Remote Sensing and Machine Learning for Urban Air Quality and Heat Island Monitoring

Maria Antonia Brovelli[1], Jesus Rodrigo Cedeno Jimenez[1], Afshin Moazzam[1], Vasil Yordanov[1], Alberto Vavassori[1]

[1] Politecnico di Milano, Department of Civil and Environmental Engineering, Piazza Leonardo da Vinci, 32, Milan, Italy - (maria.brovelli, jesusrodrigo.cedeno, afshin.moazzam, vasil.yordanov, alberto.vavassori)@polimi.it

Keywords: Remote Sensing, Machine Learning, Urban Air Quality, Local Climate Zone, Urban Heat Island

Abstract

This study presents a dual-strategy approach to monitor urban environmental stressors, conducted within the ASI-MUR-funded Space It Up! project, focusing on atmospheric pollution and the urban heat island (UHI) effect. First, we developed a scalable machine learning (ML) framework for estimating ground-level concentrations of NO_2, SO_2, and CO in Milan using Sentinel-5P satellite data, ERA5 reanalysis, CAMS forecasts, and ARPA Lombardia ground measurements. Data preprocessing pipelines were optimized by switching to Google Earth Engine, reducing retrieval times and enabling operational scalability. Despite known satellite retrieval limitations in winter months for SO_2, model performance remained robust, with normalized RMSE values consistently below 0.85. For CO, a Deep Attention Network achieved the best results (NRMSE = 0.4879), demonstrating the adaptability of the framework across pollutants. Additionally, a comparative analysis of low-cost air quality sensors showed high performance from AirGradient devices, particularly for $PM_{2.5}$ and temperature, though significant inter-brand discrepancies were observed for CO_2. Second, we implemented an advanced LCZ classification method integrating hyperspectral PRISMA imagery, Sentinel-2 data, and urban canopy parameters (UCPs). Applied to the Metropolitan City of Milan, the proposed workflow achieved substantial improvements over existing methods, with an overall accuracy increase up to 16% when utilizing PRISMA data compared to the state-of-art LCZ Generator approach. We also presented ongoing efforts to further improve the proposed methodology, including the automation of data retrieval and training and test sample creation. The methodology is being applied across multiple urban areas worldwide by also testing other ML techniques. Together, these methodologies provide a comprehensive and reproducible framework for urban environmental monitoring.

1. Introduction

Urban environments are increasingly challenged by environmental stressors such as air pollution and the urban heat island (UHI) effect, which threaten public health, urban livability, and climate resilience. The rapid growth of cities, coupled with climate change, exacerbates these phenomena, particularly in dense metropolitan areas where anthropogenic emissions and impervious surfaces are concentrated. Monitoring and mitigating these effects require spatially explicit and temporally consistent data.

Monitoring trace atmospheric pollutants such as carbon monoxide (CO), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2) allows scientists and policy makers to assess both environmental quality and public health risks. These gases, primarily emitted through fossil fuel combustion, industrial processes, and biomass burning, play a central role in urban air pollution, photochemical smog formation, and climate forcing (Manisalidis et al., 2020). Due to their significant impacts on respiratory and cardiovascular health (WHO, 2024), regulatory bodies have developed air quality standards to limit ambient concentrations. However, dense monitoring networks are generally limited to high-income regions, while large parts of the globe, particularly in Africa, Latin America, and parts of Asia, remain underinstrumented (Smith et al., 2025). This disparity impacts negatively global-scale air quality assessment and the development of equitable pollution mitigation strategies.

On the other hand, understanding and mitigating UHI is essential for sustainable urban development and improving the quality of life for citizens and ecosystem well-being (Irfeey et al., 2023). However, traditional approaches that simply compare urban and rural areas often fail to capture the complex spatial variability and morphological influences on UHI intensity (Liu et al., 2023). The Local Climate Zone (LCZ) classification system offers a robust and standardized framework to analyze urban climate by categorizing landscapes based on surface structure, land cover, and human activity, thus enabling precise mapping of UHI patterns and their drivers (Zhou et al., 2022).

This paper, conducted within the framework of the Space It Up! (SIU) project funded by the Italian Space Agency (ASI) and the Italian Ministry of University and Research (MUR), presents two complementary lines of research: (1) the estimation of ground-level concentrations of air pollutants using machine learning (ML) techniques, satellite data (Sentinel-5P), and ERA5 reanalysis data, and (2) the classification of LCZs to characterize urban morphology and its relationship with urban thermal patterns. The proposed work uses recent advances in Earth Observation (EO) technologies and data-driven methodologies to produce results on both fronts, which can be further explored in the future to investigate potential interactions between these two phenomena.

The remainder of the paper is structured as follows. Section 2 describes the satellite data and ML methodology for pollutant estimation, including the evaluation of low-cost sensors. Section 3 details the LCZ classification approach and the ongoing improvements in the methodology. Section 4 presents the conclusions from both research lines.

2. Satellite data and air quality monitoring

Earth observation (EO) platforms have become essential tools for addressing global air quality data gaps by delivering con-

sistent, large-scale atmospheric measurements. The Sentinel-5 Precursor (Sentinel-5P) mission, launched by the European Space Agency (ESA) in 2017, allowed scientists to monitor atmospheric pollutant concentrations with higher resolution than before. Its onboard TROPOspheric Monitoring Instrument (TROPOMI) detects NO_2, SO_2, CO, and other trace gases at high spatial resolution (5.5×3.5 km^2) with daily revisit capability (Veefkind et al., 2012). Compared to earlier missions such as OMI or SCIAMACHY, TROPOMI offers enhanced spatial detail, enabling improved monitoring of pollution patterns in urban and industrial areas (Gu et al., 2025).

