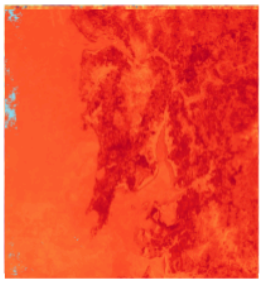


UHI Predicted By MC

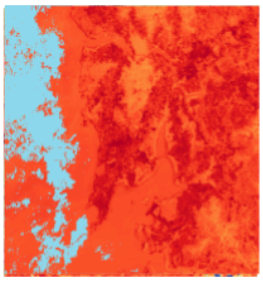
UHI Predicted By RFR



Base UTFVI for 2022

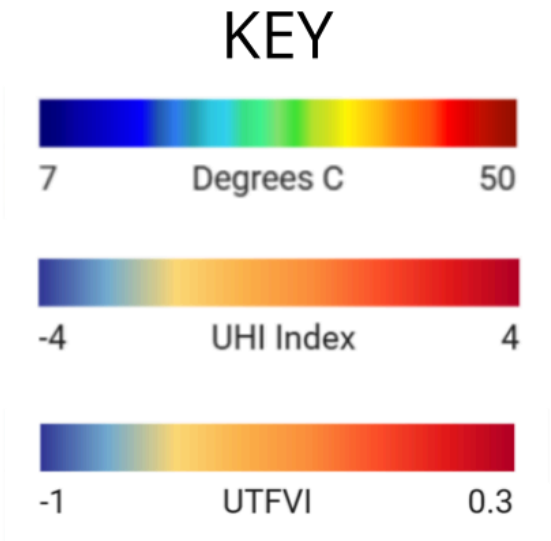


UTFVI Predicted By CA

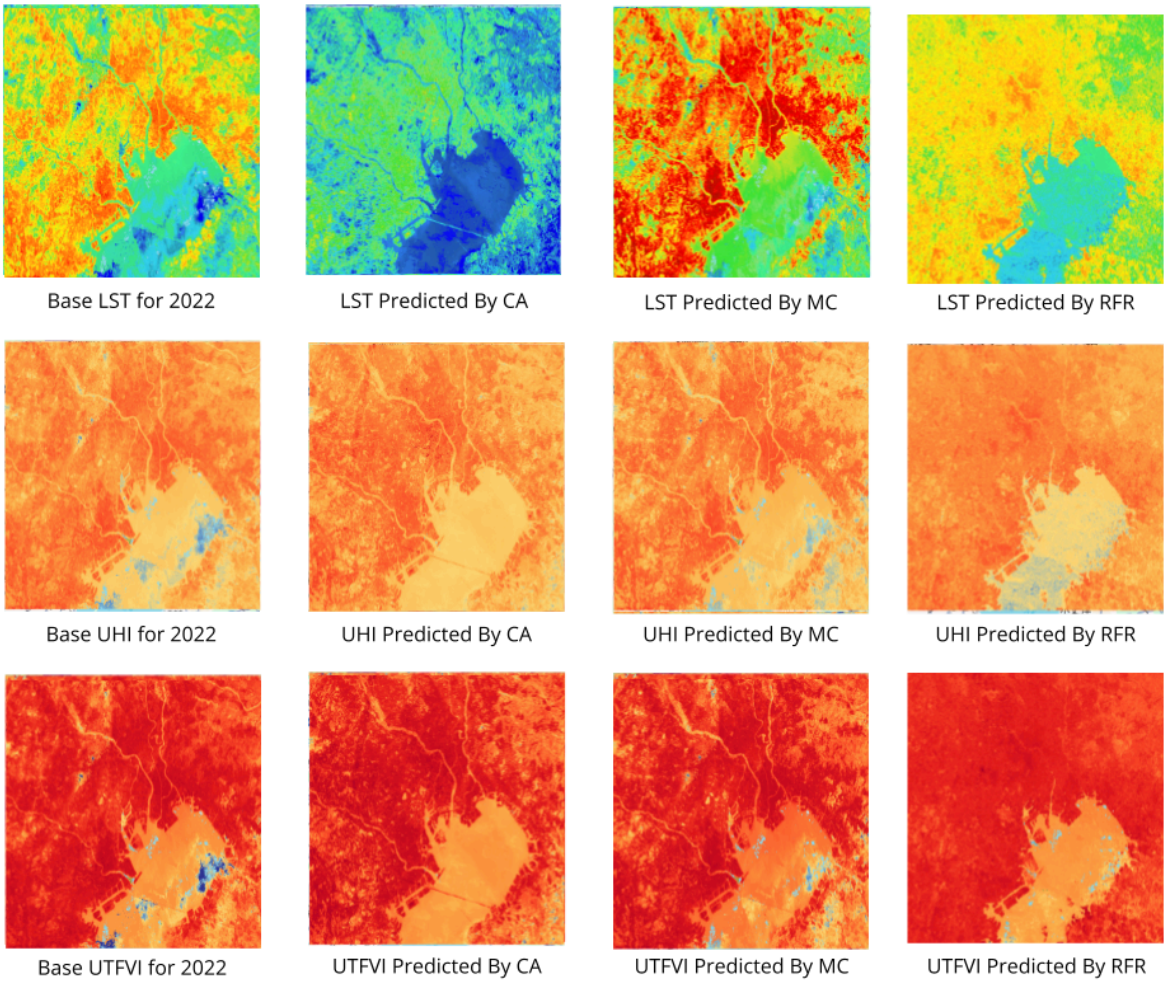


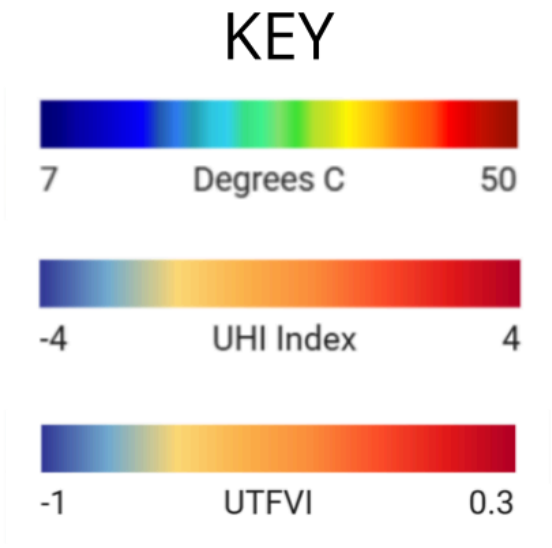
UTFVI Predicted By MC

UTFVI Predicted By RFR

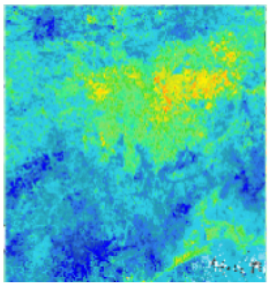


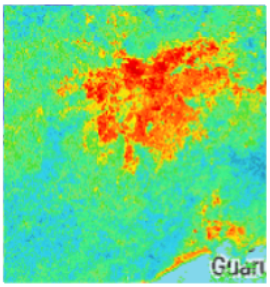
Tokyo, Japan

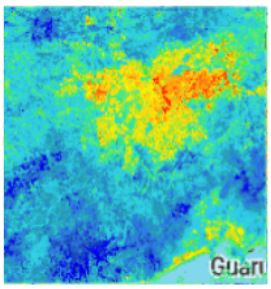

Base LST for 2022
LST Predicted By CA
LST Predicted By MC
LST Predicted By RFR
Base UHI for 2022
UHI Predicted By CA
UHI Predicted By MC
UHI Predicted By RFR
Base UTFVI for 2022
UTFVI Predicted By CA
UTFVI Predicted By MC
UTFVI Predicted By RFR

