

Data-driven analysis of Urban Heat Island phenomenon based on street typology



Monica Pena Acosta ^a, Faridaddin Vahdatikhaki ^{a,*}, João Santos ^a, Sandra Patricia Jarro ^b, Andries G. Dorée ^c

^a Department of Construction Management and Engineering, University of Twente, Horsttoren Z-204, Drienerlolaan 5, Enschede 7522 NB, the Netherlands

^b Campus del Puente del Común, Universidad de la Sabana, Km. 7, Autopista Norte de Bogotá, Chía, Cundinamarca, Colombia

^c Department of Construction Management and Engineering, University of Twente, Horsttoren Z-219, Drienerlolaan 5, Enschede, 7522 NB the Netherlands

ARTICLE INFO

Keywords:

Urban heat islands
Data-driven modeling
Street typology
UHI mechanism

ABSTRACT

This study explores the intricate relationship between diverse street types and the urban heat island (UHI) phenomenon - a major urban issue where urban regions are warmer than their rural counterparts due to anthropogenic heat release and absorption by urban structures. UHI leads to increased energy consumption, diminished air quality, and potential health hazards. This research posits that a sample of representative streets (i.e., a few streets from each type of street) will be sufficient to capture and model the UHI in an urban context, accurately reflecting the behavior of other streets. To do so, streets were classified into unique typologies based on (1) socio-economic and morphological attributes and (2) temperature profiles, utilizing two clustering methodologies. The first approach employed K-Prototypes to categorize streets according to their socio-economic and morphological similarities. The second approach utilized Time Series Clustering K-Means, focusing on temperature profiles. The findings indicate that models retain strong performance levels, with R-Squared values of 0,85 and 0,80 and MAE ranging from 0,22 to 0,84 °C for CUHI and SUHI respectively, while data collection efforts can be reduced by 50 to 70 %. This highlights the value of the street typology in interpreting UHI mechanisms. The study also stresses the need to consider the unique aspects of UHI and the temporal variations in its drivers when formulating mitigation strategies, thereby providing new insights into understanding and alleviating UHI effects at a local scale.

1. Introduction

Urban heat island (UHI) is a phenomenon in which urban areas are warmer than their surrounding rural areas, due to the presence of buildings, pavement, and other infrastructure that absorbs and retains heat (Howard, 1833). This can lead to a range of negative impacts, including increased energy use for cooling, reduced air quality, and heat-related illness and mortality (Akbari & Kolokotsa, 2016). The effects of UHI are typically more pronounced during the night when the temperature difference between urban and rural areas is greatest. This is because the heat absorbed by urban infrastructure during the day is slowly released back into the atmosphere at night, resulting in higher temperatures in urban areas (Oke, 1982).

According to the energy budget equation of the urban canopy layer (Oke, 1988b) the underlying causes for the development of UHI can be divided into four key components: (1) absorption of short and long-wave

radiation, (2) transpiration from buildings and infrastructure, (3) release of anthropogenic heat from inhabitants and appliances, and (4) airflow blocking effect of urban geometries. However, the effects of these are unique to the morphology and socio-economic factors of each city. The design and layout of the built environment lead to the creation of distinct areas within a city experiencing varying temperatures, known as Urban Local Climate Zones (ULCZs) (Stewart & Oke, 2012). These zones have their own interplay between morphological, and socio-economic parameters, which results in different intensities of UHIs throughout the city.

Physic-based methods, which involve simulating thermodynamic processes, are commonly used to model and study UHI in the built environment (Grimmond et al., 2010). These methods require detailed information about the materials, geometries, and local weather considerations in order to accurately represent the built environment (Adilkhanova et al., 2022). While these methods offer a detailed

* Corresponding author.

E-mail address: f.vahdatikhaki@utwente.nl (F. Vahdatikhaki).

understanding of the UHI, they require specialized expertise and resources to conduct the simulation. This can make the modeling process complex and difficult, especially when considering the real-world applications and usability of the detailed models (Mirzaei, 2015). Yet, based on the energy budget equation, one-size-fits-all “recipes” for how to address this phenomenon have been suggested and implemented worldwide (Manoli et al., 2019). These include (1) the implementation of green infrastructure (Wong et al., 2021), such as increased vegetation, green roofs and facades, cool roofs, reflective pavements, wall coatings, and window films (Salvati et al., 2022), and (2) modifying the street geometry to enhance wind flow (Aleksandrowicz et al., 2017). While these strategies can be effective in mitigating UHI at the micro-level, their effectiveness varies based on the specific characteristics of the built environment to which they are going to be applied (Qi et al., 2020). For instance, increasing the albedo of paved surfaces could be counter-effective on streets with a small height-to-width ratio (H/W) that consists mostly of light-colored building facades. This is because solar radiation bounces between the buildings and the pavement in an urban canyon (Salvati et al., 2022). In practical terms, this context-dependency means that effective strategies should be tailored to individual contexts and that it is important to refrain from using one-size-fits-all off-the-shelf mitigation solutions. Yet, this would require precise and user-friendly UHI simulation and modeling tools that can support urban decision-making on a case-by-case basis. As mentioned earlier, the existing physics-based modeling tools, albeit accurate, are hard to use by non-experts. This creates a practical gap in terms of the absence of tools required for UHI-conscious urban decision-making. One attempt to bridge this gap is the development of the UHI-DS tool (Ding et al., 2023), a web-based decision-support tool that helps in informing urban policy and planning practices by suggesting potential building and urban interventions to mitigate UHI. Despite the utility of the UHI-DS tool, there remains a significant potential for data-driven approaches to simplify UHI modeling and decision-making further.

Data-driven approaches have shown significant potential in reducing the complexity and user-friendliness of physics-based models (Adilkhanova et al., 2022). By identifying patterns in the data, these methods can provide a more efficient and user-friendly way to analyze the phenomena without using heat exchange models or advanced numerical simulations (Adilkhanova et al., 2022; Lyu et al., 2022; Oukawa et al., 2022; Peng et al., 2022; Venter et al., 2021; Wang et al., 2020). The authors have previously presented a framework for the development of an easy-to-use data-driven UHI modeling tool that can support urban planners (Pena Acosta et al., 2021). Data-driven models have also been successfully used to map different ULCZs across the globe (Feng & Liu, 2022). However, the majority of these models utilize satellite or aerial imagery to assess Surface UHI (SUHI). In doing so, the data-driven assessment and modeling of Canopy UHI (CUHI) have been marginalized. To give some background about different types of UHI, it should be noted that the different thermodynamic processes within the built environment led to various types of UHI. CUHI is primarily caused by the reduction of transpiration and evaporation within the urban canopy, which reduces the cooling effect of vegetation and leads to higher temperatures. CUHI is typically measured by means of weather stations that collect data at a high temporal resolution but are only located in a few specific areas (thus, limited spatial resolution) (Ching et al., 2018; See et al., 2015). SUHI, on the other hand, is caused by the absorption of solar radiation by the urban fabric. SUHI is measured using land surface temperature (LST) data obtained from thermal infrared (TIR) satellite imagery, resulting in a higher spatial resolution but lower temporal resolution. As a result, the amount of data available for the study of SUHI is considerably more and it can be accessed more readily. It is exactly the difference in the temperature measurement method that introduces an imbalance in the data-driven study of CUHI and SUHI. From the standpoint of mitigation strategies, it is crucial to study CUHI and SUHI simultaneously (Du et al., 2021; Peng et al., 2022; Sun et al., 2020;

Venter et al., 2021; Yang et al., 2020), as it is the combination of the two that UHI effects in cities. Nonetheless, the majority of current studies do not take into the interplay between CUHI and SUHI mostly because the available data are not at the same spatial and temporal resolutions.