Sentinel-5P has been used in many applications, including tracking COVID-19 lockdown-related NO_2 reductions in Europe, South America, and India (Levelt et al., 2021), identifying SO_2 emission hotspots (Fioletov et al., 2016), and detecting CO from biomass burning in tropical forests (Landgraf et al., 2016). However, since TROPOMI measures total column densities rather than surface concentrations, translating these to ground-level values remains challenging. Retrieval limitations and atmospheric complexity can result in discrepancies with in situ data, particularly in regions with clouds or low emissions (Griffin et al., 2019, Van Geffen et al., 2020). To address this, integration with meteorological reanalysis and ML is required, an approach explored in the following sections.

2.1 Machine Learning Approaches for Surface-Level Estimation

As specified before, although satellite instruments such as Sentinel-5P offer high spatial coverage of trace gas distributions, they primarily retrieve total or tropospheric vertical column densities. These must be transformed to estimate near-surface concentrations, which are more relevant for health assessments and regulatory policies. Given the complexity of this translation, ML methods offer a flexible, data-driven alternative for mapping satellite-based columns to ground-level pollutant concentrations (Zhou et al., 2024). As shown in previous work (Cedeno Jimenez and Brovelli, 2023), ML models can combine satellite retrievals, meteorological reanalysis, and in-situ measurements to capture nonlinear relationships between predictors and surface levels. Algorithms such as Random Forests (RF), Support Vector Regression (SVR), and Gradient Boosting (GB) have been used for this task, while deep learning approaches continue to grow for modeling spatiotemporal dependencies (Li et al., 2017, Cai et al., 2025). A study in China found that an XGBoost model using TROPOMI NO_2, meteorological inputs, population density, and road networks reached an R^2 of 0.83, with an RMSE of 7.58 g/m^3 and a mean error of 5.56 g/m^3 (Liu, 2021).

Cedeño Jimenez and Brovelli (2024) demonstrated the feasibility of estimating ground-level NO_2 using only remote sensing inputs, including Sentinel-5P and ERA5 meteorological variables (`https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels`). The model, validated over the Metropolitan City of Milan (MCM) and transferred to Mexico City, incorporated boundary layer height (ABLH), surface temperature, and wind dynamics to produce robust estimates. Comparative tests against TimeGPT (`https://www.nixtla.io/docs`) confirmed that traditional ML models, when well-tuned, outperform generative time series methods in this application.

2.1.1 Data Access Optimization. Given the success of this framework in estimating ground-level NO_2, the first phase of this study focused on improving the existing Python-based geospatial pipeline for data collection and pre-processing, with the goal of enhancing its scalability to other regions. Although previously tested in Mexico City (Cedeno Jimenez and Brovelli, 2023), challenges remained in the data acquisition process. The Copernicus Browser's API (`https://dataspace.copernicus.eu/analyse/apis`) was initially used to download Sentinel-5P data, but this method posed limitations in terms of time, storage, and processing requirements. Moreover, user quotas (`https://documentation.dataspace.copernicus.eu/Quotas.html`) restricted request frequency, concurrent sessions, and parallel downloads. Additionally, Copernicus only provides Level-2 data, which demands further processing such as pixel quality filtering and binning.

To address this, the study compared the Copernicus Browser with Google Earth Engine (GEE, `https://earthengine.google.com`), a cloud-based geospatial platform offering Level-3 Sentinel-5P data binned at 1 km × 1 km. DIAS platforms like CREO DIAS and WekEO were excluded as they provide only Level-2 data and lack bulk download options. A comparative analysis was performed between Copernicus and GEE data: Copernicus data was processed from Level-2 to Level-3 using a 75% quality assurance threshold, while GEE data was upscaled to 5.5 km × 3.5 km to match the original resolution. After temporal and spatial alignment by pixel ID and date, a merged dataset enabled the computation of relative errors. Results showed an average difference of 5% and a Pearson correlation of 98.30%. The discrepancies were mainly due to GEE's unpublished processing algorithm and a 1.20 km grid shift.

To evaluate the impact on model performance, we applied the ML model trained on Copernicus data using GEE as input. The resulting average NRMSE was 58%, compared to 56% with Copernicus, indicating only a minor degradation. This confirmed that GEE remains a valid and efficient alternative for operational NO_2 estimation, reducing resource demands while maintaining reliable performance.

2.1.2 Framework extension to additional trace gases. By improving the processing framework to use a data source other than the Copernicus Browser, the time required to access satellite data was drastically reduced. This enhancement enabled the Python pipeline to be used for trace gases beyond NO_2, eliminating the need for the previously cumbersome data download and processing procedure. We expanded our analysis to include CO and SO_2 alongside NO_2 due to their environmental and health relevance. CO, resulting from incomplete combustion, has a longer atmospheric lifetime and more uniform spatial distribution compared to NO_2, and has been linked to respiratory mortality even at ambient levels below regulatory limits (Allred et al., 1989). Additionally, SO_2 is typically emitted in brief, concentrated plumes from stationary sources such as power plants and volcanic activity, and short-term exposure has been associated with neural and respiratory effects (Meo et al., 2024). Including both pollutants allowed the framework to address trace gases with varying behaviors and impacts on urban populations, increasing its scalability and applicability for real-world air quality monitoring.

The ML pipeline was first adapted for the estimation of ground-level sulfur dioxide (SO_2). This followed the same scalable data acquisition and pre-processing structure previously implemented for NO_2, incorporating daily SO_2 total column retrievals from Sentinel-5P via GEE, meteorological variables from ERA5, Copernicus Atmosphere Monitoring Service (CAMS)

reanalysis SO_2 estimations, and in-situ measurements from Regional Environmental Protection Agency (ARPA) Lombardia. All variables were harmonized temporally with the satellite overpass window. The model incorporated derived features such as boundary layer height, surface pressure, and solar radiation to capture temporal variability and vertical atmospheric structure. These additions helped correct for the limited vertical sensitivity of satellite-derived SO_2 measurements. Among all the wide variety of ML models that were trained and tested, the best-performing configuration was an ensemble of Random Forest and Gradient Tree Boosting combined using a voting mechanism.