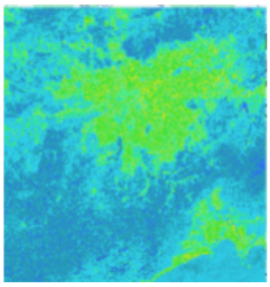

KEY
7
Degrees C
50
-4
UHI Index
4
-1
UTFVI
0.3

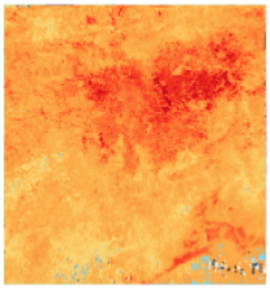
São Paulo, Brazil

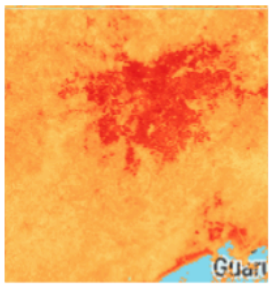

Base LST for 2022

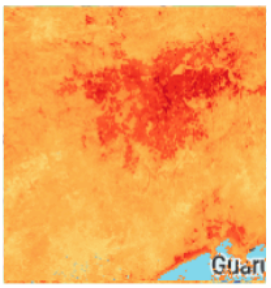

LST Predicted By CA

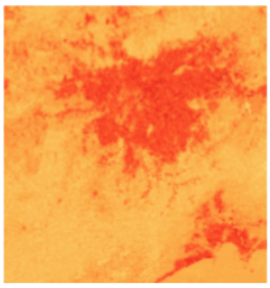

LST Predicted By MC

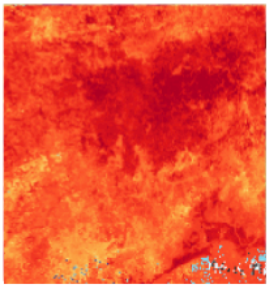

LST Predicted By RFR

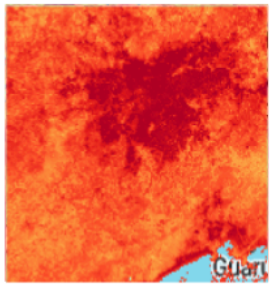

Base UHI for 2022

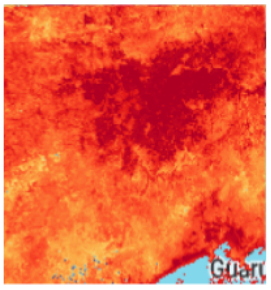

UHI Predicted By CA

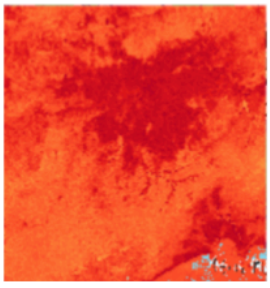