In previous studies, the authors demonstrated a systematic and rigorous data collection strategy that can address the problem of inconsistent data for the study of CUHI and SUHI fields (Pena Acosta et al., 2022, 2023a). Using consistent and well-structured street-level air and surface temperatures at the same spatiotemporal resolution enables the application of machine learning (ML) techniques to uncover hidden relationships between urban factors and temperature changes within the built environment. Such data-driven models would enable us to better understand and strategize for the mitigation of UHI.

Given the potential of ML models to accurately explain UHI, it is interesting to explore the extent to which these modes are transferable across different urban contexts. This is important because the generalizability of models allows a limited number of UHI models to be used for many urban contexts, relieving urban authorities from the need to collect data per city. In previous research, the authors have performed a comprehensive generalizability/transferability analysis of UHI data-driven models (Pena Acosta et al., 2023a). The results indicate that the UHI model of one city is not transferable to another city regardless of morphological similarity and geographic proximity. This finding, once again, attests to the context-specificity of the UHI models.

Since the data-driven UHI models showed low transferability across different urban contexts, it is still relevant to investigate the extent to which UHI demonstrates similar behavioral patterns within each city. In other words, it is not clear how generalizable UHI models are when they are developed for specific street types in the same city. The mapping of the UHI mechanism to street typology offers the advantage that the urban planner can develop specific sets of mitigation strategies per street type. In doing so, they will be able to highly standardize their mitigation policy without falling back on the one-size-fits-all approach. To the best of the authors' knowledge, there is no insight into the relationship between street typology and UHI mechanisms.

Based on the above, this paper builds on the previous work of the authors to investigate the relationship between street typology in an urban context and UHI driving mechanisms. The main hypothesis tested in this research is that a sample of representative streets (i.e., a few streets from each type of street) will be sufficient to capture and model the UHI in an urban context. If proven, this means that urban authorities only need to focus on collecting data from a limited number of streets, which makes the data collection easier and cheaper. This in turn contributes to making the whole practice of data-driven modeling of UHI more accessible and available to a wider range of cities. To clarify, street typology in the context of UHI refers to streets that have similar socio-economic and morphological characteristics. The novelty of this research lies in two aspects: (1) it explores the use of street typologies to study UHI. Each typology represents a unique combination of urban elements, allowing for the identification of how different street features influence temperature patterns. This has the potential to simplify data collection by selecting a sample of streets from each group that accurately represents the behavior of the entire group, and (2) it identifies the defining features of each group, which then serve as representative features of each typology. This means that the identified features can be used to characterize and differentiate the various street types, shedding light on temporal variations in temperature patterns.

In order to provide a comprehensive overview of the research, the rest of the paper is organized as follows: Section 2 presents a brief introduction to the current state of the art. Section 3 presents a summary of the dataset used, as well as explaining the research methodology. Section 4 presents the results. Finally, Section 5 offers a discussion of the findings and final conclusions drawn from the research.

2. Literature review

The relationship between the urban form and UHI has been widely established in the current body of knowledge (Oke, 1988a). Previous studies have utilized statistical methods and satellite imagery to investigate the connection between various aspects of the built environment, its geometry and the intensity of UHI (Unger, 2004). These studies have examined concepts such as fractal geometry, urban density, spatial connectivity, and city size (Chakraborty et al., 2020; Dewan et al., 2021; Du et al., 2021; Huang et al., 2020; Liu et al., 2022; Shen et al., 2021; Wang et al., 2020). However, while helpful in understanding the factors contributing to UHI, the existing literature tends to focus on studying cities as a whole, potentially leading to generalized assumptions that may not always be accurate (Manoli et al., 2019; Pena Acosta et al., 2023a).

In contrast, Stewart and Oke (2012) provided valuable insights into the intraurban nature of UHI. Their approach defined distinct areas within a city that demonstrate unique responses to the local climate. They introduced the concept of ULCZs initially designed for air temperatures. Each ULCZs represents specific urban microclimates that are influenced by how different urban elements, such as buildings, roads, vegetation, and water bodies, are arranged and composed. The fundamental idea behind ULCZs is to classify the urban landscape into zones that display similar climatic behaviors. Nevertheless, there are drawbacks to this approach: (1) the categories are determined based on general characteristics, focusing mainly on global and regional factors. This may lead to an oversight of local influences; (2) the categorization used in the ULCZs system is built under the premise that it is universal and can be applied to all urban contexts globally. However, the previous study of the authors demonstrated that this premise does not necessarily hold, and it is better to develop the local climate zones per urban context, and (3) the ULCZs were primarily designed for CUHI and may not fully encompass the interaction between CUHI and SUHI in the resulting zones.

One potential approach to address these limitations is by examining UHI at the street-level. Streets at the ground level offer distinct microclimates that respond differently to the local climate (Geng et al., 2023). As discussed by the authors in previous research (Pena Acosta et al., 2021), by considering street-level data, it is possible to capture the intricate relationship between various urban elements, such as buildings, roads, vegetation, and water bodies, and their impact on local temperatures. For instance, Mahroof et al. (2020) successfully identified typologies based on street design. However, they used a limited set of parameters, particularly related to the ULCZs classification, such as openness, surface properties, and façade geometry. While these parameters are relevant, they tend to simplify the complex nature of the built environment.

With the vast amount of data being generated by the built environment, there is a unique opportunity to harness its potential (Adilkhanova et al., 2022; Creutzig et al., 2019). In the past decade, numerous data-driven methods have focused on the predictability (Shen et al., 2022) and quantification (Li et al., 2019; Mohammad & Goswami, 2021) of UHI. However, only a few studies have utilized data-driven techniques to understand the similarities between urban elements that result in similar UHI patterns (Chakraborty & Lee, 2019; Chao et al., 2020; Dang & Kim, 2023; Litardo et al., 2020; Peng et al., 2022; Ullah et al., 2023).

Litardo et al. (2020) conducted a spatial distribution analysis to estimate urban morphology parameters, properties of materials, and anthropogenic heat emissions in the streets, by applying the K-Means clustering technique to twenty-eight randomly selected urban samples. Their approach involved a combination of physics-based and statistical analysis to investigate the parameters influencing the simulation outcomes. The results of the clustering process successfully revealed four distinct urban clusters. However, the study was limited to a specific set of parameters and did not explore additional potential factors, such as

socio-economic features.

To apply proper clustering for the study of UHI, it is important to consider both categorical and numerical features because some of the socio-economic factors are categorical in nature, e.g., land use. For this K-Prototypes algorithm, which is a variant of the K-Means algorithm, is proposed (Ahmad and Dey (2007)). Another important factor to consider is that given the fact that the temperature measurement is time-series data, it is important to use an appropriate cluster method that can handle time-series data. The Time Series K-Means approach extends traditional K-Means to handle temporal data (Aghabozorgi et al., 2015). This method identifies patterns and clusters similar data points in time series by considering the sequential nature and inherent temporal structures of the data. Unlike traditional K-Means, which treats each data point as independent, Time Series K-Means takes into account the order and continuity of data points. As a result, it groups together time series that have similar shapes or movements over time, providing a more meaningful clustering of temporal data. This makes it especially suitable for analyzing both CUHI and SUHI, where temporal trends are significant. Despite its potential, the usage of Time Series K-Means in the context of UHI is still not widespread (Liu et al., 2018).

Based on the above, the question remains as to what extent clustering urban elements can lead to street typologies that provide a deeper understanding of the drivers of UHI. In this context, alternative clustering methods such as K-Prototypes and Time Series K-Means could potentially enhance our understanding of UHI, providing a more comprehensive view of the influencing factors. In summary, by analyzing morphological and socioeconomic features at the same spatiotemporal resolution for both CUHI and SUHI, could be possible to provide a deeper understanding of the mechanism of UHI at the street-level and pave the way for informed urban heat resilience.

