# Urban Heat Island Prediction

## 1) INTRODUCTION

The Urban Heat Island (UHI) effect is an issue resulting 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. [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. [3]

To measure this effect, we use an **UHI Index** = (Temperature at a given location) / (Mean temperature for all locations).

The goal of this project is to **develop a digital model to predict the locations and severity of the UHI effect and to understand the drivers of this phenomenon, in New York City** (Manhattan and The Bronx).

## 2) DATA

For this project, the spatial coordinate system used is epsg:4326.

**Response variable**

Data was collected by CAPA Strategies using a ground traverse with vehicles and bicycles on 07/24/2021 between 3pm and 4pm. This data collection effort resulted in 11,229 data points having a unique UHI index. 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 at the time of day consistent with the data collection. 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 for consideration of natural variability.

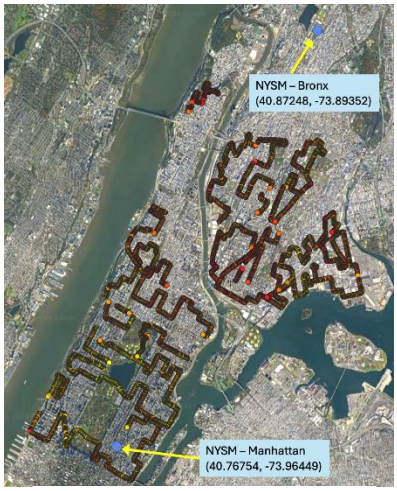



*Figure 1: Visualization of UHI Index values across NYC on 07/24/2021*

**Predictor variables**

- EuropeanSentinel-2 optical satellite data (Multispectral)

Extraction of specific band values of Sentinel-2 satellite imaging data with their precision:

| B01 | B02 | B03 | B04 | B06 | B08 | B11 |
|---|---|---|---|---|---|---|
| Coastal Aerosol | Blue | Green | Red | Red Edge (740 nm) | NIR (833 nm) | SWIR (1.6 um) |
| 60m | 10m | 10m | 10m | 20m | 10m | 20m |

Values are extracted via API Calls with the planetary_computer and then stored as GeoTIFF images.

The data were averaged over a **three-month** period, using only images with less than 10% cloud cover. This avoids cloud bias and gives a more accurate representation of the observed surface.

For each of them, I created averaged buffered value of a circle of radius 50 m, 100 m, 150 m, 200 m, 250 m, 300 m, 350 m, 400 m, 450 m, 500 m, 600 m, 700 m, 800 m, 900 m and 1000 m.