UHI Predicted By MC

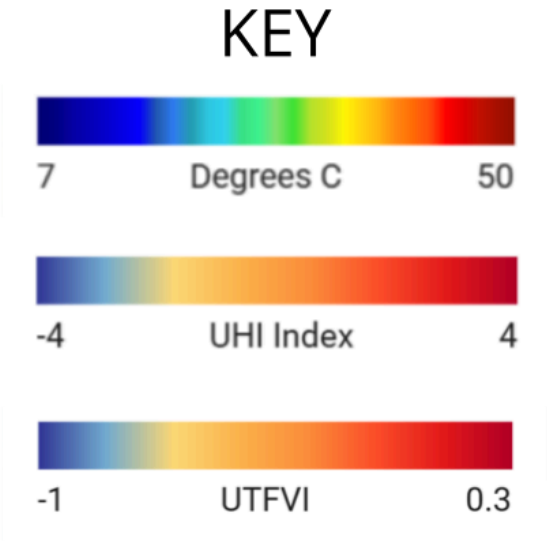

UHI Predicted By RFR


Base UTFVI for 2022

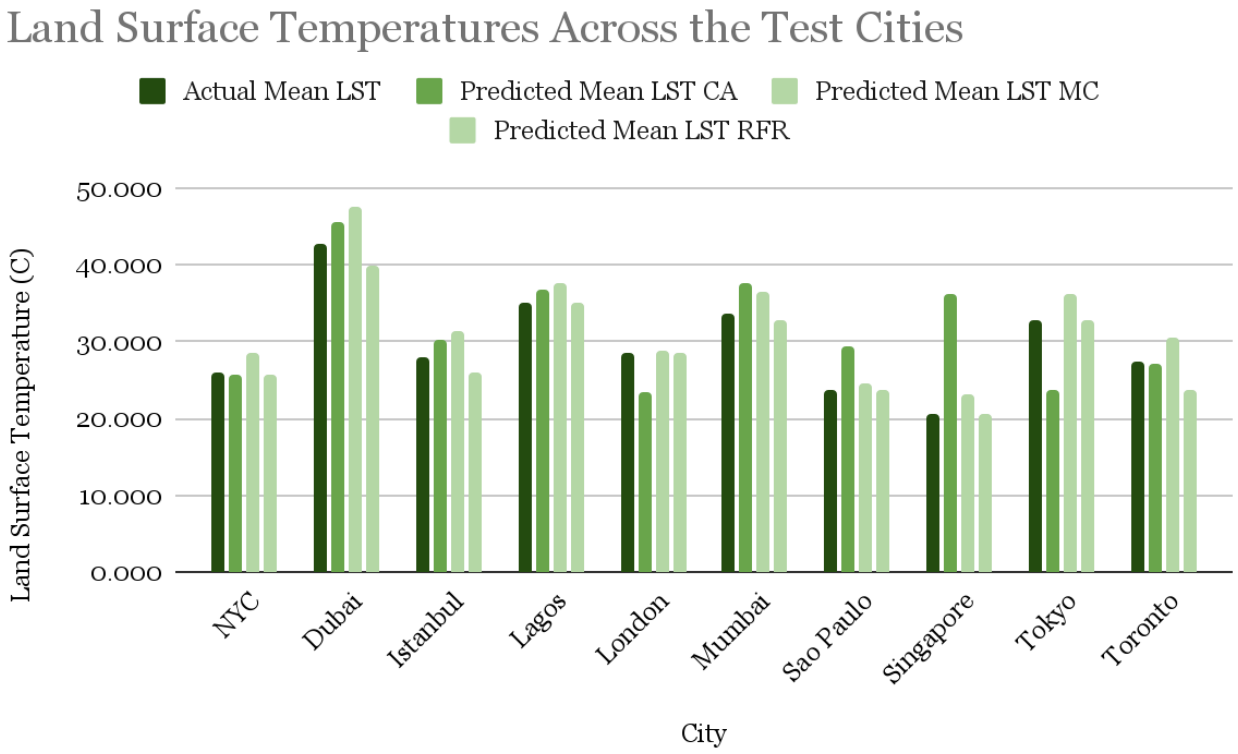

UTFVI Predicted By CA


UTFVI Predicted By MC


UTFVI Predicted By RFR

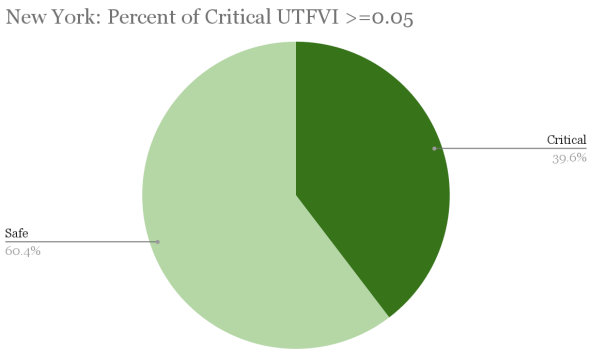


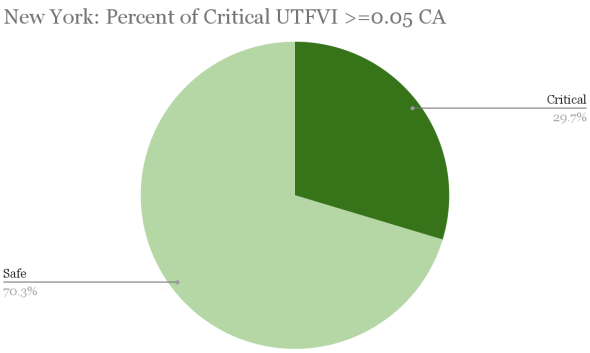
Land Surface Temperature Averages

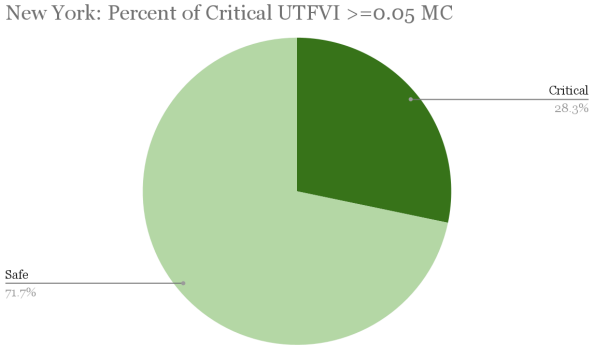


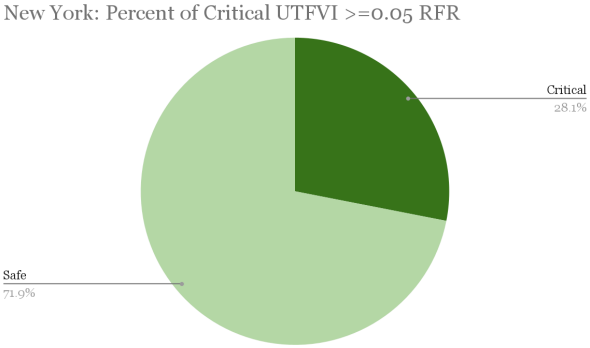
Percent of Critical UTFVIs

New York



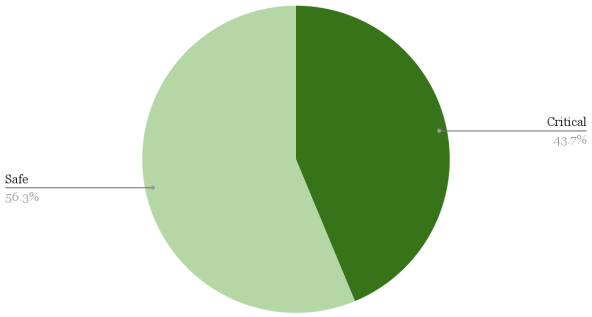




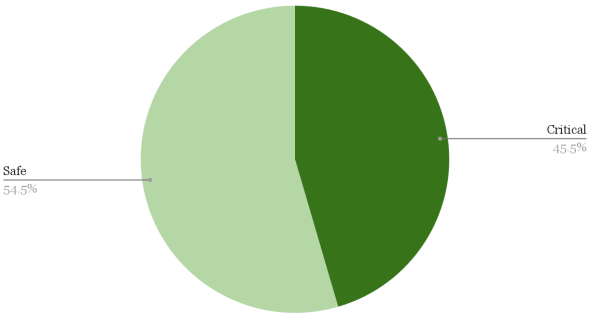