3. Research methodology

The research methodology used in this study consisted of four main phases, as shown in Fig. 1. The first stage pertains to the data collection phase. Here, data used to generate different types of street typologies was collected, cleaned if necessary, and structured to feed into the next stage. In the second phase, the dataset was examined to identify common patterns in streets. To achieve this, two distinct clustering approaches were applied: (1) Clustering streets based on the feature domain using K-Prototypes. In this approach, streets were classified based on their similarities in terms of socioeconomic and morphological features. In this approach, streets ending up in the same cluster are expected to have similar physical characteristics but may not necessarily have a similar temperature profile; (2) Clustering based on the label domain using Time Series Clustering K-Means. Contrary to the previous approach, here the streets were classified based on their temperature profile (both surface and canopy) similarities. In other words, streets ending up in a cluster have rather similar temperature profiles while they may not necessarily have a high degree of similitude with respect to socioeconomic and morphological features. The resulting clusters obtained from both approaches were assessed on the absolute error (MAE) and the highest R-squared (R^2) value. This dual approach towards clustering made it possible to investigate the impact of using representative samples for large groups of streets based on either street features or temperature profiles. It is important to mention in the second approach, once the clusters are identified based on their temperature profile, the governing features that explain this classification were identified. These governing features can be used to narrow down the scope of data collection by urban authorities. In the third phase, a machine learning approach was employed to develop a model capable of revealing hidden patterns within each street type. Lastly, the fourth phase involved conducting feature analysis, wherein each generated street group was dissected based on their similarities. These phases are discussed in detail in the subsequent sections.

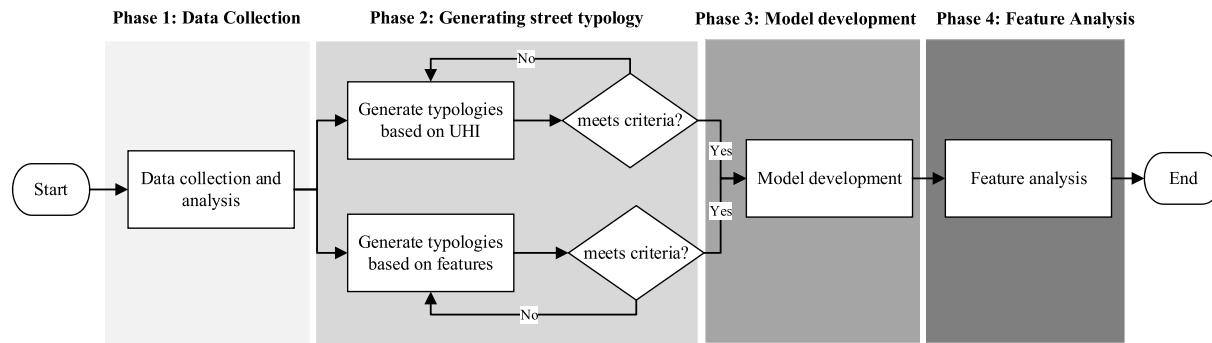


Fig. 1. Illustrates the research methodology employed in this study. The process initiates with data collection and analysis, followed by the creation of street typologies using two distinct clustering techniques: one based on inherent street features and the other on temperature profiles. The methodology concludes with the ML model development and a comprehensive feature analysis to study the resultant street typologies.

3.1. Data

The data used in this research were explained extensively in the previous work of the authors (Pena Acosta et al., 2023b). However, for completeness, a brief explanation is provided here as well. The city of Apeldoorn, the Netherlands, is the 11th largest municipality in the country with a population of 165,611 as of 2022 and an area of 341.2 km². With its moderate oceanic climate, Apeldoorn presents a unique combination of urban and natural features, making it an interesting location to study UHI. This is because not all cities in the Netherlands experience the same intensity of UHI. For instance, older cities with a high proportion of brick buildings retain more heat than newer cities with more energy-efficient buildings. Similarly, cities with more green spaces, and lower populations experienced lower intensities of UHI (Van Hove et al., 2011). The dataset collected by the authors consisted of two main parts: publicly available cadastral datasets (PDOK, 2023) and time-dependent environmental parameters. The cadastral datasets capture the characteristics of the built environment, and socio-economic factors such as the types of buildings, the land use of different areas within the city, and the population density. As summarized in [Table 1](#),

Table 1

Summary statistics of all the features categorized by type, and category, showing the mean and standard deviation values for the whole dataset.

Category	Type	Feature	Mean	Stdev
Urban morphology	Street characteristics	Width	9,77	3,47
		H/W Orientation	0,97 N/A	0,82
	Street use in terms of %	Bicycle	0,2	0,34
		Vehicle	0,78	0,36
		Pedestrian	0,02	0,11
	Building characteristics	Average height	7,87	4,23
		Maximum height	5,8	4,6
	Densities	StdDev height	1,2	1,62
		Building	0,13	0,12
		Vegetation	0,17	0,18
Socio-economic	Street characteristics	Water bodies	0,01	0,03
		Material	N/A	
	Social	Surface color		
		Land use		
		Average population	57,67	33,39
Environmental	Environmental Temperatures	RH	68,25	16,92
		Ref Air	14,31	6,82
		Ref Surface	12,2	10,27
		Air	14,46	6,93
	UHI	Surface	11,74	10,26
		Canopy	0,15	0,9
		Surface	-0,46	2,97

the urban morphology of the streets is characterized by attributes such as street width, height-to-width ratio (H/W), and street use percentages for bicycles, vehicles, and pedestrians. Additionally building characteristics, including average height, maximum height, and standard deviation of height. Furthermore, the data on building densities for buildings, vegetation, and water bodies. Regarding the socio-economic parameters, features such as street materials, surface colors, land use, and population was used. For instance, regarding orientation of streets, data was collected on whether streets were oriented East, North, Northeast, South, or Southeast. Additionally, the type of street materials used, such as Hot Mix Asphalt (HMA), Concrete tiles, Concrete pavers, Clay brick, or Anti-skid surfacing, were included. Surface colors of the streets were also documented, with options including Black, Gray, Red, Red coating, and Yellow-gray.

The time-dependent environmental parameters were captured by a mobile unit built by the authors to collect geo-referenced and timestamped air and surface temperature data. As shown in [Fig. 2](#), the mobile unit was equipped with a sensor kit, including a GPS rover, thermologger, display, thermal camera, and processing responsible for

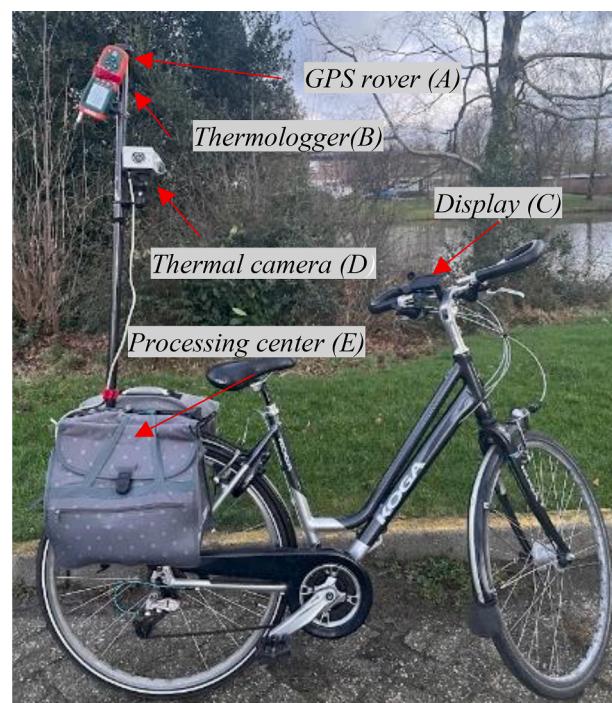


Fig. 2. Bicycle-based mobile urban data-gathering station. Adapted from Pena Acosta et al. (2023).

storing the data. The measurements were taken every second with a constant cycling speed of 8 km/h, meaning data was collected at a 2 m spatial resolution. To compute the average street temperatures a two-phase process was employed: initially, for each street segment, data points falling outside the typical range (those below the 10th or above the 90th percentile) were classified as outliers and omitted. The second phase involved averaging the temperatures from the remaining data points, thus calculating the mean street temperature. The data was collected with a frequency of three measurements per day at 5:30 UTC, 10:30 UTC, and 16:30 UTC, twice a week between March and September 2021, and once per week between October 2021 and February 2022. This frequency allowed for a detailed examination of the evolution of diurnal patterns, from morning to afternoon to early evening, and for exploring the distinct relationships between urban features and varying time periods throughout the day. The 8 km-long route was composed of 105 streets with distinctive urban morphology and socioeconomic parameters. The data campaign resulted in a total of 137,325 measurements.