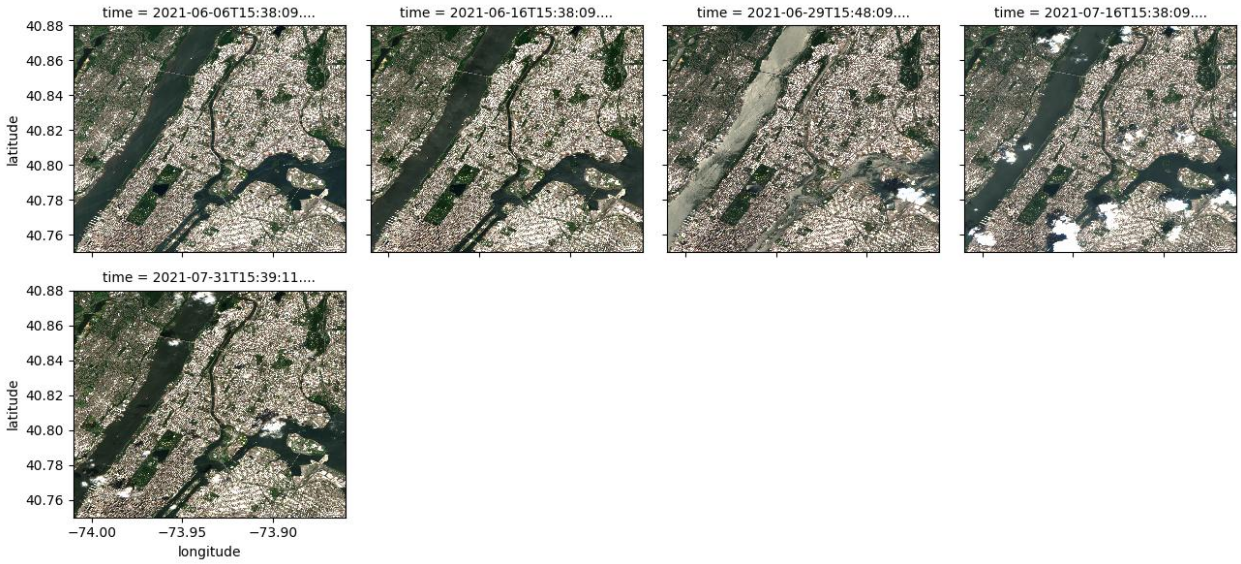



*Figure 2: RGB view of Manhattan and the Bronx at different time frame*

- NASA Landsat Optical Satellite Data (Surface Temperature)

I used the same method as the previous satellite (except special buffered part) to extract the Surface Temperature data from Lansdat satellite: Band 11 = Surface Temperature (lwir11) with a precision of 100m.

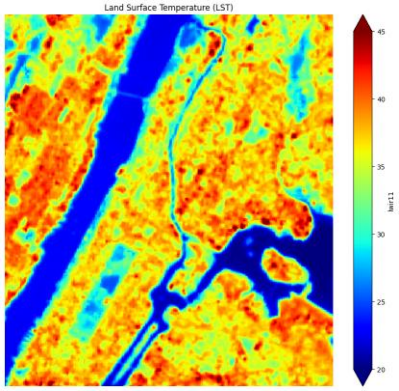



*Figure 3: Surface temperature of Manhattan and the Bronx measured by Landsat Satellite*

- The Building Footprint Data

These data were extracted from NYC Open Data : [https://data.cityofnewyork.us/City-Government/Building-Footprints-Map-/3g6p-4u5s](https://data.cityofnewyork.us/City-Government/Building-Footprints-Map-/3g6p-4u5s)

This CSV file contains the coordinates of all the buildings in Manhattan and the Bronx, associated with their elevation

## 3) METHODOLOGY

In order to model the UHI effect in Manhattan and the Bronx, I used ideas and techniques from these two articles:

- *Voelkel, J., & Shandas, V. (2017). Towards Systematic Prediction of Urban Heat Islands: Grounding Measurements, Assessing Modeling Techniques [4]*
- *Shandas, V., Voelkel, J., Williams, J., & Hoffman, J., (2019). Integrating Satellite and Ground Measurements for Predicting Locations of Extreme Urban Heat [5]*

These articles study a similar use case as they are trying to predict UHI Index in Portland, Washington, Richmond and Baltimore.

In the article [4], the authors tried three different models: a Multiple Linear Regression, a Classification and Regression Tree and a Random Forest. The predictive results were a lot better using the **Random Forest**, and as there was a lot of things to explore of the data preprocessing, I decided to focus on this model.

- Features

To detect key elements of the city like vegetation, buildings or water, it's interesting to go further the bands data and add **spectral index products** using mathematical combinations of bands. For example, in this model, I used the:

- Normalized Difference Vegetation Index NDVI = (NIR-Red) / (NIR+Red)
- Normalized Difference Buildup Index NDBI = (SWIR-NIR) / (SWIR+NIR)
- Normalized Difference Water Index NDWI = (GREEN-NIR) / (GREEN+NIR)