Dubai

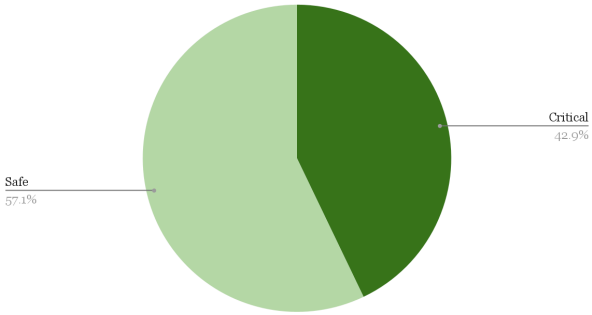
Dubai: Percent of Critical UTFVI >=0.05



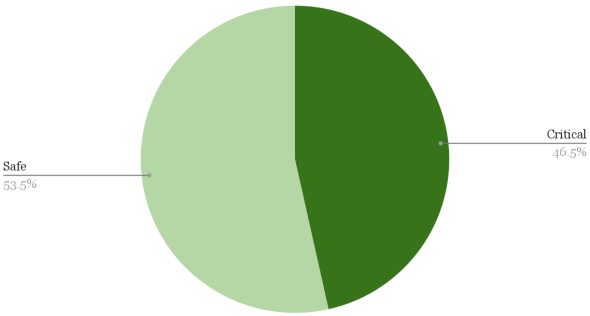
Dubai: Percent of Critical UTFVI >=0.05 CA



Dubai: Percent of Critical UTFVI >=0.05 MC

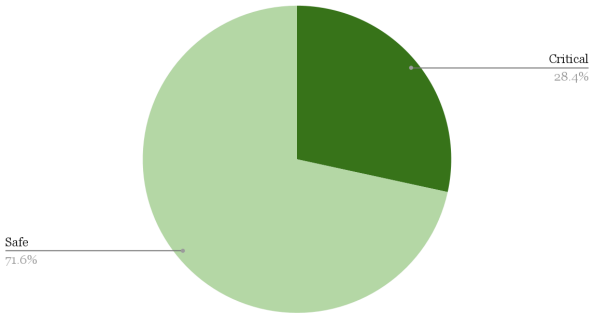


Dubai: Percent of Critical UTFVI >=0.05 RFR

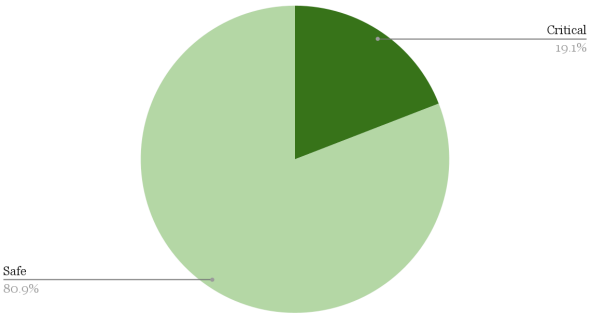


Istanbul

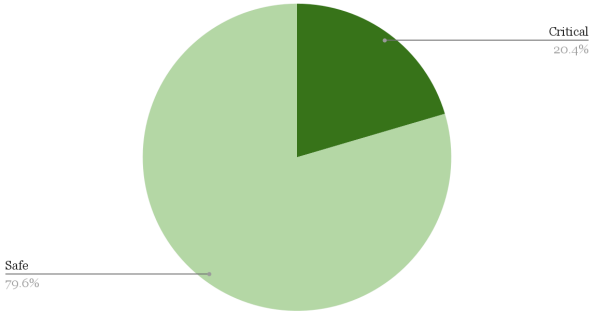
Istanbul: Percent of Critical UTFVI >=0.05



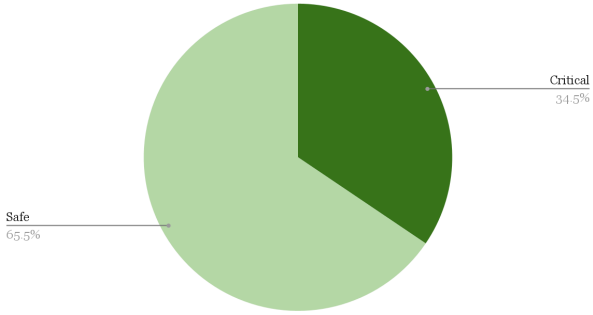
Istanbul: Percent of Critical UTFVI >=0.05 CA