A key challenge encountered in modeling SO_2 was the seasonal absence of valid satellite retrievals year by year during the months of December and January for the MCM. Figure 1, shows there an example of the data gap in the SO_2 measurements retrieved from GEE in Europe for the months of December 2024 and January 2025. This data gap is primarily caused by increased noise and reduced measurement quality at high solar zenith angles and swath edges, leading to frequent quality flag rejections in the standard SO_2 product (Fioletov et al., 2020). Retrieval noise increases substantially at zenith angles greater than 75°, producing consistent wintertime gaps across years (Fioletov et al., 2020). Despite this issue, the SO_2 model demonstrated robust performance, achieving NRMSE values consistently below 0.85 when validated against ARPA ground stations. Although the future inclusion of predictors such as NO_2 may further enhance performance, as explored in other studies (Yang et al., 2022), the current model's accuracy falls within the standard deviation of ground-based measurements, supporting its use under partial data availability.

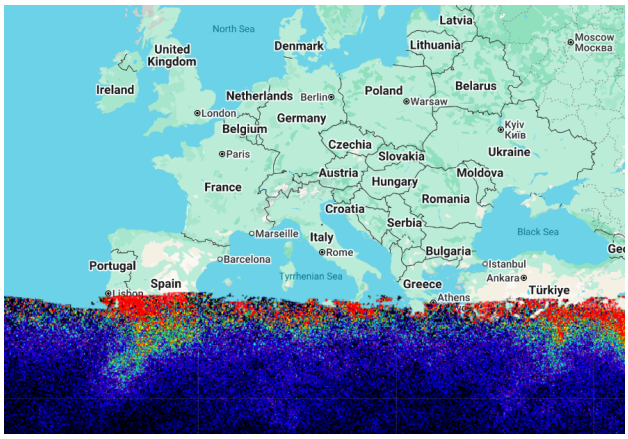

Figure 1. SO_2 Sentinel-5P image gaps during the winter period from Google Earth Engine from December 1st 2024 till January 15th 2025.

In parallel, the framework was also adapted to estimate ground-level carbon monoxide (CO) concentrations. Following the same structure, we integrated Sentinel-5P GEE data, ERA5 meteorological indicators, CAMS reanalysis products, and ARPA Lombardia in-situ measurements. These datasets were harmonized to the same spatial grid and temporally aligned with the Sentinel-5P overpass window. Additional temporal and derived features, such as normalized CO and solar-thermal contrast, were computed to enrich the dataset and account for atmospheric memory effects. This configuration ensured that the model could account for both short-term and persistent spatial patterns typical of CO distribution.

The modeling process was conducted using both random and chronological validation splits. Among untuned models, the Random Forest initially provided the best performance, with a test NRMSE of 0.6363. Performance improved with the inclusion of CAMS CO reanalysis and meteorological variables like boundary layer height and surface radiation. An ensemble combining Support Vector Regression (SVR) and Gradient Boosting (GB) reduced the NRMSE to 0.5819. SHAP-based feature importance analysis confirmed the high predictive value of CAMS CO, boundary layer height, and solar-thermal indicators. After eliminating low-contribution features, the Deep Attention Network (DAN) achieved the best overall performance, with an NRMSE of 0.4879 validated through 20 shuffle-split iterations. While small improvements were seen after removing extreme values, these were retained to preserve the model's ability to detect peak pollution events. These results confirm the reliability and adaptability of the framework for pollutants with different emission profiles, reinforcing its value for large-scale and scalable air quality monitoring.

2.2 Low-cost ground sensors

Traditional air quality monitoring systems rely on static stations equipped with certified reference instruments, which provide highly accurate data. However, these stations are expensive to install and maintain, resulting in sparse geographic coverage that often fails to capture localized pollution gradients - especially in smaller cities or underdeveloped regions where health risks may be underestimated (Brauer et al., 2016). To address these limitations, there has been a rapid rise in the use of low-cost, compact, and user-friendly air pollution sensors (Jiao et al., 2016, Castell et al., 2017). These new platforms enable more widespread and frequent monitoring, offering high spatial and temporal resolution data in near-real-time. Such capabilities can supplement existing networks, support the creation of detailed air quality maps, facilitate personal exposure assessments, and encourage greater public participation in environmental monitoring. Despite these advantages, low-cost sensors face significant challenges regarding data quality and reliability. Their performance can vary widely between units and under different environmental conditions. Issues such as chemical interference, cross-sensitivity to other gases, and sensitivity to meteorological factors like temperature and humidity can affect sensor accuracy. As a result, field calibration against reference-grade instruments is essential to ensure data validity (Castell et al., 2017). While low-cost sensors may not yet meet the strict data quality requirements for regulatory compliance or precise scientific exposure assessments, they provide valuable relative and aggregated information. However, a significant challenge associated with these emerging technologies is their often questionable data quality and highly variable performance, both between different sensor units and under varying environmental conditions (Jiao et al., 2016, Castell et al., 2017). Low-cost sensors can suffer from chemical interference, cross-sensitivity to other gases, and are notably affected by meteorological conditions such as temperature and relative humidity, which can alter particle properties or sensor response, necessitating the critical need for field calibration against reference instrument (Raheja et al., 2023).

2.2.1 Comparative assessment of low-cost sensor measurements To complement satellite-based approaches, low-cost ground-sensors were deployed together to monitor CO_2, NO_x, $PM_{2.5}$, and PM_{10}. As an initial step, we collocated these sensors and systematically evaluated their performance

relative to one another for the shared measured parameters. For this purpose, we deployed four sensors: two AirGradient units (https://www.airgradient.com/) and two Temtop M2000C units (https://www.temtopus.com). The AirGradient sensors are open-source, low-cost devices designed for indoor and outdoor air quality monitoring, capable of measuring $PM_{2.5}$, CO2, temperature, and humidity. The Temtop M2000C sensors are portable commercial monitors widely used for real-time air quality assessment, also providing measurements of $PM_{2.5}$, CO_2, temperature, and humidity. By deploying all four sensors together at the same location, we were able to directly compare their measurements for these common parameters and assess their relative performance under identical environmental conditions.