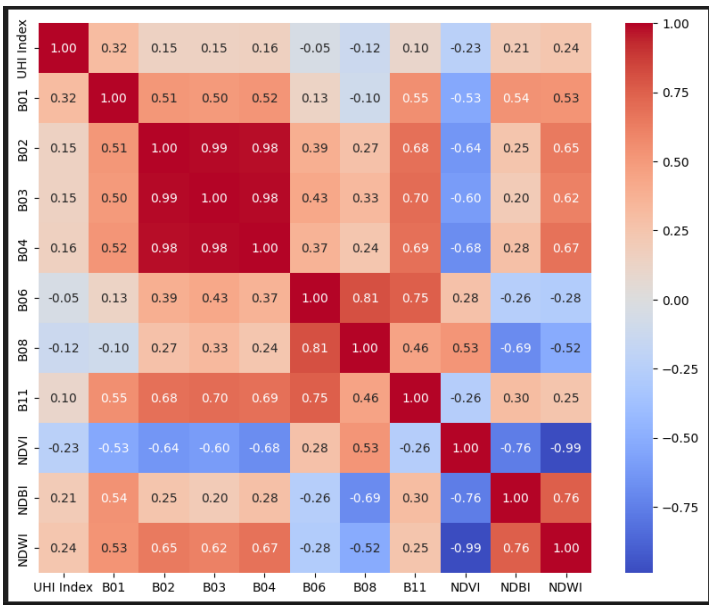



*Figure 4: Correlation Matrix between the UHI Index, the bands and the spectral indices*