Istanbul: Percent of Critical UTFVI >=0.05 MC

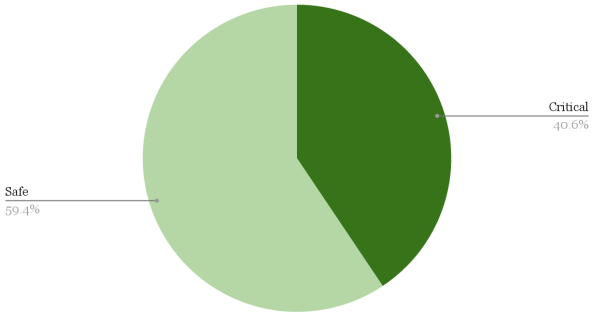


Istanbul: Percent of Critical UTFVI >=0.05 RFR

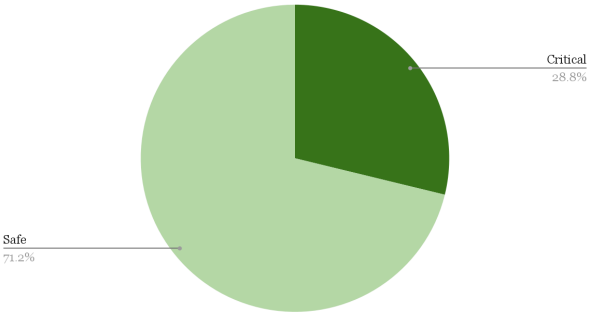


Lagos

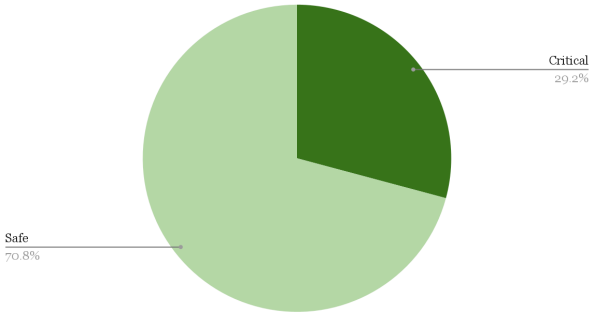
Lagos: Percent of Critical UTFVI >=0.05



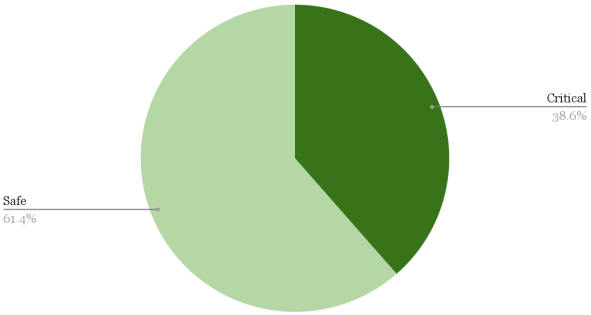
Lagos: Percent of Critical UTFVI >=0.05 CA



Lagos: Percent of Critical UTFVI >=0.05 MC

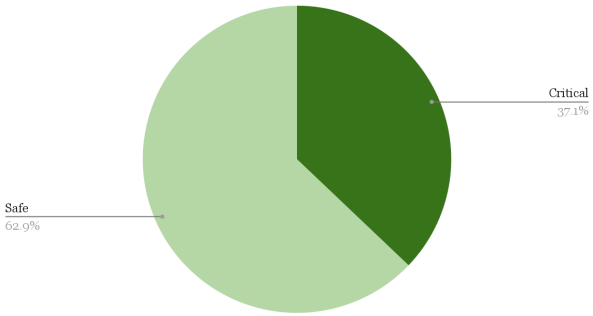


Lagos: Percent of Critical UTFVI >=0.05 RFR

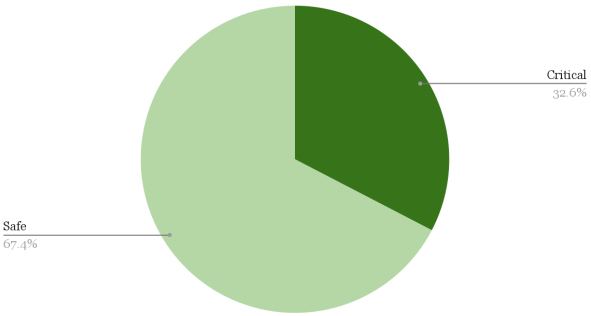


London

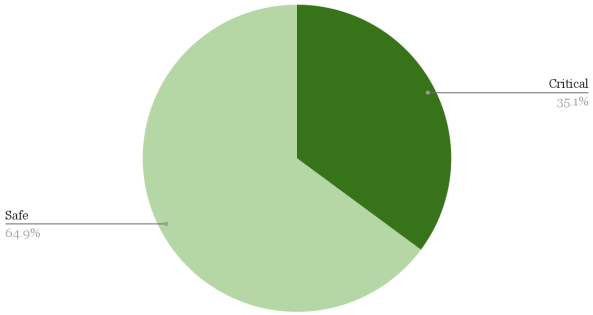
London: Percent of Critical UTFVI >=0.05



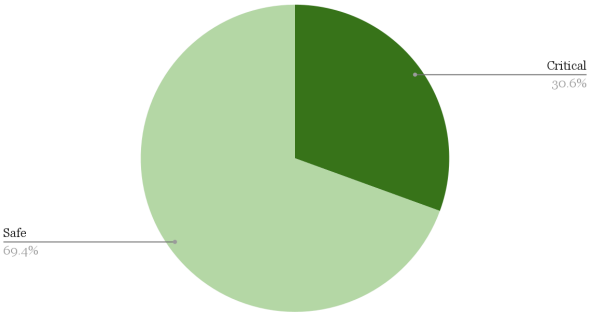
London: Percent of Critical UTFVI >=0.05 CA



London: Percent of Critical UTFVI >=0.05 MC

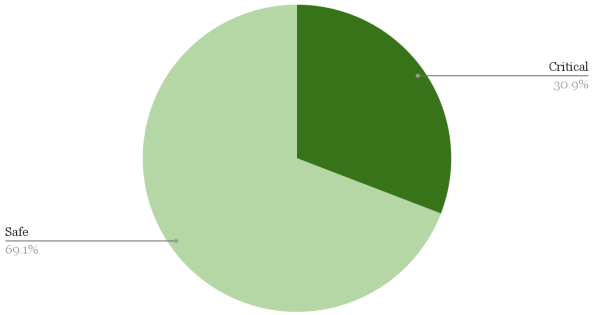


London: Percent of Critical UTFVI >=0.05 RFR

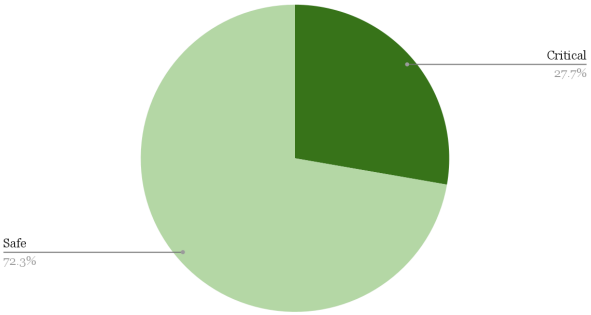


Mumbai

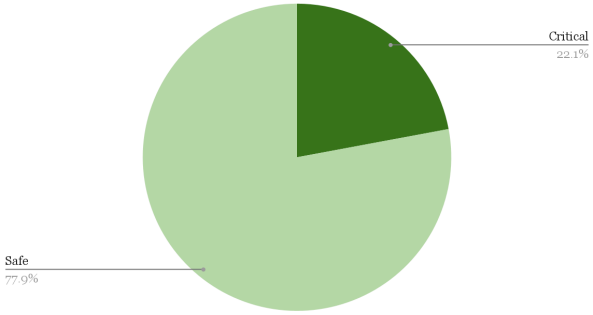
Mumbai: Percent of Critical UTFVI >=0.05