To evaluate the reliability and agreement of low-cost environmental sensors across multiple physical parameters, a comparative analysis was performed for four measured variables: $PM_{2.5}$, CO_2, Temperature, and Humidity. The analysis covered two sensor models from each brand AirGradient (AG01, AG02) TemTop (TT01, TT02), employing a range of statistical metrics including Root Mean Square Error (RMSE), Normalized RMSE (NRMSE), Mean Absolute Error (MAE), Mean Bias Error (MBE), Pearson correlation coefficient, and paired t-tests to quantify inter- and intra-brand agreement.

Particulate Matter ($PM_{2.5}$) - The $PM_{2.5}$ comparisons revealed excellent intra-brand consistency, especially for AG sensors (AG01-AG02), with a remarkably low RMSE (1.31 µg/m^3), low relative error (NRMSE = 0.09), and near-perfect correlation (Pearson = 0.998). TT sensors (TT01–TT02) also exhibited strong agreement, though with higher RMSE (2.35 µg/m^3) and NRMSE (0.16). The inter-brand comparisons showed generally strong correlations (Pearson $\geq$ 0.969), but higher RMSEs and significant biases, particularly involving TT02. Notably, AG01–TT02 yielded the highest RMSE (3.70 µg/m^3) and a statistically significant mean difference ($p < 0.0001$), suggesting a systematic offset. These findings emphasize that, while trend agreement between brands is high, absolute $PM_{2.5}$ values may vary considerably and require calibration.

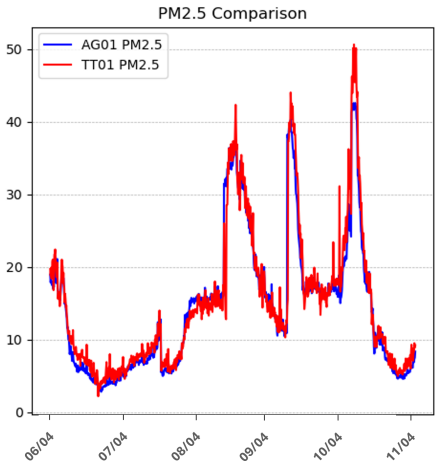

Figure 2. $PM_{2.5}$ time series from AirGradient and TemTop sensors.

Carbon Dioxide (CO_2) - In contrast to $PM_{2.5}$, CO_2 comparisons revealed substantial discrepancies across and within brands, particularly among TT sensors. The AG01–AG02 pair showed strong agreement (RMSE = 4.87 ppm, NRMSE = 0.01, Pearson = 0.970) and no significant mean difference (p = 0.078), indicating good internal consistency. TT01–TT02, however, displayed poor correlation (Pearson = 0.005), despite a relatively low NRMSE (0.05), likely due to high noise or inconsistencies in signal dynamics. Inter-brand comparisons showed large RMSEs (up to 30.85 ppm), low or negligible correlation coefficients, and highly significant mean differences ($p < 0.001$), highlighting a lack of agreement and potential incompatibility between AG and TT CO_2 measurements in their current uncalibrated form.

Temperature was the most consistent variable across all sensor pairs, with low RMSEs, high correlations (Pearson $\geq$ 0.846), and minimal bias. The AG01–AG02 comparison again performed best (RMSE = 0.35 °C, NRMSE = 0.02), supported by nearly perfect correlation (0.999) and only a slight bias (p = 0.041). TT01–TT02 followed with slightly higher RMSE (0.59 °C) but no significant difference in mean (p = 0.984). Inter-brand comparisons (AG vs TT) showed higher RMSEs (1.73-1.84 °C), though still within acceptable ranges for general environmental monitoring. The strong correlation and minimal bias errors across all pairs confirm that temperature is robustly measured by all sensor models, with only minor calibration differences.

Humidity comparisons revealed moderate to high agreement, with the best performance observed again in the AG01–AG02 pair (RMSE = 0.90%, NRMSE = 0.02, Pearson = 0.992). TT01–TT02 showed lower relative accuracy (RMSE = 2.47%, NRMSE = 0.17) and more notable deviation. Inter-brand comparisons were characterized by high RMSEs (up to 7.74%) and statistically significant differences ($p < 0.001$), although correlations remained reasonably strong (Pearson = 0.882–0.902). The consistently higher mean values for TT sensors suggest systematic overestimation of humidity compared to AG sensors, reinforcing the need for cross-sensor calibration in multi-sensor deployments.

3. Local Climate Zones (LCZs) for urban heat island analysis

The UHI is the phenomenon that describes how urban areas have higher temperatures compared to their surroundings. The initiation of this phenomenon can be dated to the start of urbanization and industrialization of human habitats. Solar radiation absorbed by surfaces and buildings creates re-radiation, while anthropogenic activities are also sources of heat (Shahmohamadi et al., 2011). These heat sources are considered the main reason for recording higher temperatures in urban areas (Rizwan et al., 2008), where human presence is more constant and dense, compared to their surroundings.

Traditionally, the UHI effect has been studied by relying on the simple urban-rural dichotomy. However, this simple division lacks sufficient detail to establish an adequate relationship between the complex urban environment and the UHI effect. For example, an urban area, which is usually considered a city or town, has diverse morphological features and land cover types, and each could have a different impact on the thermal environment. It is therefore often inappropriate to classify all these land cover types into the same category. To enable more detailed analysis, the LCZ classification system was developed by I. D. Stewart and T. R. Oke in 2012. The system comprises 17 standard classes, each characterized by uniform sur-

face cover, material properties, structure, and human activity (Stewart and Oke, 2012).