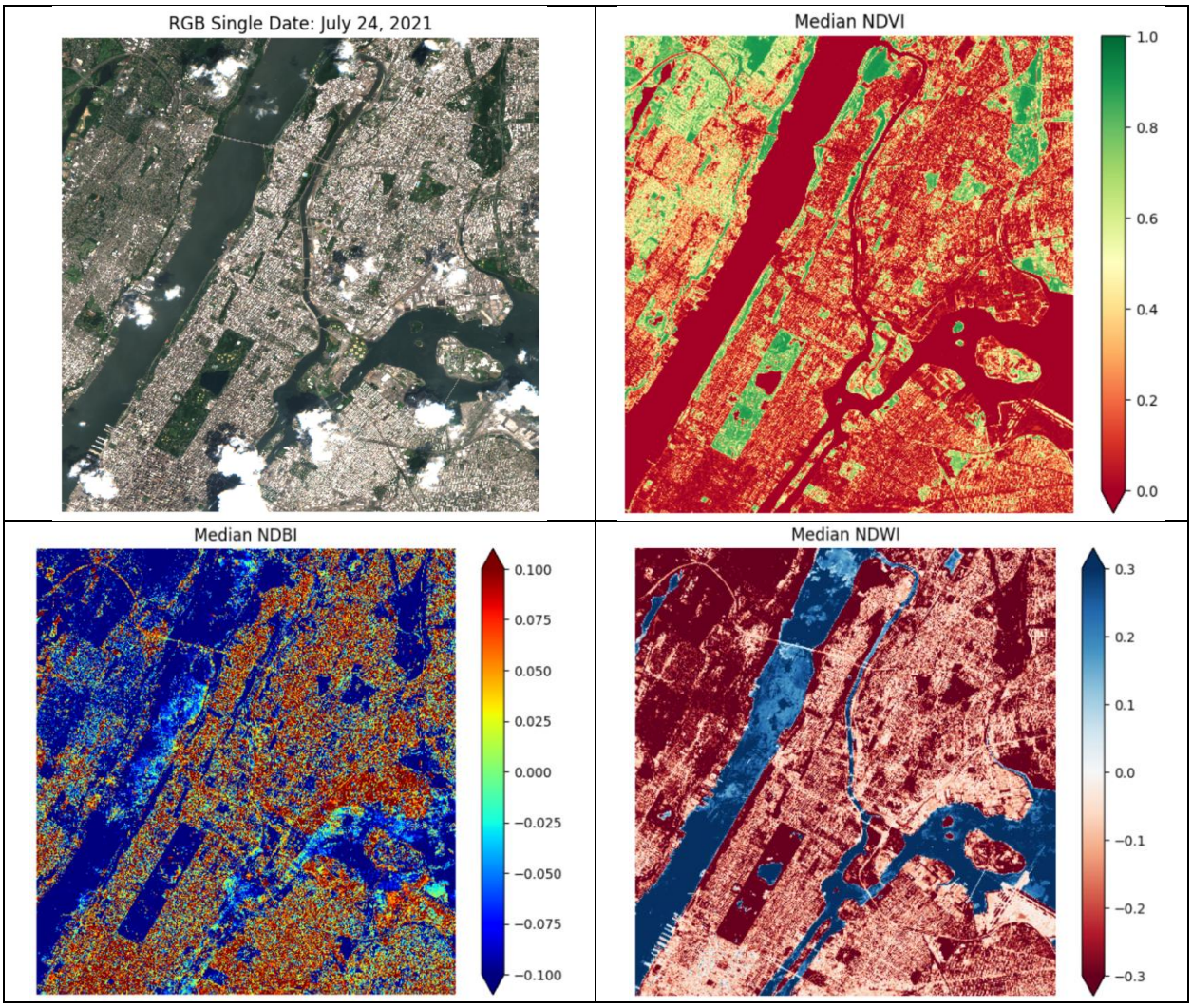



*Figure 5: Visualization of the RGB and the spectral indices (NDVI, NDBI, NDWI)*

As I explained above, for each value of the bands, I created a focal buffer. It represents the average of the value in a circle of radius 50 m, 100 m, 150 m, 200 m, 250 m, 300 m, 350 m, 400 m, 450 m, 500 m, 600 m, 700 m, 800 m, 900 m and 1000 m.

I also did the same for the different Index, and this brings **a multi distance - multi spectral analysis**.

For the Landsat data, I extracted the surface temperature.

For the building data, I chose to use the sum of areas and the sum of volumes (more like an average height) of the building using the same distance buffer as the bands. For example:

- Sum_area_50 = Sum of all buildings 'area in a circle of 50m / Total area
- Sum_volume_50 = Sum of all (buildings 'area * buildings 'height) in a circle of 50m / Total area

Then, I joined my features and my predictor variable, and I removed the duplicates.

- Model Building

To build the model, I used the Scikit learn library.

I split my data into a 70% training and 30% test.

For my feature scaling, I used a **Standard Scaler**. I tried to use different ones, but I didn't have better results.

To avoid overfitting, I used **cross-validation** method.

The metric used was the **R-Squared** one.

As I explained before, the model chosen was a **Random Forest Regressor**. It's a nonparametric machine learning technique that predicts continuous values by averaging the results of multiple decision trees, each trained on a random subset of the data.

Below is the evolution and impacts of the improvements I brought to the model.

For all improvements, the model is a Random Forest Regressor with n=100 estimators and I used a Standard Scaler. As we can see in the evolution of the scores, the work on the data preparation is really impactful here.

| | **Comments** | **Train** | **Test** | **Submission** |
|---|---|---|---|---|
| 1 | First model: Use of features B01, B04 and NDVI on a single day<br>-> We can see that the model overfits | 0.9113 | 0.3706 | 0.5108 |
| 2 | Adding median data: instead of values of a single day, we use the median values of multiple days | 0.9160 | 0.3703 | 0.519 |
| 3 | Adding the NDBI index | 0.9103 | 0.3259 | 0.4929 |
| 4 | Adding the Buffer of 50 meters radius with only B01, B04 and NDVI<br>-> We can see that adding the buffer is a good improvement for the model | 0.9402 | 0.5735 | 0.557 |
| 5 | Adding LST data (surface temperature) with all the previous data of step 4<br>-> Also here, the surface temperatures are helping a lot | 0.9498 | 0.6441 | 0.6232 |
| 6 | Adding all the buffers of all the bands to the previous data<br>-> Main improvement, the notion of buffer here, especially on larger radius is improving the model | 0.9825 | 0.8901 | 0.881 |
| 7 | Adding the building data to the previous data | 0.9916 | 0.9404 | 0.9394 |
| 8 | Changing parameter of the model n=200 estimators | 0.9920 | 0.9413 | |
| 9 | Using a test size of 10% | 0.9939 | 0.9550 | 0.9544 |
| 10 | Final Submission with all the training set | 0.9944 | | **0.9596** |

## 4) RESULTS

In this use case, I reached a prediction score of 95,96% which is pretty good value. In fact, in the articles, they reached scores between 96.44% and 98.03% for the cities they were modelling, and they explained that UHI in the afternoon are the harder to predict because of the non-land use variable.