Furthermore, as presented in the previous work of the authors (Pena Acosta et al., 2022) all features were scaled down to a street-level resolution. To achieve this, each street was delineated by the stretch of road between two intersections. As shown in Fig. 3, a 15 m buffer zone was established around the center of the street section. This buffer zone is considered the street jurisdiction, within which all surrounding features were compiled and analyzed by calculating the proportion of the buffer area that coincides with each specific feature. The 15-m buffer zone was employed to account for the different street layouts, ranging from the city center with minimal setback to residential areas with more prominent front spaces.

Additionally, the authors calculated CUHI and SUHI by selecting a temperature reference location from the data collection path to ensure that the reference point had the same spatial and temporal resolution. Fig. 4 shows an example of the data collected by the authors. Moreover, Fig. 5 presents a visualization of the variation in both Canopy and Surface UHI over the course of the data collection campaign. The figure

is designed in the form of a grid (heatmap) which provides a color-coded representation of temperature values. On the vertical axis (y-axis) of this heatmap, the streets that were part of the study are listed. There are a total of 105 streets, and each one has its own dedicated row on the grid. On the horizontal axis (x-axis), the months during which the data was collected are delineated, starting from March 2021, and concluding in March 2022. Each individual cell or grid on this heatmap corresponds to a specific street (from the y-axis) during a particular month (from the x-axis).

3.2. Street typology

The purpose of the street typology analysis is to identify groups of streets that share similar UHI profiles. This typology was created, as discussed previously, by taking two different approaches. The first approach looked at the configuration of the streets, which are the urban morphology and socio-economic factors. Then, it identified similarities among these features and grouped them into distinct sets. This was achieved through the use of the K-Prototypes algorithm. Conversely, the second method examined temperature fluctuations, represented by the labels, i.e., CUHI and SUHI. This second approach was conducted by implementing the K-Means Time Series. Hence, feature-based clustering relies on a wide range of morphological and socio-economic features to classify streets. It aims to capture the diversity of streets based on measurable attributes such as street width, building height, population density, and vehicle usage. By considering these features, it is intended to create typologies that reflect the physical and social characteristics of streets. However, it's important to note that streets with similar feature profiles may still exhibit variations in UHI patterns due to localized environmental factors. In contrast, UHI-based clustering focuses on the temporal patterns of UHI over time. It seeks to categorize streets based on their thermal characteristics, emphasizing the temperature differences between streets during specific time windows. This approach aims to capture the variations in heat patterns that may not be fully explained by morphological and socio-economic features alone. Thus, while

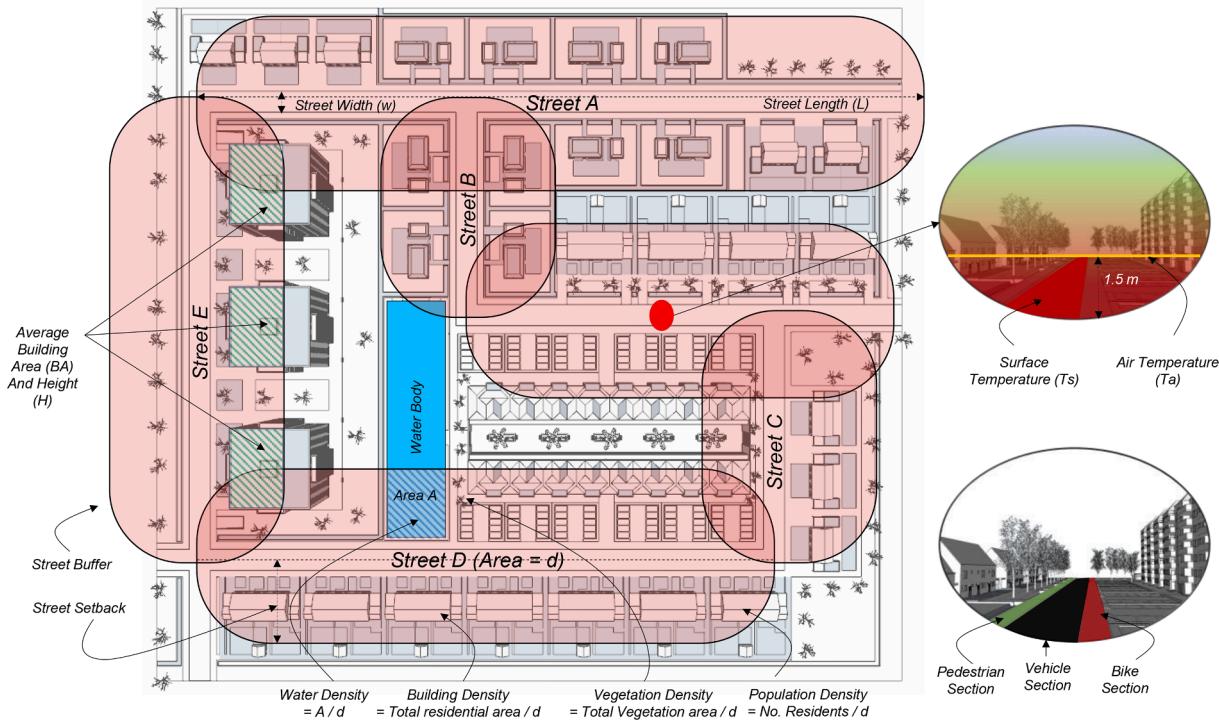


Fig. 3. Graphical representations of the buffer area where all urban methodology and socio-economic features were calculated Adapted from Pena Acosta et al. (2023).

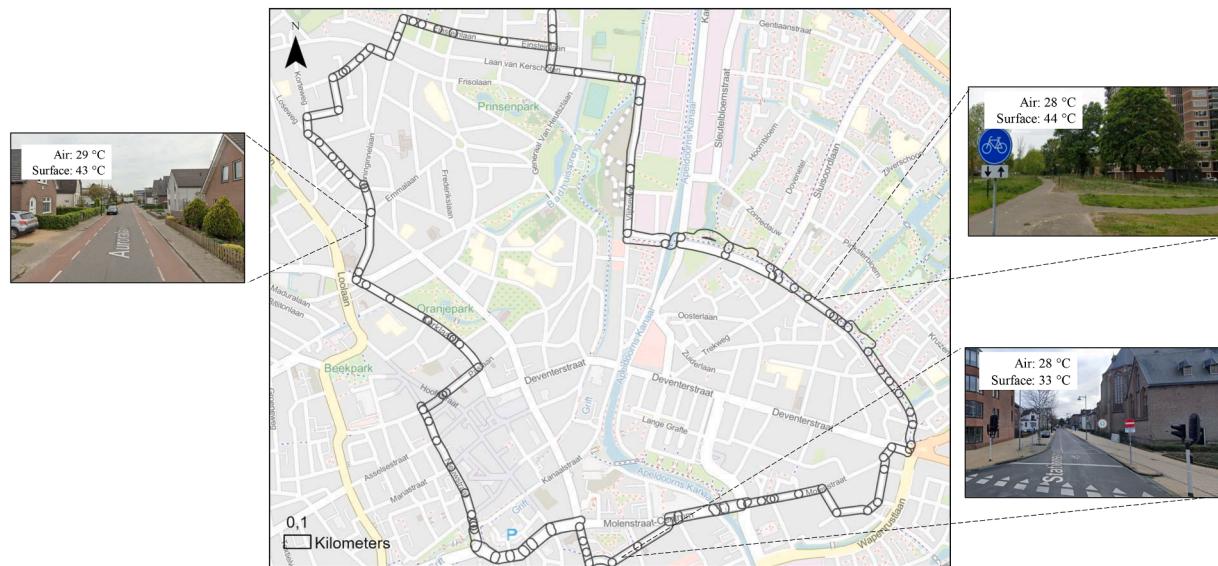


Fig. 4. Examples of different type of streets from which the clusters are built and analyzed. The air and surface temperatures were collected on the 14 of June 2021.

