Guidance and Suggestions for Participants of the EY Urban Heat Island Data Challenge

University of Maryland – Information Challenge 2025 (IC25)
March 1-8, 2025

EY Background

EY is a global professional services organization that exists to build a better working world, helping create long-term value for clients, people and society and build trust in the capital markets. Working across assurance, consulting, law, strategy, tax and transactions, EY teams ask better questions to find new answers for the complex issues facing our world today. Each year, EY runs a global data challenge focused on AI and global sustainability issues. The 2025 version (HERE) is open from January 20 to March 20. Feel free to join our challenge!

IC25 Data Challenge Background

This document provides background, guidance and suggestions for participants working on the EY Urban Heat Island Data Challenge as part of the 2025 University of Maryland (UMD) Information Challenge (IC25). The IC25 data challenge is NOT the same as the global EY Data Challenge (noted above) but there are similarities. Those similarities and differences will be mentioned in the remaining portion of this document.

This challenge focuses on a phenomenon known as the Urban Heat Island (UHI) effect. This issue can result in temperature variations between rural and urban environments that exceed 10-degrees Celsius in some cases and can cause significant health-, social-, and energy-related issues. In many industrialized countries, heat events account for more than all other natural hazards combined. [References 1,2] Urban areas are most susceptible to heat stress due to the high density of buildings, lack of vegetation (green space), lack of water bodies, and waste heat from industry and transportation. Those particularly vulnerable to heat-related problems include young children, older adults, outdoor workers, and low-income populations. According to the World Health Organization (WHO), over 55% of the world's population live in urban areas with that proportion expected to increase to 68% by 2050. According to the Intergovernmental Panel on Climate Change (IPCC), climate change is expected to cause increasing temperatures which combined with population increases will put more citizens in danger from the negative health effects of extreme heat.

Challenge Goals

Though this topic has been widely studied and documented it is still unknown to many people and there is a need for increased awareness and accurate open-source models. The Urban Heat Island (UHI) effect is typically modeled using coarse satellite-based measurements of surface temperatures such as those produced by NASA's Landsat mission. Such models do not reflect the near-surface micro-climate air temperature that most impact the human population. Therefore, there is a need for simplified, yet accurate, UHI models to allow urban planners and city managers to better understand the location and severity of UHI issues in their city and the drivers of urban heating. Such models will bring attention to this important global sustainability problem and may force decisions to change existing urban plans or influence urban planning for the future. In addition, such models can be used to understand the impact of natural areas and support their preservation and future expansion.

The primary goal of this data challenge will be to develop a digital model to predict the locations and severity of the UHI effect over a region in Montgomery County, Maryland and to understand the drivers of this phenomenon. This machine learning model will be developed using near-surface air temperature data in an index format (target dataset) and optical satellite data (feature data) from the European Sentinel-2 mission. The satellite data can be used to assess the impact of vegetation, water and urbanization on local urban heating. A secondary goal of the challenge is to address the practical application of the output model for local decision-makers, scaling such solutions to other locations, considerations for additional datasets that could improve model accuracy, and socioeconomic impact.

Challenge Approach

Participants will be provided with two core datasets to build their models. These datasets include ground-level air temperature collected over the region on a single day in 2022 and multispectral satellite data from the European Sentinel-2 mission. These datasets will be used to develop a machine learning model to predict UHI "hotspots" at micro-scales (meters) across the city. Additionally, the model should be designed to discern and highlight the key features that contribute significantly to the existence of these UHI "hotspots" within city environments.

Participants will be given ground-level air temperature data in an index format, which was collected in the late afternoon on August 7, 2022 by CAPA Strategies using automobile traverses across Montgomery County, Maryland. This dataset includes traverse points (latitude and longitude) and the corresponding UHI Index values for 34,502 data points. These UHI Index values will be the "target" parameter dataset for your model. Participants will then use Sentinel-2 satellite data as the "feature" dataset within their model. As a note, participants are NOT allowed to use any additional datasets for their model. This is different than the EY global data challenge but simplifies the data challenge for completion within a 1-week time window required for the IC25 session.