As I explained, I spent most of my time understanding the data and preparing them, rather than focusing on the model part.

For the data part, the main improvements were the add of the buffers, the surface temperature and the building data.

Now, it's interesting to look at the features importance of the Random Forest:

| | **Features** | **Importance score** |
|---|---|---|
| 1 | sum_volume_1000 | 0.23897 |
| 2 | sum_area_900 | 0.055775 |
| 3 | B11_buffer_1000 | 0.04077 |
| 4 | sum_volume_700 | 0.035691 |
| 5 | sum_volume_250 | 0.031521 |
| 6 | sum_area_1000 | 0.023703 |
| 7 | B03_buffer_1000 | 0.022755 |
| 8 | sum_area_800 | 0.019038 |
| 9 | NDBI_buffer_1000 | 0.016833 |
| 10 | sum_volume_450 | 0.016317 |
| 11 | sum_volume_900 | 0.016194 |
| 12 | sum_area_700 | 0.016119 |
| 13 | LST | 0.014892 |
| 14 | sum_volume_800 | 0.014225 |
| 15 | B06_buffer_1000 | 0.013217 |

First, we can see that the feature sum_volune_1000 dominates by far. In a global way, we can see that the features associated to the buildings are the most important here. The B11 band presence is interesting because it's an infrared band that it used a lot for environmental monitoring like vegetation one. SWIR bands are sensitive to cellulose allowing the detection of vegetation and assessing their health. [6]

Secondly, we can see that the importance score quickly drops, meaning that the model can also be simplified with less features, to be faster for training and prediction.

What are the limits of my model and what I could have explored more:

- Adding non-land use variables: humidity, wind speed and direction, albedo, urban canyons, … It would also have been interesting to add a notion of solar flux mixed with the shadow of building.
- Differentiate canopied and non-canopied vegetation by analyzing their height to differentiate their nature and see if one has more impact than another (tress vs. grass impact)
- Exploring other models
- Landsat data also have limitations: it only has a precision of 100m which often mixes the type of surface (building, road, land) and the acquisition time is about 11:30 am which does not exactly match the time of the day when ground traverse data was collected (3:00 pm to 4:00 pm).
- Simplification and selection of features

## 5) CONCLUSION

Emerging tools, like satellite and climate data are now essential tools to analyze, monitor and assess our human impact over large areas. They are helping to evaluate the risks and develop appropriate strategies and solutions against hazards and vulnerabilities.

In this used case, we saw that lower-density urban areas are often cooler than high-density, and that vegetated/watered areas are cooler than urbanized areas. Also, more than identifying the causes, the model can predict the localization of the UHI and their scale, which can be useful for the city in case of heatwave.