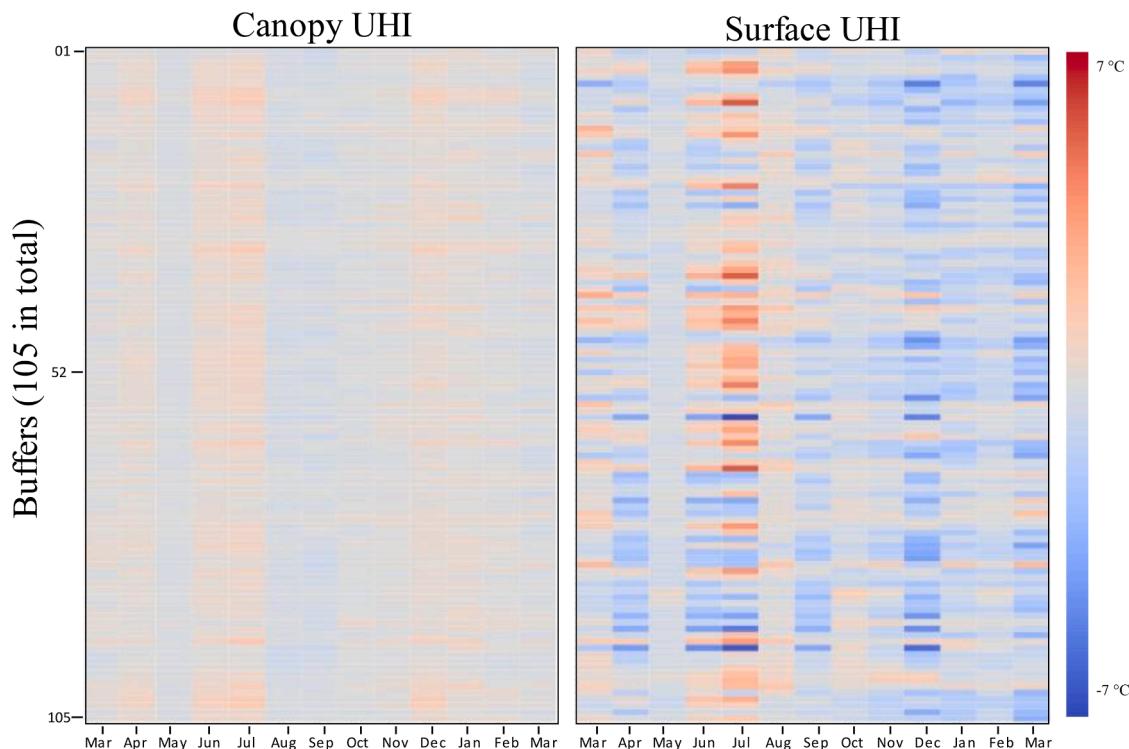


Fig. 5. Comparative heat map of monthly air and surface UHI across urban streets: Each vertical grid represents a specific street, and the temporal progression is captured on the X-axis, spanning a year. Notably, not all streets exhibit uniform UHI patterns. Some streets display pronounced UHI variations, while others remain more consistent.

feature-based clustering accounts for physical and social attributes, UHI-based clustering considers the dynamic nature of temperature variations.

For both approaches, two main steps were taken. First, each numerical feature had to be normalized. This is because the algorithms calculate the distances between data instances. In the case of the K-Prototypes algorithm, these distances were calculated by a combined measure that incorporates both numerical and categorical features. For numerical features, squared Euclidean distance was used (Yuan & Yang, 2019), while for categorical features, the Hamming distance was

employed (Ahmad & Dey, 2007), which is the measure of the minimum number of substitutions needed to change one string into the other.

For the K-Means Time Series algorithm, the distances are calculated by employing a dynamic time warping (DTW) distance (Aghabozorgi et al., 2015), which allows to compare and align the time series data. It is important to keep in mind that in the context of this research, the time series data are temperature measurements that represent values over time, with one sequence corresponding to CUHI and the other SUHI. These two sequences were used as input for the clustering. Table 2 shows an example of the structure of the data in terms of the streets. The values

Table 2

Example structure of the dataset for the UHI-based clustering approach: The values represent the average of three daily measurements for CUHI and SUHI, in degrees Celsius (°C).

Street ID	22-Mar-21		24-Jun-21		11-Oct-21		6-Dec-21		5-Feb-22	
	CUHI	SUHI	CUHI	SUHI	CUHI	SUHI	CUHI	SUHI	CUHI	SUHI
112,380,812	0,257	0,211	...	0,538	-1841	...	0,051	-0,073	...	-0,523
116,146,981	0,297	-0,029	...	0,744	-0,464	...	0,120	-0,693	...	-0,392
116,147,138	0,245	-0,127	...	1104	1302	...	-0,263	-0,532	...	-0,561
.
.
.
116,147,315	0,260	0,407	...	1000	1820	...	-0,294	-0,205	...	-0,569
116,147,709	0,253	-0,083	...	0,945	-0,042	...	-0,023	-1312	...	-0,201
116,147,757	0,176	-0,498	...	0,632	-0,098	...	-0,419	-2227	...	-0,303
116,147,896	0,400	-0,280	...	0,947	-1910	...	0,342	-0,708	...	-0,439
116,148,586	0,479	-0,404	...	1263	0,714	...	0,071	-1109	...	-0,375
116,148,799	0,540	0,229	...	1192	2586	...	-0,060	-2058	...	-1086
116,149,031	0,288	0,060	...	0,726	-1343	...	-0,433	-1224	...	-1100

presented in the table represent the mean of three measurements taken per day for CUHI and SUHI. The choice to average the values was made to maintain a consistent approach to data handling across the dataset. Once the data was normalized, the second step involves determining the ideal number of clusters. Since this number is specific to the dataset,

there is no universal method for identifying it (Halkidi et al., 2001). In this research, as shown in Fig. 6, the dataset was divided into clusters ($k = 2$ to 12), and each cluster's performance was assessed using two metrics: R^2 and MAE. The number of clusters that achieves the best performance, characterized by the lowest MAE and the highest R^2 , was selected as the final number of clusters for further analysis. This approach allowed for the identification of the optimal value of k for clustering. Once the number of clusters was determined, the entire dataset was divided into each group of typologies for model development and feature analysis.