Required Skills

Participants in this challenge can benefit from a basic understanding of data science and Python programming, but there are no prerequisites for participation. Participating in this challenge will improve one's skills in machine learning, Artificial Intelligence (AI), data science, and working with satellite datasets. The data challenge has been designed to entice beginners and those less familiar with AI and Python programming.

Computing Requirements

This data challenge was designed to run on a local computer with common computing resources (e.g., 4 cores, 32 GB memory). The configuration should include a Python programming environment and a code development tool (e.g., Jupyter). It is also possible to participate in this challenge using common cloud-based environments, such as those available from Microsoft (Azure), Google (Google Cloud, Earth Engine) or GitHub (Codespaces). For more information about setting up a local Python computing environment and a Jupyter notebook development environment, please review the document on this subject in your package.

Participants will be given core datasets for training and testing their models, a sample benchmark Python notebook, a sample Sentinel-2 satellite Python notebook, and a document addressing Python and Jupyter environments. To get started, participants should create their notebook environment and ensure they are able to run the two sample notebooks to completion without any execution issues. After completing this step, participants should then consider steps to improve their model.

Model Development and Evaluation

Participants will develop a machine learning / artificial intelligence (AI) model [example in Reference 5] that can accurately predict UHI index values at specific locations. To get started, participants are provided with a sample benchmark Python notebook that will demonstrate a simple UHI prediction model. This sample model is designed to use UHI index data from the "target" dataset that was compiled from ground traverse data. Sentinel-2 satellite data spectral bands are used as the "feature" dataset in the sample model. The model uses a common 70/30 training and testing split to evaluate model performance. The sample model produces poor results (R-squared score of ~0.10) to allow significant improvements by data challenge participants. Some suggestions for improved model performance include consideration of: Sentinel-2 statistical band combinations or spectral indices, proximity to vegetation, proximity to water, proximity to dense urbanization, bounding box sizes around target data locations, regression algorithms, and hyperparameter tuning. As a note, participants are NOT allowed to use additional datasets for their model and may only use Sentinel-2 satellite data. In addition, participants are prohibited from using latitude and longitude as "feature" variables in their model. Using these data will instantly produce very high scores (>0.95) but the results are a spatially autocorrelated model that is not applicable to other regions and thus not generalized. The goal of your model is to use only satellite data as a driver of urban heat island response.

In the end, participant models will be tested against known UHI index values (validation dataset) for specific locations not included in the target dataset. Predictions on the validation dataset shall be saved in a CSV file and uploaded to the challenge platform to get a score on the ranking board, which you can improve over the course of the challenge with subsequent model revisions and submissions. Model performance will be based on the coefficient of determination (R-squared) score for your model.

Business Plan Development and Evaluation

In addition to completing a UHI model, participants will be required to develop a practical "business plan" that describes how their AI model could be applied by local city managers to address the impacts and concerns of urban heating. This business plan, along with a summary of the AI model approach and performance, shall be included in the final presentation to be evaluated on March 8.

Participants should consider the following in their presentation:

- How might local governments or urban planning decision-makers use your model?
- How might your model and approach be scaled to other cities?
- How could your model be used to address socioeconomic impact on vulnerable communities?
- How could your model be used to address impacts on energy demand?
- What additional datasets might you consider for improving model accuracy if given more time and resources

Participants should use a strategic and well-structured approach while infusing creativity and considering generative-AI tools for completeness and enhanced impact.

Participant Package Contents

1. Participant Overview (**IC25_Challenge_Overview.docx**) – this document

2. Benchmark Notebook (**Benchmark_IC25.ipynb**) – a Python notebook that provides a starting point for participants

3. Sentinel-2 Sample Notebook (**Sentinel2_IC25.ipynb**) – a Python notebook that outputs a GeoTIFF file with specific spectral bands. This output is used by the Benchmark Notebook.

4. Training Dataset (**Training_Data_IC25.csv**) – a CSV file containing **34502** locations (latitude and longitude) and their corresponding Urban Heat Island (UHI) index values for model training.

5. Submission Template (**Submission_Template_IC25.csv**) – a CSV file containing **3834** locations (latitude and longitude) for participant model validation testing.

6. Python and Jupyter Setup and Operation (**Python_Jupyter.docx**) – a document summarizing how to set up Python computing and Jupyter Lab notebook development environments.