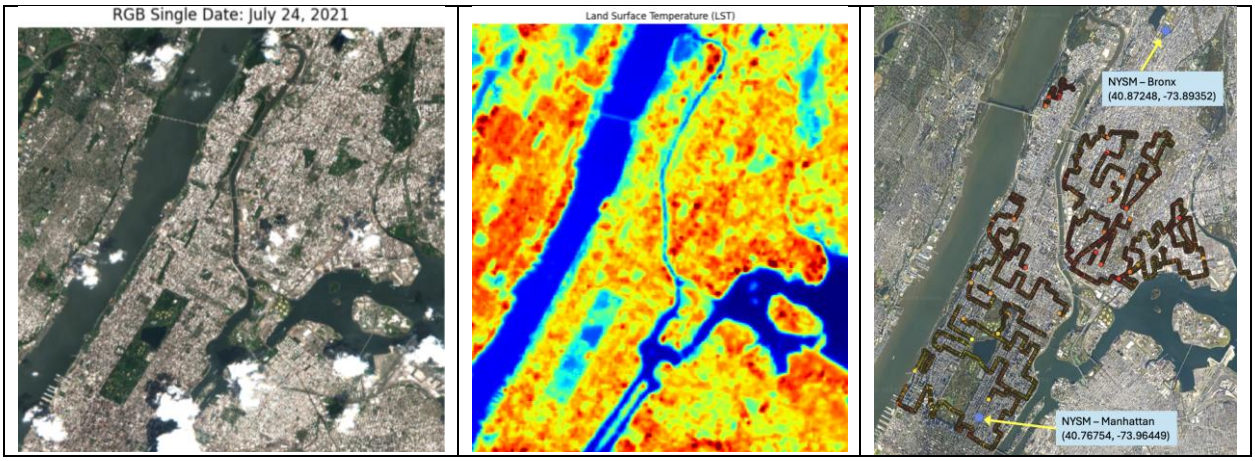



*Figure 6: Comparison of RGB, Landsat Surface Temperature and UHI Index Views*

This work was done under the 2025 EY Open Science AI&Data Challenge Program, with over 10,000 participants from 115 countries.

With a final score of 95.96%, I was ranked 93/380 among external participants.

## 6) REFERENCES


[1] PoumadèreM, Mays C, Le Mer S, BlongR. 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] 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

[5] 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

[6] Hively, W.D.; Lamb, B.T.; Daughtry, C.S.T.; Serbin, G.; Dennison, P.; Kokaly, R.F.; Wu, Z.; Masek, J.G. Evaluation of SWIR Crop Residue Bands for the Landsat Next Mission. Remote Sens. 2021, 13, 3718. https://doi.org/10.3390/rs13183718

**ANNEXE**

## Overview of Sentinel-2 Mission

Sentinel-2 is a European wide-swath, high-resolution, multi-spectral imaging mission. The full mission specification of the twin satellites flying in the same orbit but phased at 180°, is designed to give a high revisit frequency of 5 days at the Equator.

Each of the satellites in the Sentinel-2 mission carries a single payload: **the optical Multi-Spectral Instrument (MSI) that samples 13 spectral bands: four bands at 10 m, six bands at 20 m and three bands at 60 m spatial resolution.** The orbital swath width is 290 km.

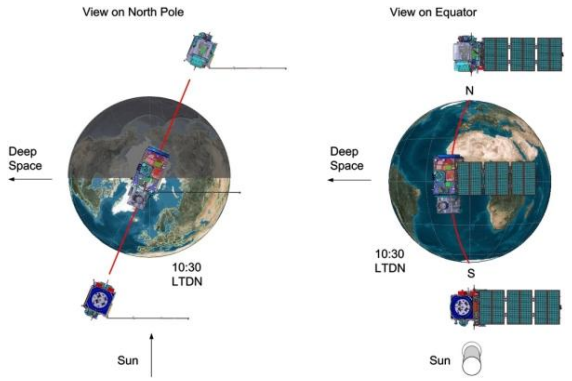



Figure 1: The Twin-Satellite Sentinel-2 Orbital Configuration [Credits: Astrium GmbH]

The Sentinel-2 twin satellites carry on the legacy of SPOT and LANDSAT by continuing to provide similar types of image data and contributing to ongoing multispectral observations. These satellites are used to support a variety of services and applications offered by Copernicus, including land management, agriculture, forestry, disaster control, humanitarian relief operations, risk mapping, and security concerns.