3.1 LCZ Classification

To study the thermal regime of urban areas, the LCZ classification system has been increasingly utilized in recent years. For LCZ map production, methodologies can be categorized into three distinct approaches: Remote Sensing (RS) based, Geographic Information System (GIS) based, or hybrid combinations of both techniques (Huang et al., 2023). Vavassori et al. (2024) proposed a novel hybrid GIS and Remote Sensing based workflow (Figure 3) for LCZ mapping leveraging hyperspectral satellite imagery, a data source that has received limited attention for this application. The proposed methodology integrates hyperspectral PRISMA imagery, multispectral Sentinel-2 data, and urban canopy parameters (UCPs), including building height, impervious surface fraction, and sky view factor, to generate comprehensive LCZ maps (Vavassori et al., 2024).

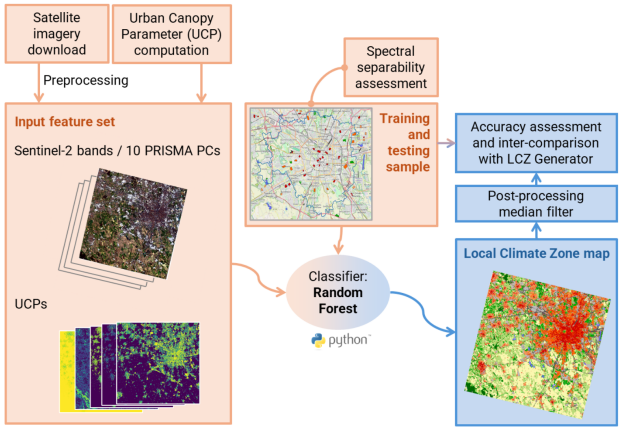

Figure 3. Workflow of proposed LCZ map generation methodology by Vavassori et al. (2024)

The first step of the methodology involves the preprocessing of Sentinel-2 Level-2A and PRISMA Level-2D bottom-of-atmosphere reflectance products. Due to the known georeferencing accuracy in PRISMA imagery (nominally better than 200 m), co-registration with temporally corresponding Sentinel-2 data represents a critical preprocessing step. For this purpose, Sentinel-2 bands B02-B07, B8A, B11, and B12, provided at 20 m spatial resolution by the Copernicus program, are resampled to 30 m to achieve resolution compatibility with PRISMA imagery. The open-source GeFolki algorithm, implemented in Python, is employed to correct geometric distortions in PRISMA image georeferencing. This process requires spectral bands with comparable wavelengths from both sensors. Specifically, the Sentinel-2 red band is paired with the corresponding PRISMA band at 575.49 nm central wavelength. The algorithm generates a displacement matrix containing vertical and horizontal pixel corrections, which is subsequently applied to all Short-Wave Infrared (SWIR) and Visible-Near Infrared (VNIR) PRISMA bands. Following geometric correction, PRISMA imagery undergoes Principal Component Analysis (PCA) using the scikit-learn Python library to reduce dimensionality and eliminate redundant spectral information to optimize computational efficiency for the next processing steps.

Five UCPs, describing the urban morphology and imperviousness, are computed at 20 m and 30 m resolution (to match the spatial resolution of Sentinel-2 and PRISMA imagery, respectively), and integrated to the satellite spectral data. UCPs comprise Sky View Factor, Impervious Surface Fraction, Building Surface Fraction, Tree Canopy Height, and Building Heights, each contributing essential morphological information for LCZ discrimination. Sky View Factor is computed with the SAGA GIS software from the ALOS Digital Surface Model (DSM) processed through SAGA GIS software. Impervious Surface Fraction is derived from the Copernicus Imperviousness Density layer, providing validated soil sealing information across European territories. Building-related parameters are extracted from regional geo-topographic databases (DBGT), while Tree Canopy Height data is obtained from the Global Sentinel-2 Canopy Height product. All UCP layers undergo normalization to the [0-1] range to ensure compatibility with satellite reflectance data.

Training and validation samples are systematically collected by relying on multi-source ancillary data, to ensure an accurate representation of each LCZ class. The sampling strategy employs 30 m resolution RGB PRISMA imagery for land cover interpretation, 5 m panchromatic PRISMA data for precise boundary delineation, and building height information for built-up class discrimination. Sample distribution adheres to spatial and thematic balance principles, ensuring fair representation across LCZ classes and the study area extent. Spectral separability of LCZs is also assessed by computing the Jeffries-Matusita (JM) distance metric. JM values approach 2.0 for completely separable spectral signatures and 0.0 for identical signatures. To mitigate issues arising from high inter-band correlation in PRISMA data that may result in singular covariance matrices, spectral signatures are sampled at 10-band intervals.

The RF ensemble learning algorithm serves as the classification approach, applied consistently to both datasets (PRISMA and Sentinel-2) using the training sample polygons. PRISMA classification is based on the first 10 principal components, accounting for approximately 100% of original spectral variance; Sentinel-2 classification incorporates all selected spectral bands. In both cases, satellite data is augmented with UCP layers as supplementary feature vectors. Hyperparameter tuning follows a systematic grid search approach using GridSearchCV with repeated 5-fold cross-validation. The optimization procedure evaluates multiple parameter combinations, including estimator quantity, maximum features per split, and split quality criteria (Gini impurity versus information entropy). Training data is split into 80% training and 20% validation subsets for cross-validation.

Classification outputs undergo postprocessing using a 3×3 pixel median filter to reduce classification noise and merge isolated LCZ pixels into spatially coherent class regions. This smoothing procedure enhances map continuity while preserving overall classification accuracy. Comprehensive accuracy assessment employs standard metrics derived from confusion matrices, including Overall Accuracy (OA), precision, recall, and F1-score. Inter-comparison analysis between PRISMA and Sentinel-2 derived maps, and LCZ Generator reference maps, shows agreement. For comparative analysis, all products are resampled to 10 m resolution using nearest-neighbor interpolation and geometrically aligned to a common reference grid. Statistical validation implements stratified random sampling protocols, selecting 1,500 pixel pairs per comparison based on Cochran's formula for large population statistics. This sampling framework ensures adequate representation across LCZ classes while maintaining statistical significance and validity.