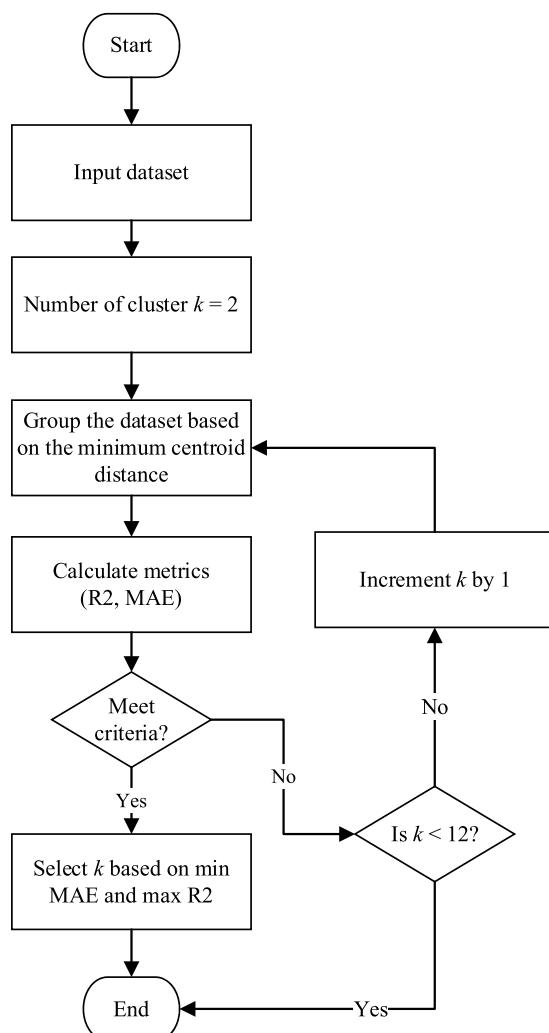


Fig. 6. Flowchart illustrating the clustering optimization process: Iterating through cluster numbers (2–12), assessing performance with R^2 and MAE metrics, and determining the optimal cluster count based on these evaluative criteria.

3.3. Model development

Machine learning algorithms have gained momentum in the field of UHI research. One of the reasons for this is their ability to process vast amounts of data and uncover hidden patterns that may not be readily noticeable. Random Forest (RF) algorithm in particular has proven to be very effective in UHI research (Kim & Brown, 2021) because of its ability to handle non-linear relationships and its robustness to overfitting. The details of how a data-driven model is developed for UHI data are extensively explained in the previous work of the authors (Pena Acosta et al., 2021). For each group of streets (i.e., street typologies) identified in Section 3.2, a RF model was trained in order to predict CUHI, and SUHI. The goal here was to design highly accurate models. To evaluate the performance of the models, 30 % of the data set was set aside for testing. The model's predictions were compared to these data by calculating the MAE and R^2 values for the predicted versus actual UHI value (i.e., the temperature differential). During the optimization of the models, various hyperparameters were adjusted to prevent overfitting and underfitting and to identify the highest-performing models.

3.4. Feature analysis

To evaluate the feature importance for a given model several, approaches can be taken. For instance, RF regression models provide feature importance analysis by measuring the average decrease in variance due to splits on a particular feature. This gives an idea of which features are more useful to the model for making predictions, but it does not consider the interactions between features, and it can be sensitive to the scale of the features. As an alternative, Permutation Importance is a model-agnostic method that works by randomly shuffling the values of a specific feature and then measuring the change in the model's performance. The overall idea behind it is that if a feature is important for the model, then shuffling its values should have a significant impact on the performance of the model while shuffling the values of a less important feature should have a smaller impact. The feature importance score is calculated as the decrease in the model performance metric (e.g., MAE, R^2) due to the feature permutation. The output of the feature analysis,

therefore, is specific to the performance of the model. This is because it only measures the main effect of each feature and does not consider interactions between features. On the other hand, SHAP feature importance is based on the concept of Shapley values, which provide a way to fairly distribute a value among a group of individuals. In the case of SHAP feature importance, the value being distributed is the model's output, and the individuals are the features. This method considers both the main effect and the interactions between features, and it provides a better picture of how each feature contributes to the model's output. In this study, SHAP feature importance was utilized to investigate the relationship between features and various types of UHI based on the clusters. In order to improve the accuracy of the results and account for the stochastic nature of the models, the process involved calculating the SHAP values for each RF model in thirty runs. Only the average of these results was considered in the analysis. Subsequently, the average value of feature importance was ranked and plotted for further analysis.

4. Results

4.1. Street typology

As described in [Section 3.2](#), this paper took two different approaches for the classification of streets (1) Feature-based clustering examines the similarities among a wide range of morphological and socio-economic elements, and (2) UHI-based clustering analyses the patterns of UHI over time and classifies streets based on their temperature differential patterns. The results showed that, as shown in [Table 3](#), Three clusters (i.e., $k = 3$) yielded the most effective classification for both SUHI and CUHI. [Table 4](#) summarizes the mean and standard deviation for each feature within every cluster. The average values for CUHI and SUHI should be interpreted as an overarching glimpse of the dataset, rather than as detailed representations of the behaviors and mechanisms within each cluster. The interplay of each socio-economic and morphological feature is examined in [Sections 4.1.1](#) and [2](#). The negative values of SUHI is due to the selection of the reference point. As mentioned in the earlier work of the authors ([Pena Acosta et al., 2023b](#)), the reference point was selected from one of the coolest point on the data collection route to ensure resolution consistency. However, as explained in the earlier work, the choice of the reference point does not significantly impact the identified mechanism of the UHI by ML. [Fig. 7](#) visualizes the time series patterns resulting from the feature-based clustering approach. The x-axis represents the time window, while the y-axis displays the ranges for both types of UHI. Likewise [Fig. 8](#) visualizes the time series patterns but this time for the UHI-based clustering approach. Furthermore, [Fig. 9](#) presents the summary of the three clusters per method. [Fig. 10](#) presents Kernel Density Estimates (KDE) of the CUHI, and SUHI, respectively, resulting from the two clustering approaches. Each subplot in the figures corresponds to a specific cluster label (1, 2, or 3) and displays the KDE of the CUHI/SUHI distributions for each clustering method. The y-axis represents the probability density, while the x-axis presents the ranges of either CUHI, or SUHI. In addition, each plot displays the Kolmogorov-Smirnov (KS) statistic values calculated for each pair of clusters, in order to compare the similarities between the distributions of the target variables (i.e., CUHI and SUHI).

For CUHI, Cluster 01, and 02 have the smallest KS value, indicating a high level of similarity in the CUHI distributions between the clusters resulting from the two methods. Conversely, Cluster 02 has the highest

KS value, suggesting that its CUHI distribution shows some differences from both Cluster 01 and Cluster 03. For the SUHI, the clusters show similar differences, with Cluster 02 having a KS value of 0.08, indicating that its SUHI distribution shows more differences from both Cluster 01 (KS value of 0.03) and Cluster 03 (KS value of 0.02). Clusters 01 shared 13 streets, Clusters 02 shared 16 streets, and Clusters 03 shared 24 streets.

4.1.1. Feature-based clustering

1. **Cluster One - Wide Airflow Streets with High Population:** This cluster is distinctive for its broader streets, averaging 10,29 m in width, and a higher H/W ratio, with an average of 1,12. The literature suggests that the combination of these features contributes significantly to improving airflow ([Ali-Toudert & Mayer, 2007](#); [Rizwan et al., 2008](#)). In terms of socio-economic attributes, it has a higher average population compared to the other clusters.
2. **Cluster Two - Vehicle-Dominant Streets with Lower Population:** The most notable characteristic of this cluster is the high predominance of vehicle-oriented roads, with vehicle use averaging at 0,93 %. On the other hand, Cluster One, has only an average vehicle usage of 0,15 %. However, this cluster is less densely populated, with an average population of 43,59.
3. **Cluster Three - High-Rise Streets with Intense Vehicle Use and Slightly Higher Temperatures:** This cluster stands out due to its higher buildings, with the highest average height of 9,68 m among all clusters. It also records the highest level of vehicle use, averaging 0,95 %. Additionally, it has slightly higher surface temperatures and CUHI when compared to the other clusters.