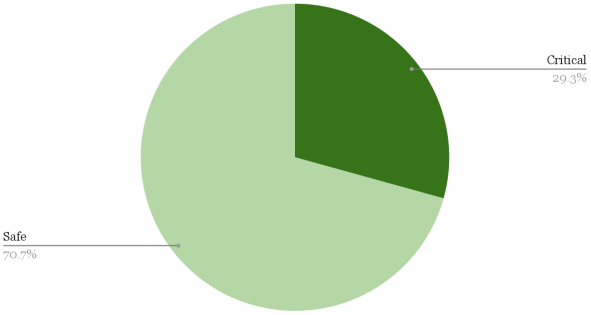
Mumbai: Percent of Critical UTFVI >=0.05 CA



Mumbai: Percent of Critical UTFVI >=0.05 MC

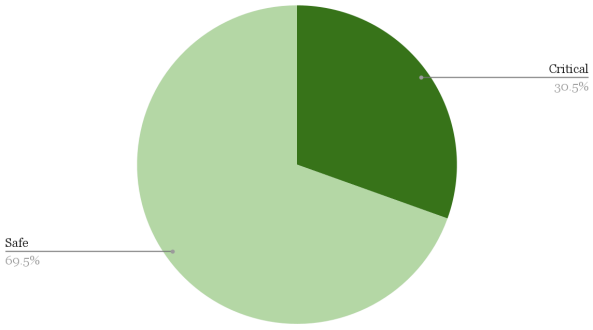


Mumbai: Percent of Critical UTFVI >=0.05 RFR

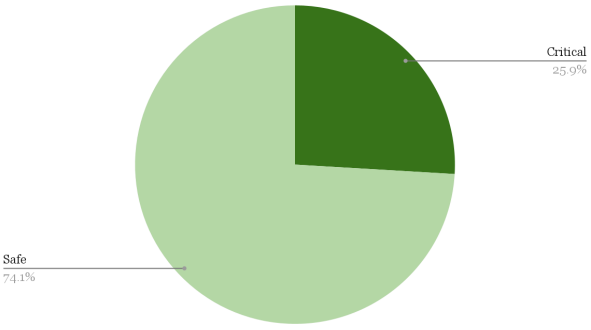


Sao Paulo

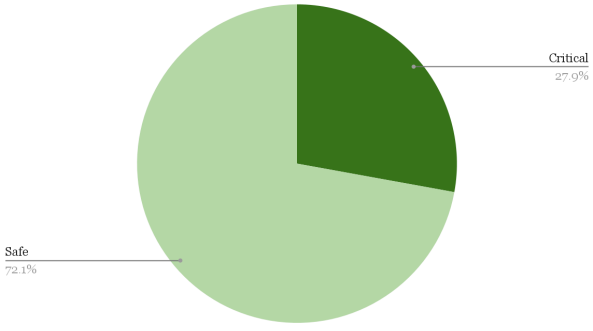
Sao Paulo: Percent of Critical UTFVI >=0.05



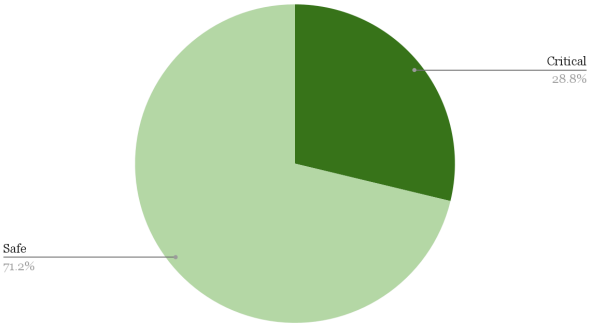
Sao Paulo: Percent of Critical UTFVI >=0.05 CA



Sao Paulo: Percent of Critical UTFVI >=0.05 MC

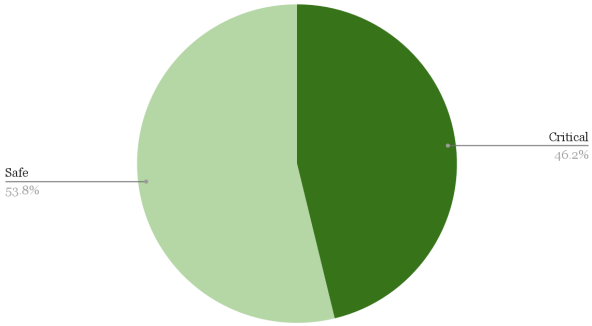


Sao Paulo: Percent of Critical UTFVI >=0.05 RFR

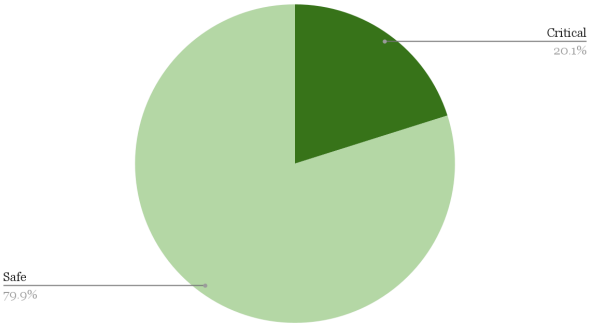


Singapore

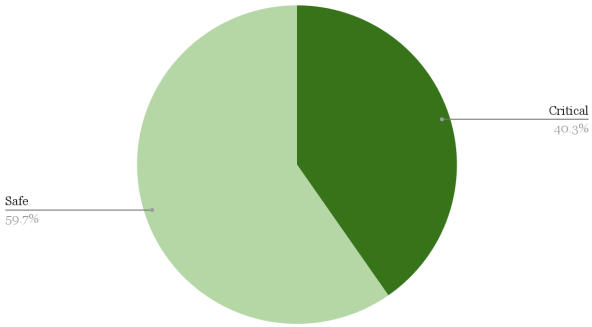
Singapore: Percent of Critical UTFVI >=0.05