Instructions for Participants / Students

a) Understand the problem statement

The objective of this data challenge will be to develop a digital model to predict the locations and severity of the UHI effect and to understand the drivers of this phenomenon. This machine learning model will be developed using near-surface air temperature data in an index format (target dataset) and optical satellite data (feature data) from the European Sentinel-2 mission. The satellite data can be used to assess environmental conditions such as proximity to vegetation (green space), proximity to water, and local urban density which are all known to contribute to the effects of urban heating.

b) Review the datasets

This challenge will use temperature data collected over the test region in 2022 which has been converted to an Urban Heat Island (UHI) index. Please review the file "*Training_Data_IC25.csv*" to understand the content of this training dataset. This dataset consists of 34,502 locations (latitude and longitude) with a UHI index for each location.

c) Set up the development environment

Students should review the file "*Python_Jupyter.docx*" for suggestions on how to set up a Python environment and a Jupyter Lab notebook development environment. There are also

some helpful Jupyter Lab operational tips. Though a paid cloud computing environment will work, it is not necessary for this challenge.

d) Run the sample Sentinel-2 notebook

Students should first run the sample Sentinel-2 satellite data notebook (Sentinel2_IC25.ipynb) to create a simple GeoTIFF output file for specific spectral bands. This output is used by the Benchmark Notebook. Prior to running the sample notebook for the first time, it is likely the Python environment will need to install specific libraries to perform the required code functions. Students should review the first block of code to review the list of required Python libraries. For example, these libraries include: numpy, xarray, matplotlib, rasterio, and planetary_computer. After successfully running the notebook, the student should create a GeoTIFF output file that will be used by the benchmark notebook (next step).

e) Run the Benchmark Model notebook

The benchmark notebook (Benchmark_IC25.ipynb) should be the starting point for a student model. We have demonstrated a basic machine learning workflow, designed to help students gain practical experience with real-world datasets. This example uses several features from the Sentinel-2 satellite dataset as predictor variables to produce a moderate result. This notebook is simplified and intended as a learning tool for students to experiment with remote sensing data and machine learning techniques. This notebook should be considered as a starting point only and students should extend this model to build a more robust model by running multiple experiments.

IMPORTANT REQUIREMENT: Using latitude and longitude as "feature" variables in your model is prohibited. Using these data will quickly produce a very high score (>0.95) but the results are a spatially autocorrelated model that is not applicable to other regions and thus not generalized. The goal of your model is to use only satellite data as a driver of urban heat island response.

f) Make the prediction on the validation dataset

After building the model, students need to validate the model performance on the dataset of different locations. To validate the model, students need to use their model to make UHI index predictions for specific locations found in the "Submission_Template_IC25.csv" file.

The final submissions (Predicted_Data_IC25.csv) should be uploaded to the EY Data Challenge platform. Please register at https://challenge.ey.com and then select the dedicated University of Maryland IC25 Data Challenge. This will allow you to upload your CSV file with your model predictions. The system will automatically score your model against a "ground truth" dataset and give you an R-squared regression score on the leaderboard. A maximum of 5 submissions per day are allowed. As a reference, the sample notebooks will produce a starting R-squared score of about 0.10, which is quite low.

Dataset Detailed Information

Temperature Data

Data was collected by CAPA Strategies using a ground traverse (Figures 1 and 2) with vehicles and bicycles on August 7, 2022. This data collection effort resulted in 34,502 data points which will be the focus for this data challenge.