Results obtained for the Metropolitan City of Milan demonstrate significant improvements in LCZ classification accuracy through the proposed methodology. PRISMA hyperspectral data yielded superior performance compared to Sentinel-2 multispectral imagery, achieving an average OA increase of 5% and exhibiting reduced confusion between built-up LCZ classes, which aligns with the higher spectral separability observed in PRISMA data. The integration of UCPs proved beneficial for enhancing classification accuracy, particularly when combined with Sentinel-2 data. Furthermore, the proposed workflow demonstrated superior performance compared to the existing LCZ Generator methodology for both built-up and land cover LCZ types, with improvements reaching up to 16% in OA when utilizing PRISMA data compared to the LCZ Generator approach. Figure 4 represents an LCZ map obtained for Milan using the PRISMA image of June 17th 2023.

The complete methodological implementation and source code are publicly available through the GitHub repository (`https://github.com/gisgeolab/LCZ-ODC`).

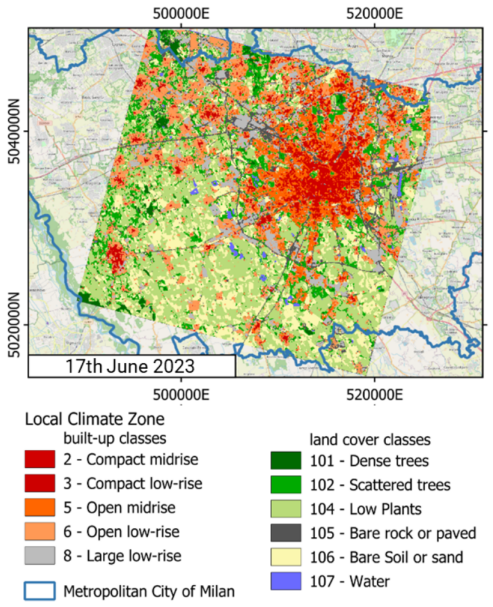

Figure 4. LCZ map over Milan relative to 17th June 2023.

3.2 Ongoing Improvements of LCZ Classification Methodology

The mentioned methodology is being successfully utilized to derive LCZ maps for different urban regions such as Milan, Wuhan (China), Cairo (Egypt), Toronto (Canada), and Ho Chi Minh City (Vietnam). However, there are ongoing efforts to enhance some critical steps of the workflow. Although the code pipeline is designed to be as automatic as possible, there are still steps that can be improved to enhance automation. More specifically, the data retrieval phase and creation of training and test samples are the main steps that could be further enhanced to reduce time consumption and complexity.

3.2.1 Automate Data Retrieval Satellite data acquisition is an essential component of the LCZ classification workflow mentioned above. It can be time-consuming for retrieval and complex for integration into the code pipeline. GEE is a cloud-based platform that contains a vast catalogue of satellite imagery and other geospatial datasets. Additionally, it provides a powerful computing infrastructure that allows for manipulation of large datasets within short timeframes. This centralized data source enables more robust access to data from different sources, and by utilizing the functionalities of the GEE API, it is possible to integrate satellite data directly into the LCZ classification code pipeline. It should be noted that as of the time of writing this manuscript, only Sentinel-2 and UCPs can be accessed via the GEE database, while PRISMA has not yet been integrated.

3.2.2 Automate Training/Test Samples Creation Training samples allow the model to learn patterns in the data, while test samples evaluate the model's performance on unseen data. It is crucial to provide high-quality samples to avoid overfitting and develop a predictive model with high accuracy. These factors highlight the importance of this step in the methodology and require precise attention. However, this step is performed manually, which can introduce human errors in the samples and does not guarantee that the user can observe all scene characteristics accurately.

As a solution, we propose to use urban spatial indices (USI), which are quantitative measures used to analyze and understand the spatial patterns and characteristics of urban areas. Since we already have LCZ maps available for Milan, we will calculate the USI for this city. By utilizing statistical measures, it is possible to determine whether any relationships exist between specific LCZ classes and the USI. Once these relationships are established, it becomes possible to use the USI to programmatically identify regions that represent good candidates for inclusion in training and test samples.

3.2.3 Training/Test Samples Quality Check To ensure accurate classification results, it is essential to have reliable training and test samples. Therefore, it is necessary to establish a standard framework for assessing the quality of samples. In this regard, samples should meet the following criteria to be accepted. As mentioned in Stewart et al. (2012), the geometric and surface cover parameters have specific thresholds within each class. Therefore, UCP values within each training and test sample should fall within the standard threshold ranges. Currently, a study is being conducted to examine the distribution of UCPs in existing training and test samples for the Milan case study. Secondly, individual polygons should have an aspect ratio below 3, and the area of the samples should be greater than 0.04 km² (Demuzere et al., 2021). Lastly, it is essential to ensure that the polygons are balanced across classes so that the algorithm receives an adequate number of samples for each class.

3.2.4 Trying Other Machine Learning Methods Although RF is a valid and widely used method for classifying LCZs, offering advantages such as high flexibility and the ability to handle complex, multivariable data without strong assumptions about input distributions, it can be limited in capturing spatial patterns, especially when compared to deep learning approaches like convolutional neural networks (CNNs) (Fung et al., 2022). To address these limitations and ensure that our classification results are robust and potentially more accurate, we are also testing other methods, including CNNs and geospatial foundation models. By comparing these different approaches, we aim to identify the most effective method for our specific application and data characteristics.