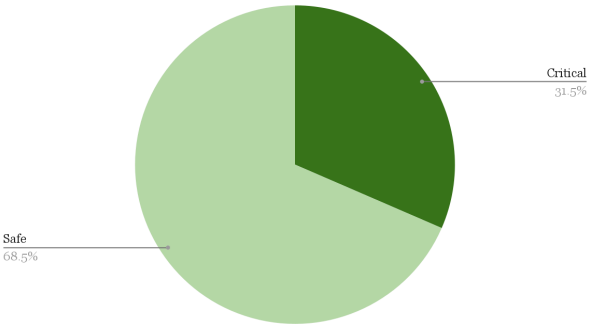
Singapore: Percent of Critical UTFVI >=0.05 CA



Singapore: Percent of Critical UTFVI >=0.05 MC

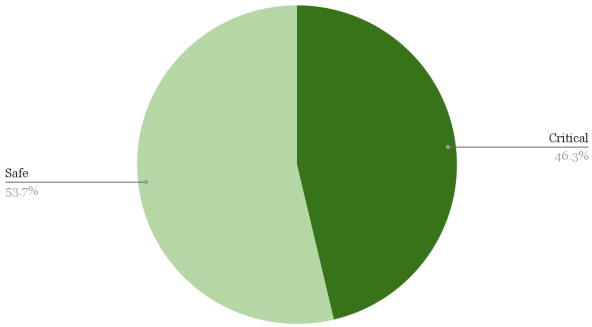


Singapore: Percent of Critical UTFVI >=0.05 RFR

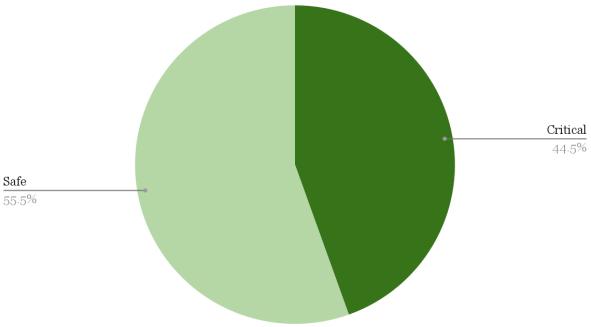


Tokyo

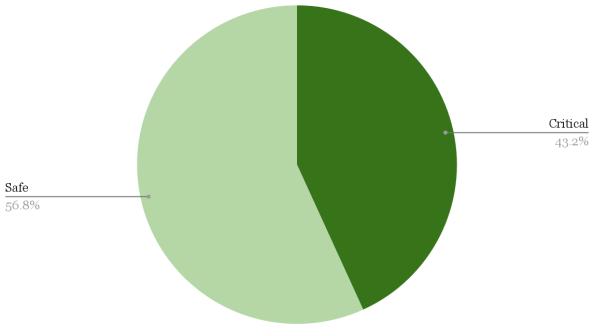
Tokyo: Percent of Critical UTFVI >=0.05



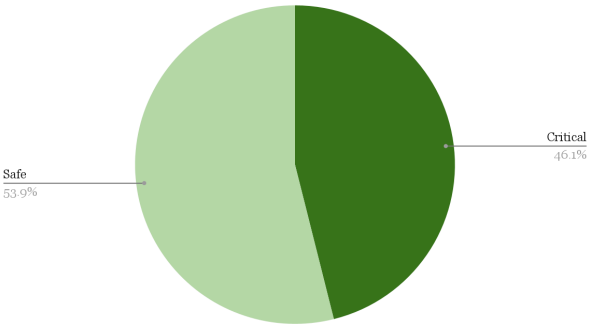
Tokyo: Percent of Critical UTFVI >=0.05 CA



Tokyo: Percent of Critical UTFVI >=0.05 MC

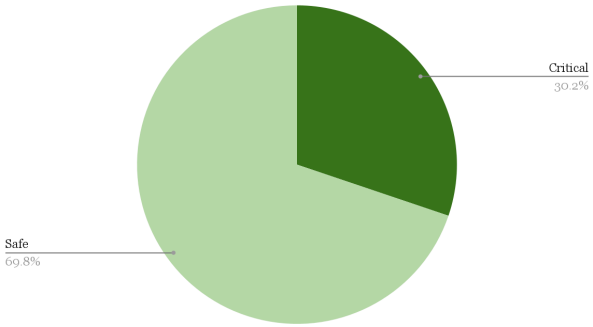


Tokyo: Percent of Critical UTFVI >=0.05 RFR

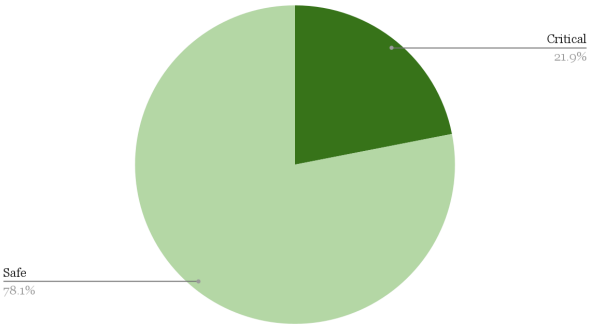


Tokyo

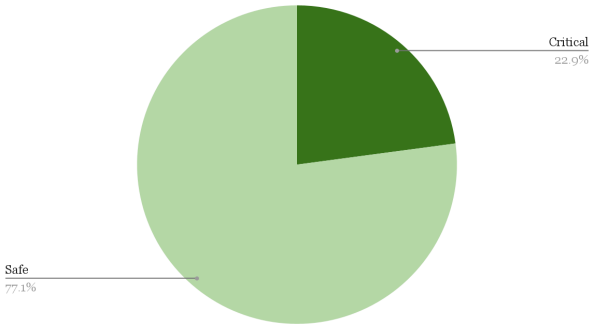
Toronto: Percent of Critical UTFVI >=0.05



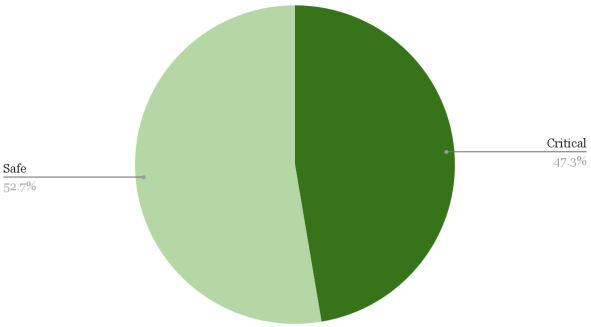
Toronto: Percent of Critical UTFVI >=0.05 CA



Toronto: Percent of Critical UTFVI >=0.05 MC



Toronto: Percent of Critical UTFVI >=0.05 RFR



Conclusion

RFR

The Random Forest Regressor worked exceptionally well for every city, averaging an accuracy of over ninety percent across all cities. The lowest accuracies were in the cities of Lagos, Dubai, and Mumbai. The cause of this lower accuracy with those cities is attributed to those imbalanced classes. These cities have much higher temperatures than the rest and do not necessarily have a wide variety of temperatures across the city due to a lack of green space, the overall climate of the city, etc. However, most cities did have a wide variety of temperatures which allowed the Random Forest to have balanced training data to predict the 2022 Land Surface Temperature. While this is a good approach to predict the Land Surface Temperature, there are some issues. It would also be helpful to add more parameters for training the model than just the NDVI and EM. This would allow for even more accurate results using Random Forest.