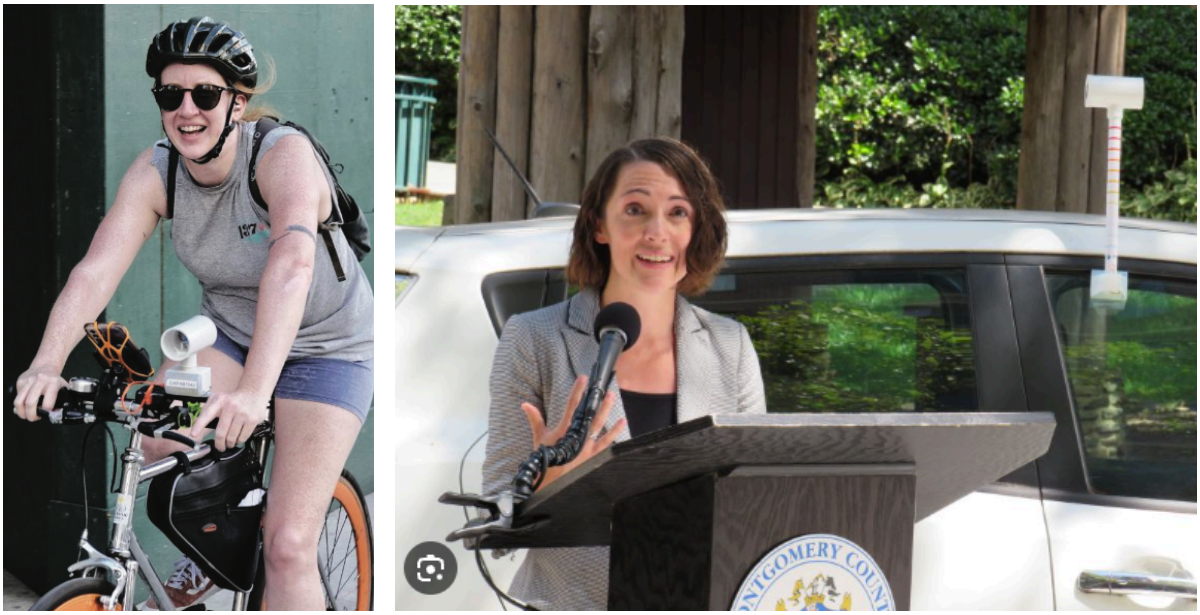

Figure 1*. Ground-level temperature data was collected by CAPA Strategies and community volunteers using temperature recording devices mounted to cars and bikes. This data collection campaign was part of the international "Heat Watch" program. Credit: CAPA Strategies, LLC.*