4.1.2. UHI-based clustering

Three clusters were also obtained when applying time series clustering to both SUHI and CUHI. Similar to feature-based clustering, Cluster One consists of 26 streets, Cluster Two consists of 27 streets, and Cluster Three consists of 52 streets. [Fig. 11 \(b\)](#) summarizes the spatial distributions of the clusters. Based on their socioeconomic and morphological features, the resulting clusters can be described as:

1. **Cluster One - The Bicycle-Friendly Narrow Streets:** This cluster has the narrowest average street width of 8,65 m, and the highest H/W ratio of 1,33. This cluster is also characterized by the highest bicycle usage (0,48 %) among the three clusters. The average building height is relatively moderate, and this cluster also shows the highest average vegetation density.
2. **Cluster Two - The Vehicular Wide Streets:** Streets in this cluster are wider on average (10,14 m) with a lower H/W ratio of 0,78, possibly indicating wider, less enclosed streets. Vehicle usage is high in this cluster (0,79 %) and it has the highest average intensity SUHI, indicating potential heat-related problems.
3. **Cluster Three - The Vehicle-Dominated Streets with Tall Buildings:** Similar to Cluster Two, streets in this cluster are also wide (average of 10,14 m) but with a slightly higher H/W ratio of 0,89. This cluster shows the highest vehicle usage (0,93 %) and the highest average building height (8,27 m). The vegetation density is comparable to Cluster Two, but lower than Cluster One.

Table 3

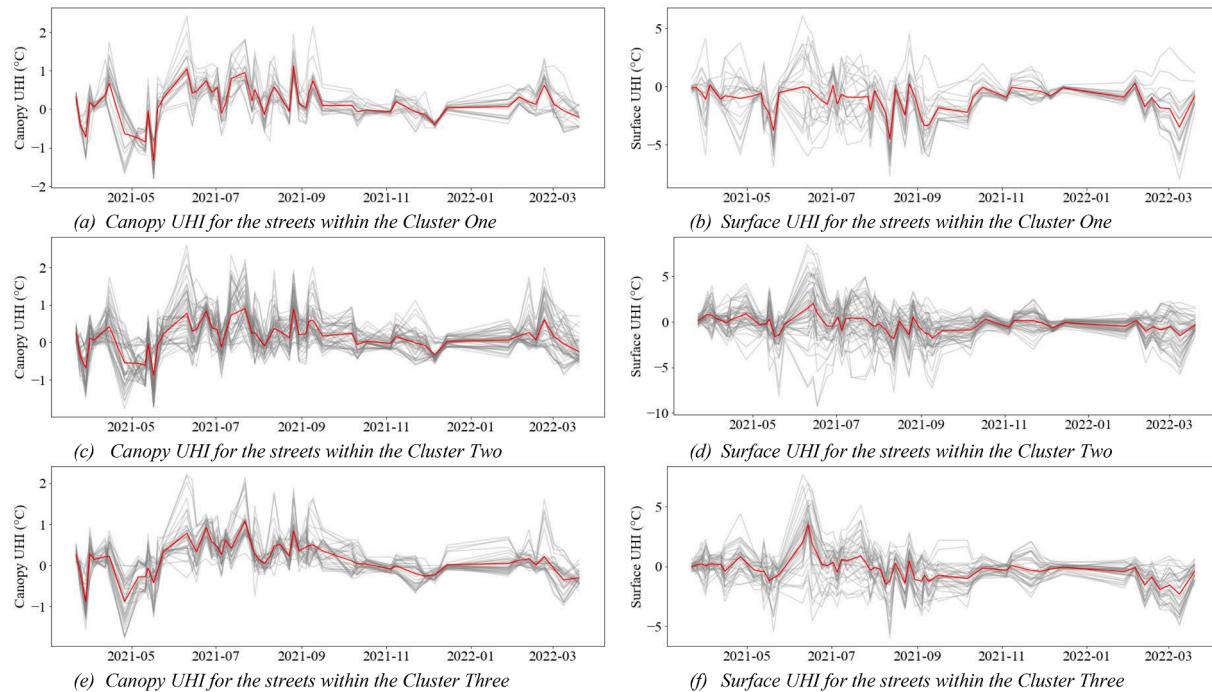
Comparison of model performances using R^2 and MAE metrics evaluated for different numbers of clusters (k) based on Feature-based and UHI-based approaches.

Num of clusters k	1	2	3	4	5	6	7	8	9	10	11	12
Feature-based	R-Squared	0,83	0,85	0,87	0,79	0,79	0,78	0,71	0,71	0,73	0,66	0,65
	MAE (°C)	0,22	0,22	0,20	0,25	0,26	0,26	0,28	0,27	0,29	0,31	0,32
UHI-based	R-Squared	0,83	0,84	0,84	0,79	0,81	0,80	0,82	0,78	0,78	0,76	0,77
	MAE (°C)	0,22	0,23	0,23	0,23	0,24	0,23	0,25	0,24	0,27	0,29	0,28

Table 4

Summary of the average and standard deviation values of all the features across the different clusters resulting from both the feature-based and UHI-based approaches.

Category	Type	Feature	Feature-based						UHI-based					
			Cluster One		Cluster Two		Cluster Three		Cluster One		Cluster Two		Cluster Three	
			21 streets	Mean	Stdev	49 streets	Mean	Stdev	35 streets	Mean	Stdev	26 streets	Mean	Stdev
Urban morphology	Street characteristics	Width	10,29	6,07	10,25	2,43	8,79	2,09	8,65	4,38	10,14	3,59	10,14	2,69
		H/W Orientation	1,12	0,93	0,7	0,51	1,26	0,96	1,33	1,31	0,78	0,46	0,89	0,54
	Street use in terms of%	Bicycle	N/A	0,84	0,16	0,07	0,14	0	0	0,48	0,43	0,21	0,3	0,06
		Vehicle	0,15	0,17	0,93	0,15	0,95	0,18	0,47	0,43	0,79	0,3	0,93	0,22
		Pedestrian	0,02	0,05	0	0,02	0,04	0,18	0,05	0,17	0	0	0,02	0,09
		Average height	8,4	3,82	6,35	4,28	9,68	3,54	7,72	4,42	7,23	3,68	8,27	4,35
	Building characteristics	Maximum height	8,7	5,03	2,85	2,8	8,19	3,78	5,85	3,8	5,82	4,89	5,76	4,8
		StdDev height	1,41	1,68	0,93	1,83	1,46	1,14	1,36	2,36	1,34	1,65	1,05	1,03
		Building	0,16	0,1	0,05	0,06	0,23	0,11	0,13	0,11	0,11	0,08	0,14	0,13
		Vegetation	0,24	0,18	0,24	0,17	0,02	0,04	0,26	0,23	0,13	0,13	0,14	0,15
	Densities	Water bodies	0,01	0,04	0,01	0,03	0	0	0	0,01	0,01	0,04	0	0,03
		Material	N/A											
Socio-economic	Street characteristics	Surface color												
		Land use												
	Social	Average population	73,25	32,83	43,59	29,69	68,04	30,55	55,41	34,31	51,8	31,8	61,85	33,17
Environmental	Environmental Temperatures	RH	68,51	16,94	68,42	16,83	67,86	17,02	68,75	16,85	67,87	17,12	68,2	16,84
		Ref Air	14,31	6,82	14,31	6,82	14,31	6,82	14,31	6,82	14,31	6,82	14,31	6,82
		Ref Surface	12,2	10,27	12,2	10,27	12,2	10,27	12,2	10,27	12,2	10,27	12,2	10,27
		Air	14,46	6,9	14,45	6,9	14,47	7	14,4	6,83	14,65	7,04	14,39	6,93
	UHI	Surface	11,01	9,77	11,91	10,22	11,94	10,58	10,74	9,04	12,42	11,2	11,88	10,29
		CUHI	0,14	1,02	0,14	0,94	0,16	0,75	0,09	1,02	0,33	1,09	0,08	0,69
		SUHI	-1,19	3,74	-0,29	2,87	-0,26	2,49	-1,46	3,75	0,22	3,01	-0,32	2,32

**Fig. 7.** Time Series representation of Canopy, and Surface UHI for the clusters resulting from the feature-based clustering. The red line indicates the overall average UHI across all streets for each cluster.