4. Conclusion

The satellite-based modeling framework developed in this study with the integration of GEE data, enabled the estimation of ground-level NO_2, SO_2, and CO concentrations using Sentinel-5P data, ERA5 meteorology, and CAMS reanalysis, demonstrating robust and scalable performance across multiple pollutants. Additionally, AirGradient sensors showed excellent internal consistency across all variables, outperforming TemTop sensors particularly for CO_2. Temperature was the most reliably measured parameter, followed by $PM_{2.5}$. CO_2 and humidity exhibited the greatest inter-sensor variability, especially across brands. Future research should focus on clarifying the causes of the poor CO_2 performance of TemTop sensors, as current results indicate significant discrepancies and low correlation compared to AG sensors. Overall, these results emphasize the need for sensor-specific calibration when integrating multi-sensor data, particularly for gases, where measurement uncertainty may be more pronounced.

The second line of research in this work demonstrates the effectiveness of an advanced, hybrid GIS and Remote Sensing based workflow for LCZ classification, leveraging hyperspectral PRISMA imagery, multispectral Sentinel-2 data, and UCPs. The proposed methodology yielded significant improvements in classification accuracy in Milan, particularly through the integration of PRISMA data and UCPs, and outperformed existing approaches such as the LCZ Generator. Activities are ongoing for data retrieval automation, training/test sample creation, implementation of robust quality control frameworks, further enhance the reproducibility and reliability of the results. Initial experiments with alternative ML methods, including deep learning models, show promise for further advancements. The workflow has been successfully applied to diverse urban contexts and is adaptable for future expansion.

Acknowledgements

This study was carried out within the Space It Up project funded by the Italian Space Agency, ASI, and the Ministry of University and Research, MUR, under contract n. 2024-5-E.0 - CUP n. I53D24000060005.

References

Allred, E. N., Bleecker, E. R., Chaitman, B. R., Dahms, T. E., Gottlieb, S. O., Hackney, J. D., Pagano, M., Selvester, R. H., Walden, S. M., Warren, J., 1989. Short-term effects of carbon monoxide exposure on the exercise performance of subjects with coronary artery disease. *New England journal of medicine*, 321(21), 1426–1432. doi.org/10.1056/NEJM198911233212102.

Brauer, M., Freedman, G., Frostad, J., Van Donkelaar, A., Martin, R. V., Dentener, F., Dingenen, R. V., Estep, K., Amini, H., Apte, J. S., Balakrishnan, K., Barregard, L., Broday, D., Feigin, V., Ghosh, S., Hopke, P. K., Knibbs, L. D., Kokubo, Y., Liu, Y., Ma, S., Morawska, L., Sangrador, J. L. T., Shaddick, G., Anderson, H. R., Vos, T., Forouzanfar, M. H., Burnett, R. T., Cohen, A., 2016. Ambient Air Pollution Exposure Estimation for the Global Burden of Disease 2013. *Environmental Science & Technology*, 50(1), 79–88. doi.org/10.1021/acs.est.5b03709.

Cai, K., Shao, Y., Lin, Y., Li, S., Fan, M., 2025. Estimating NOx Emissions in China via Multisource Satellite Data and Deep Learning Model. *Remote Sensing*, 17(7). doi.org/10.3390/rs17071231.

Castell, N., Dauge, F. R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., Bartonova, A., 2017. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International*, 99, 293–302. doi.org/10.1016/j.envint.2016.12.007.

Cedeno Jimenez, J. R., Brovelli, M. A., 2023. NO2 Concentration Estimation at Urban Ground Level by Integrating Sentinel 5P Data and ERA5 Using Machine Learning: The Milan (Italy) Case Study. *Remote Sensing*, 15(22), 5400. doi.org/10.3390/rs15225400.

Demuzere, M., Kittner, J., Bechtel, B., 2021. LCZ Generator: a web application to create Local Climate Zone maps. *Frontiers in Environmental Science*, 9, 637455. doi.org/10.3389/fenvs.2021.637455.

Fioletov, V. E., McLinden, C. A., Krotkov, N. et al., 2020. Anthropogenic and volcanic point-source SO_2 emissions derived from TROPOMI on board Sentinel-5 Precursor: first results. *Atmospheric Chemistry and Physics*, 20, 5591–5629. doi.org/10.5194/acp-20-5591-2020.

Fioletov, V. E., McLinden, C. A., Krotkov, N., Li, C., Joiner, J., Theys, N., Carn, S., Moran, M. D., 2016. A global catalogue of large SO 2 sources and emissions derived from the Ozone Monitoring Instrument. *Atmospheric Chemistry and Physics*, 16(18), 11497–11519. doi.org/10.5194/acp-16-11497-2016.

Fung, K. Y., Yang, Z.-L., Niyogi, D., 2022. Improving the local climate zone classification with building height, imperviousness, and machine learning for urban models. *Computational Urban Science*, 2(1), 16. doi.org/10.1007/s43762-022-00046-x.

Griffin, D., Zhao, X., McLinden, C. A., Boersma, F., Bourassa, A., Dammers, E., Degenstein, D., Eskes, H., Fehr, L., Fioletov, V. et al., 2019. High-resolution mapping of nitrogen dioxide with TROPOMI: First results and validation over the Canadian oil sands. *Geophysical Research Letters*, 46(2), 1049–1060. doi.org/10.1029/2018GL081095.

Gu, J., Tao, J., Chen, L., Fan, M., Tian, Y., 2025. High-Resolution Mapping of NO2 Population Exposure in China from Satellite Observations. *Environmental Development*, 101238. doi.org/10.1016/j.envdev.2025.101238.

Huang, F., Jiang, S., Zhan, W., Bechtel, B., Liu, Z., Demuzere, M., Huang, Y., Xu, Y., Ma, L., Xia, W. et al., 2023. Mapping local climate zones for cities: A large review. *Remote Sensing of Environment*, 292, 113573. doi.org/10.1007/s43762-022-00046-x.

Irfeey, A. M. M., Chau, H.-W., Sumaiya, M. M. F., Wai, C. Y., Muttil, N., Jamei, E., 2023. Sustainable mitigation strategies for urban heat island effects in urban areas. *Sustainability*, 15(14), 10767. doi.org/10.3390/su151410767.