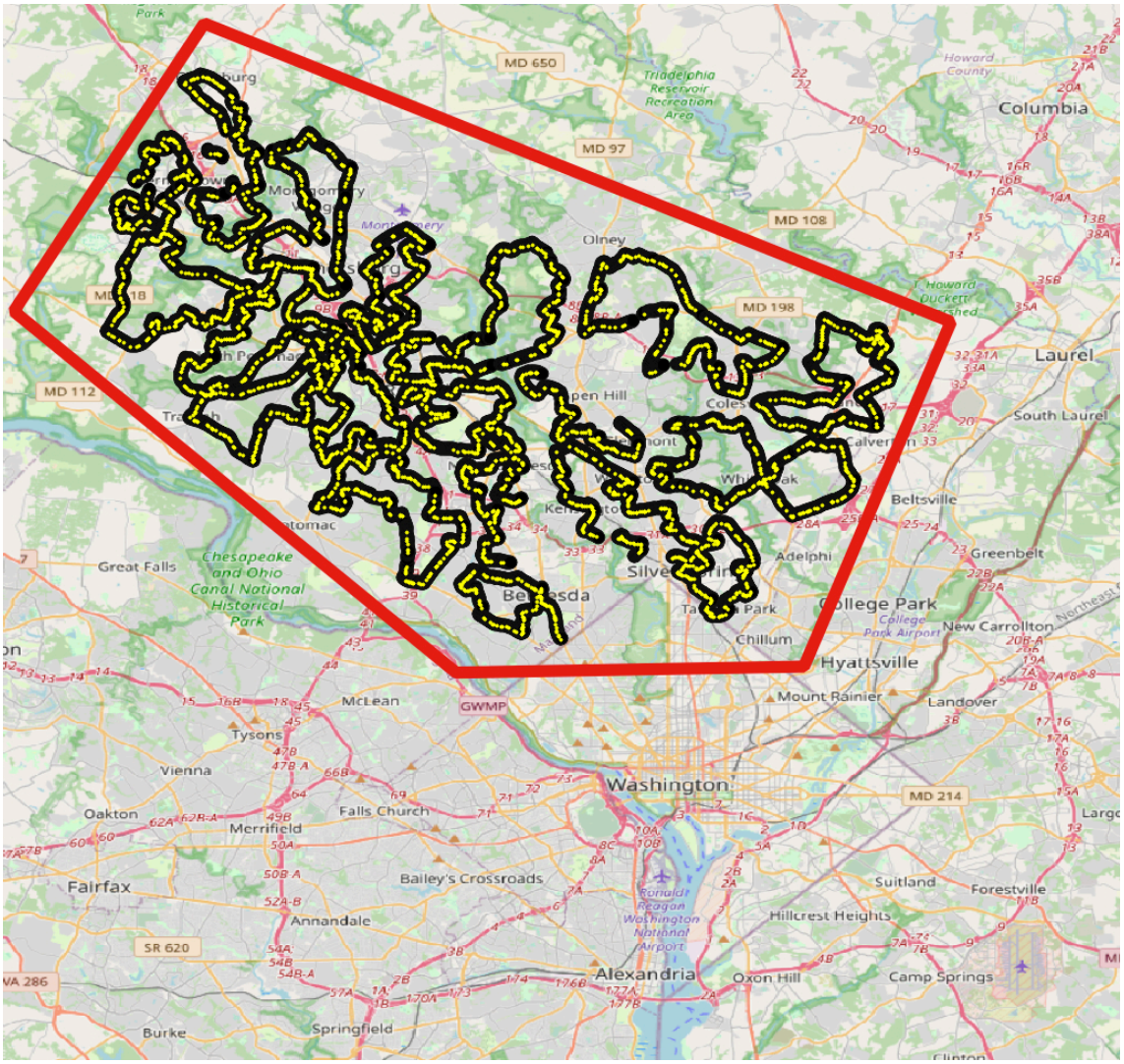

Figure 2*. Data was collected across Montgomery County, Maryland on August 7, 2022. The data (34,502 points) was converted to a UHI Index for the purpose of this data challenge. The image above shows the ground transect paths. Credit: Brian Killough, EY*

For this challenge, we have created a unique "UHI Index" for every data point location. This index reflects the local temperature at the data point location compared to the city's average temperature across all data points during the time window of the data collection. Though this is not a "perfect" approach to modelling the complex urban heating dynamics of a city, it will provide a reasonably accurate model of urban heat islands in the city. In an ideal situation, time series data would be collected at thousands of locations across the city and weather data (e.g., wind speed, wind direction, solar flux) would be added to the model to yield more accuracy and allow consideration of natural variability.

UHI Index = (Temperature at a given location) / (Mean Temperature for all locations)

The chosen UHI Index serves as a crucial metric for assessing the intensity of heat within different urban zones of the city. For comparison, most literature calculates a UHI Index based on temperature differences between inner city locations and rural locations far outside of the city. Since we did not have data from rural locations, we decided to create a unique UHI Index that reflects the variability of temperatures within our collected dataset and time collection window. As an example, a UHI Index value of 1.0 suggests the local temperature is the same as the mean temperature of all collected data points. UHI Index values above 1.0 are consistent with "hotspots" above mean temperature values and UHI Index values below 1.0 are consistent with "cooler" locations in the city. Participants will use their model to predict these UHI values across the city. Figure 3 shows a histogram of UHI values for the data challenge.