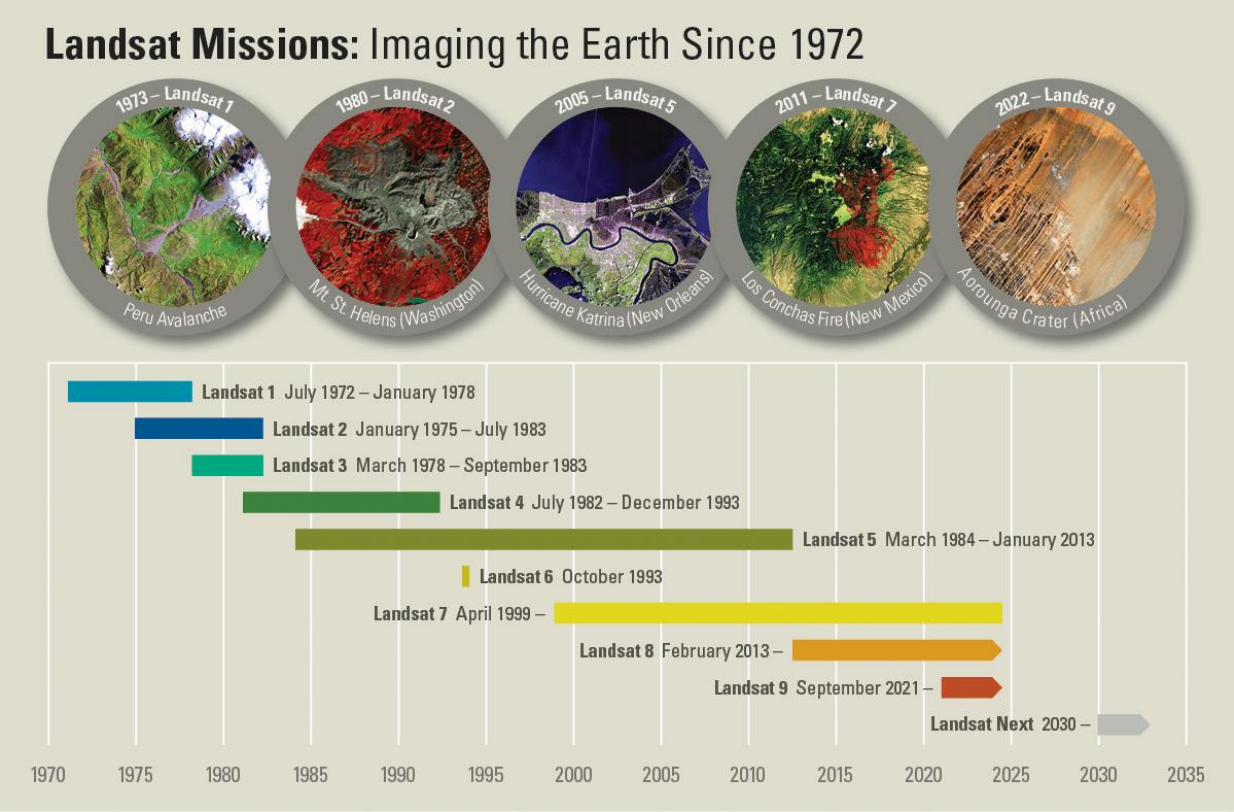
The Sentinel-2 mission consists of two identical satellites operating together, Sentinel-2B (launched in 2017) and Sentinel-2C (launched in 2024) that were launched using the European VEGA launcher. Each of these satellites weighs approximately 1.2 tonnes.

*Table 1: Summary of useful orbital information for Sentinel-2 satellites:*

| Altitude | Inclination | Period | Cycle | Ground-track deviation | Local Time at Descending Node |
|---|---|---|---|---|---|
| 786 km | 98.62 deg | 100.6 min | 10 days | +- 2 km | 10:30 hours |

[Landsat Collection 2 | U.S. Geological Survey]

The Landsat Missions are comprised of Earth-observing operational satellites that carry remote sensors to collect data and image our planet as a part of the U.S. Geological Survey (USGS) National Land Imaging (NLI) Program. Products generated from the imagery acquired by the sensors carried on the Landsat satellites are hosted at the USGS Earth Resources Observation and Science (EROS) Center in Sioux Falls, South Dakota.




Landsat Collection 2, the second major reprocessing effort on the Landsat archive, resulted in several data product improvements that applied advancements in data processing, algorithm development, and data access and distribution capabilities.

Landsat Collection 2 contains Level-1 data from Landsats 1-9, and Level-2 and Level-3 science products from Landsats 4-9.

**Landsat 8 Satellite Orbit Facts**

Orbits the Earth in a sun-synchronous, near-polar orbit (98.2 degrees inclination)

Achieved an altitude of 705 km (438 mi), Completes one Earth orbit every 99 minutes

Has a 16-day repeat cycle with an equatorial crossing time of 10:00 a.m. +/- 15 minutes

Landsat 8 carries two sensors. The Operational Land Imager sensor is built by Ball Aerospace & Technologies Corporation. The Thermal Infrared Sensor is built by NASA Goddard Space Flight Center.