Cellular Automation

For cellular automation, my hypothesis about which cities would see better results was partially correct. For cellular automation, New York City saw very comparable results to the actual results, being within 0.25 degrees of the actual temperatures. This was also the case for Istanbul, Lagos, Toronto, and Dubai as they all were within one to three degrees of the actual

temperature. However, for cities such as London, São Paulo, Singapore, and Tokyo, the model struggled quite significantly. Especially with Singapore, the model was off by over fifteen degrees. While most of these results were expected, my hypothesis was incorrect on New York City working well with the cellular model. I first expected the city to perform well with the Markov Chain model as its growth had been stunted and the slower, probability based growth model would perform better. However, in New York, urban growth is caused in "neighborhoods" and those spatial patterns are better represented by cellular automation as the model is based on neighborhood growth versus probability where the neighborhood is not taken into consideration. This model however does have its pros and cons. The model is really good at showing and simulating the chaotic growth of cities such as Lagos which have largely informal settings. However, with this chaos comes unpredictability where it can show large variability varying by cities. For example with New York City, it was very accurate down to the nearest degree. However, in Singapore, cellular automation struggled and predicted a mean temperature nearly sixteen degrees higher than it should have been.

Markov Chain Model

For the Markov Chain Model, my hypothesis about which cities would see better results was partially correct. For cities such as London, the Markov Chain model was within less than a degree of the actual land surface temperature. On the contrary, for cities such as Dubai and New York City, it struggled more than the Cellular Automation model. While, again, most of these results were expected, the model was unexpectedly better for the cities of Mumbai and Sao Paulo. Initially, the prediction was that Mumbai was going to perform better with cellular

automation due to its unpredictable growth behavior, however, the Markov Chain model proved to be slightly better. This is likely attributed to Mumbai's geography. Mumbai is an island surrounded all by water and mountains. This causes it to have limited space to grow and the city is becoming oversaturated with people and mass urbanization. For Sao Paulo, it is a similar case. The city is also starting to see more and more oversaturation of places to urbanize as it continues to follow the trend of many places in Europe and the United States. Overall, the Markov Chain Model is, from the perspective, of this study, better than cellular automation. This is because the Markov Chain model produces more precise and accurate results which are overall closer to the target value compared to cellular automation which was either really close or really far, meaning the model does not produce many outliers. It is able to accurately predict the urban heat island in already urbanized cities such as in Tokyo or London and it still does a good job in predicting the Urban Heat Islands in oversaturated urban areas.

References

Links

1. https://www.epa.gov/heatislands
2. https://waleedgeo.com/papers/waleed2023_sustainability.pdf
3. https://mathworld.wolfram.com/CellularAutomaton.html
4. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
5. https://www.khanacademy.org/computing/computer-science/informationtheory/moderninfotheory/v/markov_chains
6. https://www.ucl.ac.uk/bartlett/planning/publications/2020/dec/critical-dialogues-urban-governance-development-and-activism-london-and
7. https://www.researchgate.net/publication/330289683_Strategic_Planning_and_Urban_Development_in_New_York_City_Paris_and_Sao_Paulo_2001_-_2012
8. https://www.mckinsey.com/capabilities/sustainability/our-insights/elements-of-success-urban-transportation-systems-of-24-global-cities
9. https://www.researchgate.net/publication/353296921_International_Collaborative_Research_Smart_Global_Mega_Cities_and_Conclusions_of_Cities_Case_Studies_Tokyo_New_York_Mumbai_Hong_Kong-Shenzhen_and_Kolkata
10. https://www.sciencedirect.com/science/article/abs/pii/S0198971514000210

Technical Resources

1. Google Earth Engine Documentation

   https://developers.google.com/earth-engine

2. Towards Sustainable and Livable Cities: Leveraging Remote Sensing, Machine Learning, and Geo-Information Modelling to Explore and Predict Thermal Field Variance in Response to Urban Growth

   https://waleedgeo.com/papers/waleed2023_sustainability.pdf

3. LandSat Dataset

   https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1_L2

Prediction of Urban Heat Islands Using Machine Learning, Markov Chain Modeling, and Cellular Automation

Sadhil Mehta
12th Grade - Tippecanoe High School - Tipp City OH 45371

Abstract

As the world moves into the future, urbanization has reached unprecedented rates, leading to a rise in the Urban Heat Island effect. The Urban Heat Island effect is where a metropolitan area is hotter than the surrounding area due to human activity. This effect can lead to many economic concerns and health hazards. It is paramount to predict and monitor the spread of this effect. I aimed to predict the Urban Heat Islands in various cities across the world using the following combination of methods:

- Random Forest Regressor (RFR)
- Markov Chain Modeling (MC)
- Cellular Automation (CA)

The 2017 Land Surface Temperature, along with various other measures was fed into Random Forest Regressor, a machine learning method based on decision trees, a Markov Chain model, a simulation model based on probability, and a cellular automaton model, a simulation model based on neighborhood evolution. These approaches were assessed upon ten testing cities of various factors (eg. population, climate, urbanization rate, etc.) by analyzing the Urban Heat Island Effect (UHI) and the Urban Thermal Field Variance Index (UTFVI). Results mostly aligned with my expectations—MC worked well in stable cities, CA performed better in chaotic urban environments, while RFR worked well regardless of city. However, there was some variability in the results due to oversaturation of the cities, but overall, the methods were capable of

predicting Urban Heat Islands well, providing insight into the behaviors of UHIs in new areas of the world.