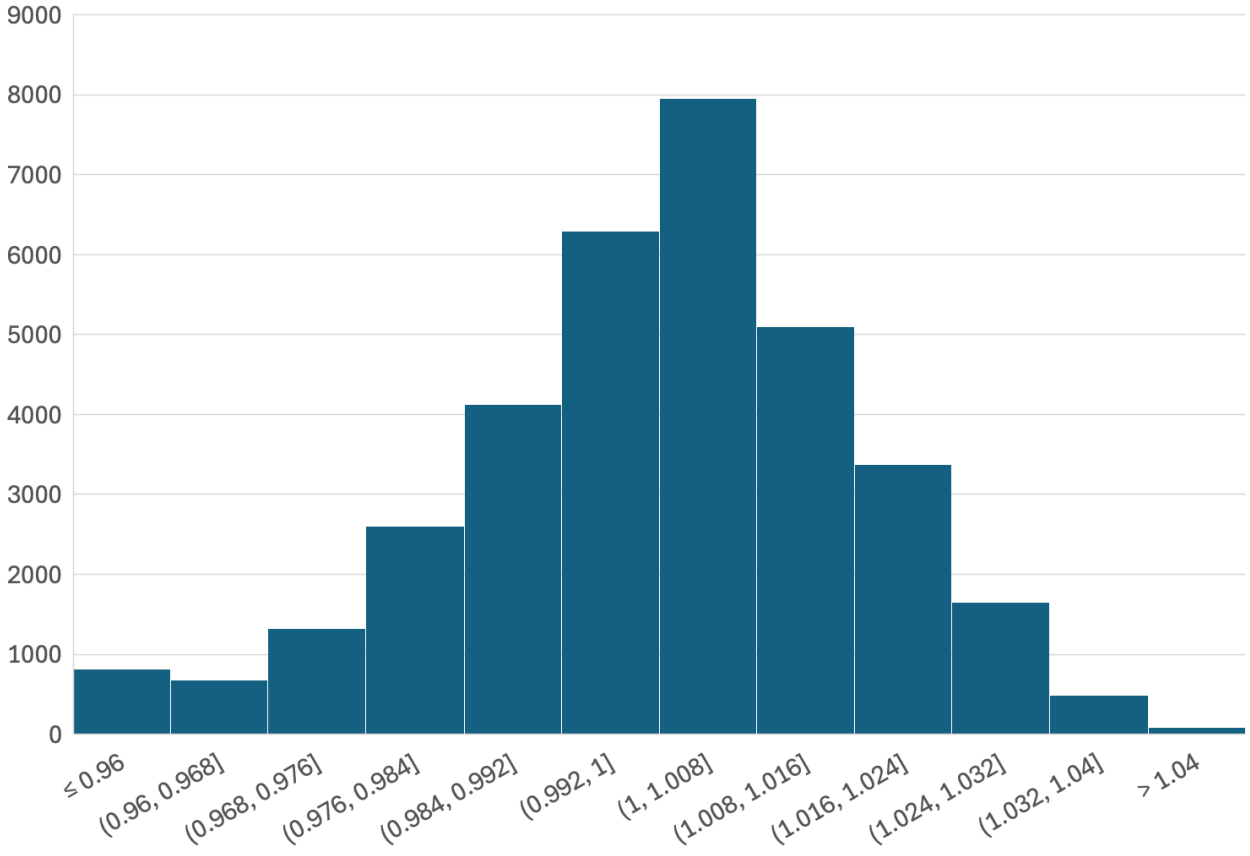

Figure 3*. Histogram of UHI values (34,502 total) representing the target dataset for the challenge. Most of the data is close to the mean temperature (UHI=1.0) but there is variability suggesting cooler regions (UHI=0.913, minimum) and hotter regions (UHI=1.054, maximum) within the bounds of the data collection region. Credit: Brian Killough, EY.*

The range of collected temperatures had a maximum difference of 12.8 Degrees-Fahrenheit (7.1 Degrees-Celsius) and mean of 90.7 Deg-F. Though this is lower than known global extremes (>10 Degrees-Celsius difference), the collected data does allow the identification of urban heat islands in the region. When converted to UHI index values (range of 0.913 to 1.054), this yielded a 14% variation in UHI values across the data collection region.

Satellite Data

The launch of the European Copernicus Sentinel-2 missions in 2015 and 2017 provides optical data at 10-meter spatial resolution and a revisit every 10 days with one mission and every 5 days with two missions. This free and open data is readily available from the Microsoft Planetary Computer (https://planetarycomputer.microsoft.com/catalog). But optical data cannot penetrate clouds, so it is necessary to filter out clouds or select scenes that have very low levels of cloud cover. For this challenge, we have provided a sample Sentinel-2 Python notebook that selects low-cloud scenes or allows the creation of a median mosaic without cloud contamination. This product can be used to assess the impacts of vegetation extent, water, or urban density on urban heating. It is well known [Reference 4] that proximity to vegetation (green space), proximity to water, and local urban density contribute to the effects of urban heating. Below is an example Sentinel-2 mosaic product showing the spatial variation of the Normalized Difference Vegetation Index (NDVI) over our data challenge region.