Jiao, W., Hagler, G., Williams, R., Sharpe, R., Brown, R., Garver, D., Judge, R., Caudill, M., Rickard, J., Davis, M., Weinstock, L., Zimmer-Dauphinee, S., Buckley, K., 2016. Community Air Sensor Network (CAIRSENSE) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern United States. *Atmospheric Measurement Techniques*, 9(11), 5281–5292. doi.org/10.5194/amt-9-5281-2016. Publisher: Copernicus GmbH.

Landgraf, J., Scheepmaker, R., Borsdorff, T., Hu, H., Houweling, S., Butz, A., Aben, I., Hasekamp, O. et al., 2016. Carbon monoxide total column retrievals from TROPOMI shortwave infrared measurements. *Atmospheric Measurement Techniques*, 9(10), 4955–4975. doi.org/10.5194/amt-9-4955-2016.

Levelt, P. F., Stein Zweers, D. C., Aben, I., Bauwens, M., Borsdorff, T., De Smedt, I., Eskes, H. J., Lerot, C., Loyola, D. G., Romahn, F. et al., 2021. Air quality impacts of COVID-19 lockdown measures detected from space using high spatial resolution observations of multiple trace gases from Sentinel-5P/TROPOMI. *Atmospheric Chemistry and Physics Discussions*, 2021, 1–53. doi.org/10.5194/acp-22-10319-2022.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T., 2017. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental pollution*, 231, 997–1004. doi.org/10.1016/j.envpol.2017.08.114.

Liu, B., Guo, X., Jiang, J., 2023. How urban morphology relates to the urban heat island effect: A multi-indicator study. *Sustainability*, 15(14), 10787. doi.org/10.3390/su151410787.

Liu, J., 2021. Mapping high resolution national daily NO2 exposure across mainland China using an ensemble algorithm. *Environmental Pollution*, 279, 116932. doi.org/10.1016/j.envpol.2021.116932.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E., 2020. Environmental and health impacts of air pollution: a review. *Frontiers in Public Health*, 8, 14. doi.org/10.3389/fpubh.2020.00014.

Meo, S. A., Shaikh, N. et al., 2024. Effect of air pollutants particulate matter ($PM_{2.5}$, PM_{10}), sulfur dioxide (SO_2) and ozone (O_3) on cognitive health. *Scientific Reports*, 14, 19616. doi.org/10.1038/s41598-024-70646-6.

Raheja, G., Nimo, J., Appoh, E. K.-E., Essien, B., Sunu, M., Nyante, J., Amegah, M., Quansah, R., Arku, R. E., Penn, S. L., Giordano, M. R., Zheng, Z., Jack, D., Chillrud, S., Amegah, K., Subramanian, R., Pinder, R., Appah-Sampong, E., Tetteh, E. N., Borketey, M. A., Hughes, A. F., Westervelt, D. M., 2023. Low-Cost Sensor Performance Intercomparison, Correction Factor Development, and 2+ Years of Ambient $PM_{2.5}$ Monitoring in Accra, Ghana. *Environmental Science & Technology*, 57(29), 10708–10720. doi.org/10.1021/acs.est.2c09264.

Rizwan, A. M., Dennis, L. Y. et al., 2008. A review on the generation, determination and mitigation of Urban Heat Island. *Journal of environmental sciences*, 20(1), 120–128. doi.org/10.1016/S1001-0742(08)60019-4.

Shahmohamadi, P., Che-Ani, A., Maulud, K., Tawil, N. M., Abdullah, N., 2011. The impact of anthropogenic heat on formation of urban heat island and energy consumption balance. *Urban Studies Research*, 2011(1), 497524. doi.org/10.1155/2011/497524.

Smith, E. K., Fournier de Lauriere, C., Henninger, E., 2025. Persistent inequalities in global air quality monitoring should not delay pollution mitigation. *Proceedings of the National Academy of Sciences*, 122(18), e2423259122. doi.org/10.1073/pnas.2423259122.

Stewart, I. D., Oke, T. R., 2012. Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, 93(12), 1879–1900. doi.org/10.1175/BAMS-D-11-00019.1.

Van Geffen, J., Boersma, K. F., Eskes, H., Sneep, M., Ter Linden, M., Zara, M., Veefkind, J. P., 2020. S5P TROPOMI NO 2 slant column retrieval: Method, stability, uncertainties and comparisons with OMI. *Atmospheric Measurement Techniques*, 13(3), 1315–1335. doi.org/10.5194/amt-13-1315-2020.

Vavassori, A., Oxoli, D., Venuti, G., Brovelli, M. A., de Cumis, M. S., Sacco, P., Tapete, D., 2024. A combined Remote Sensing and GIS-based method for Local Climate Zone mapping using PRISMA and Sentinel-2 imagery. *International Journal of Applied Earth Observation and Geoinformation*, 131, 103944.

Veefkind, P., Aben, I., McMullan, K. et al., 2012. TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sensing of Environment*, 120, 70–83.

WHO, 2024. Air quality, energy and health: Types of pollutants. https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts/types-of-pollutants. Accessed: 2025-06-23.

Yang, Q., Yuan, Q., Gao, M., Li, T., 2022. A new perspective to satellite-based retrieval of ground-level air pollution: Simultaneous estimation of multiple pollutants based on physics-informed multi-task learning. *Science of The Total Environment*, 857, 159542. doi.org/10.1016/j.scitotenv.2022.159542.

Zhou, L., Shao, Z., Wang, S., Huang, X., 2022. Deep learning-based local climate zone classification using Sentinel-1 SAR and Sentinel-2 multispectral imagery. *Geo-Spatial Information Science*, 25(3), 383–398. doi.org/10.1080/10095020.2022.2030654.

Zhou, S., Wang, W., Zhu, L., Qiao, Q., Kang, Y., 2024. Deep-learning architecture for PM2. 5 concentration prediction: A review. *Environmental Science and Ecotechnology*, 21, 100400. doi.org/10.1016/j.ese.2024.100400.