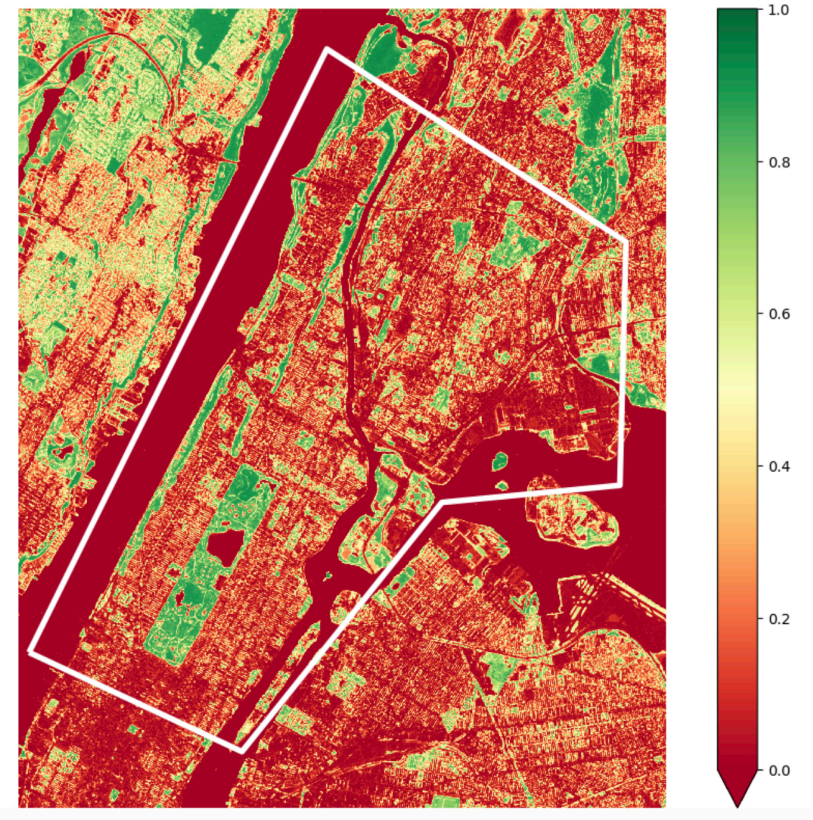

Figure 4. *Sentinel-2 Normalized Difference Vegetation Index (NDVI) over the data challenge region (Montgomery County, Maryland). This product is based on a single date near the data collection date (August 7). Areas of light green or dark green are consistent with the presence of vegetation. Areas of dark red are consistent with dense urban environments or water. This information can be used in your digital model as vegetation, urban density and water can impact local urban heating. Credit: Brian Killough, EY (using data from ESA's Sentinel-2 mission on the Microsoft Planetary Computer).*

References

[1] Poumadère M, Mays C, Le Mer S, Blong R. The 2003 Heat Wave in France: Dangerous Climate Change Here and Now. *Risk Analysis*, Vol. 25, Issue 6, Dec 2005, pp. 1483-1494. https://doi.org/10.1111/j.1539-6924.2005.00694.x

[2] Borden, K.A., Cutter, S.L. Spatial patterns of natural hazards mortality in the United States. *International Journal of Health Geographics*, 7, Article 64 (2008). https://doi.org/10.1186/1476-072X-7-64

[3] Lee S, Kim D. Multidisciplinary Understanding of the Urban Heating Problem and Mitigation: A Conceptual Framework for Urban Planning. Int J Environ Res Public Health. 2022 Aug 18;19(16):10249. https://doi.org/10.3390/ijerph191610249

[4] Shandas, V., Voelkel, J., Williams, J., & Hoffman, J., (2019). Integrating Satellite and Ground Measurements for Predicting Locations of Extreme Urban Heat. Climate, 7(1), 5. https://doi.org/10.3390/cli7010005

[5] Voelkel, J., & Shandas, V. (2017). Towards Systematic Prediction of Urban Heat Islands: Grounding Measurements, Assessing Modeling Techniques. Climate, 5(2), 41. https://doi.org/10.3390/cli5020041