4.2. Model performances

As discussed in Section 3.3, RF models were trained individually for each cluster. The optimal hyperparameters for these models were identified using a random grid search optimization method. The final

configuration is as follows: `n_estimators = 150, min_samples_split = 2, min_samples_leaf = 1, max_depth = 20, bootstrap = True, and random_state = 42`. Table 5 provides a comprehensive summary of the models' performance across the seven datasets. For the complete dataset, the RF achieved an R^2 of 0,83, indicating that approximately 83 % of

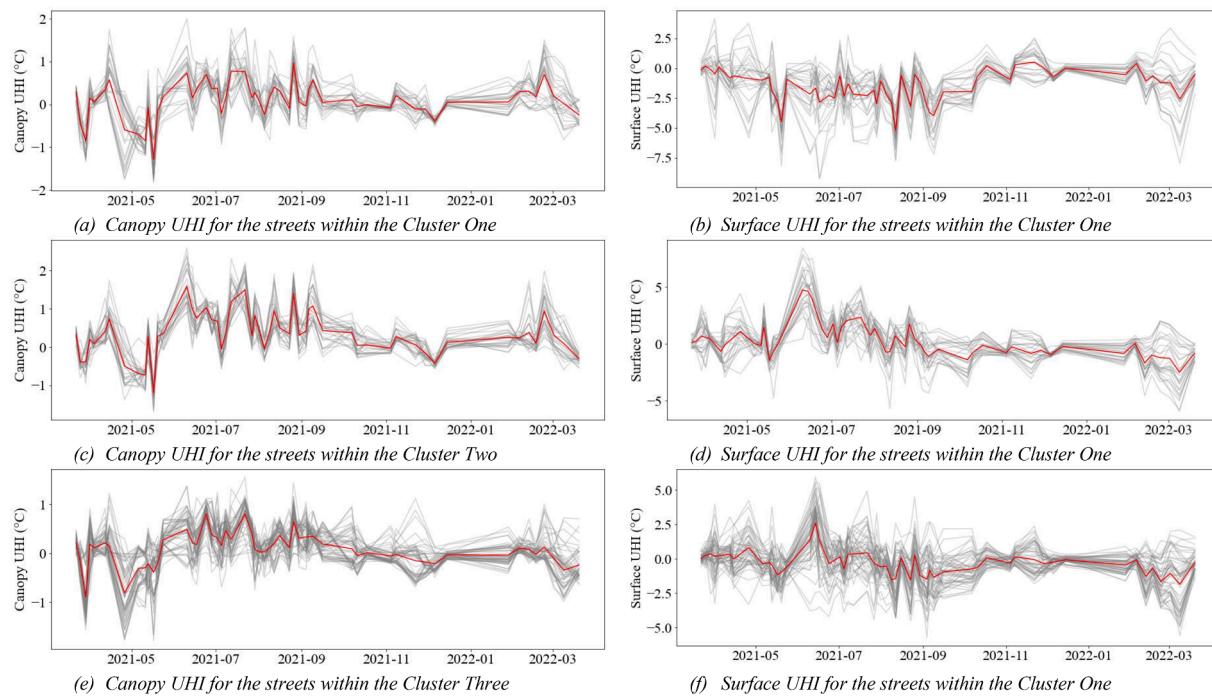


Fig. 8. Time Series representation of Canopy, and Surface UHI for the clusters resulting from the UHI-based clustering. The red line indicates the overall average UHI across all streets for each cluster.

the variance in CUHI can be explained by the features, with a MAE of 0,23, indicating an average absolute difference of 0,23 between the predicted CUHI and actual values of CUHI. Similarly, for SUHI, the model was able to achieve a performance of 0,80, indicating that approximately 80 % of the variance in SUHI can be explained by the features, with a slightly higher MAE of 0,87, which indicates an average absolute difference of 0,87 between the predicted SUHI and actual SUHI values. Overall, the model is able to predict CUHI better.

Regarding the feature-based clustering, Cluster One achieved higher R^2 and lower MAE values for both CUHI and SUHI compared to the complete dataset (0,89 for CUHI, and 0,83 for SUHI). This suggests that the dataset in Cluster One is more effective in predicting UHI compared to the entire dataset. On the other hand, Cluster Two shows slightly lower R^2 values (0,86 for CUHI, and 0,74 for SUHI) and higher MAE (0,30 for CUHI, and 1,33 for SUHI), indicating that this cluster may not capture UHI variations as well as Cluster One. As for Cluster Three, the model performs well with relatively high R^2 (0,87 for CUHI, and 0,82 for SUHI) and low MAE (0,17 for CUHI, and 0,74 for SUHI) similar to Cluster One. However, the MAE in Cluster Three is significantly worse, increasing from 0,49 to 0,74. Although the overall prediction error is still small, it is interesting to note that the MAE worsened by almost 50 %, while the R^2 value remained more or less the same.

For the UHI-based clustering (i.e., CUHI and SUHI), Cluster One performs the best with an R^2 value of 0,89 for CUHI and 0,87 for SUHI. Furthermore, Cluster One has relatively low MAE values of 0,22 for CUHI and 0,85 for SUHI, indicating a small average absolute difference between predicted and actual values. Cluster Two performs slightly lower with R^2 values of 0,85 for CUHI and 0,82 for SUHI, and slightly higher MAE values of 0,28 for CUHI and 0,85 for SUHI. Cluster Three shows a further decrease in performance with R^2 values of 0,77 for CUHI and 0,72 for SUHI, and MAE values of 0,20 for CUHI and 0,80 for SUHI.

These results suggest that the typologies may have varying levels of predictability for CUHI and SUHI. Cluster One consistently performs better in terms of predictive accuracy for both the clusters based on the features and the clusters based on the labels, while Cluster Three tends to have relatively lower performance than Cluster One.

In general, it can be observed that CUHI models perform better than

SUHI, suggesting that the UHI mechanism at the surface layer is more complex. Also, while there is not a significant difference between how streets are clustered, it seems feature-based clustering is slightly better. This is an interesting observation because it suggests that classifying streets with respect to their socio-economic and morphological features can explain UHI variation effectively.

4.3. Feature analysis

To investigate the behavior and uncover hidden patterns in the interplay between the features and UHI, SHAP values were implemented. Six datasets consisting of three clusters per type of clustering method (i.e., Feature-based, and UHI-based) were analyzed. The process involved calculating the SHAP values for each RF model 30 times to account for the stochastic nature of the models. Only the average of these results was considered. Fig. 12 graphically represents the SHAP values of the features in the CUHI and SUHI models using parallel plots. In these plots, the importance of each feature is displayed on the vertical axis, while the connected lines spanning across the clusters represent the corresponding values for each of these features.

4.3.1. Feature-based clustering

As shown in Fig. 12(a) and (b), across all three clusters for both CUHI and SUHI, the features representing the moment of the day ("part of the day"), relative humidity, and the temperature at the reference locations, were the most influential features in determining the prediction of the models. It can be interpreted as these features are significant determinants of the UHI at the street level. However, when looking at the morphological, and socio-economic factors, in the CUHI models, the feature representing the width of the road ("road width") has a more influential role in Cluster One compared to its importance in Clusters Two and Three. When looking at the SUHI model, Cluster Three presents high SHAP values for average building height, and maximum building height per street (0,21 and 0,05, respectively), indicating their importance in these clusters. Conversely, for Clusters One and Two in the SUHI model, the feature representing the percentage of the street dedicated to bicycle use ("bicycle share") has more influence, with SHAP values of

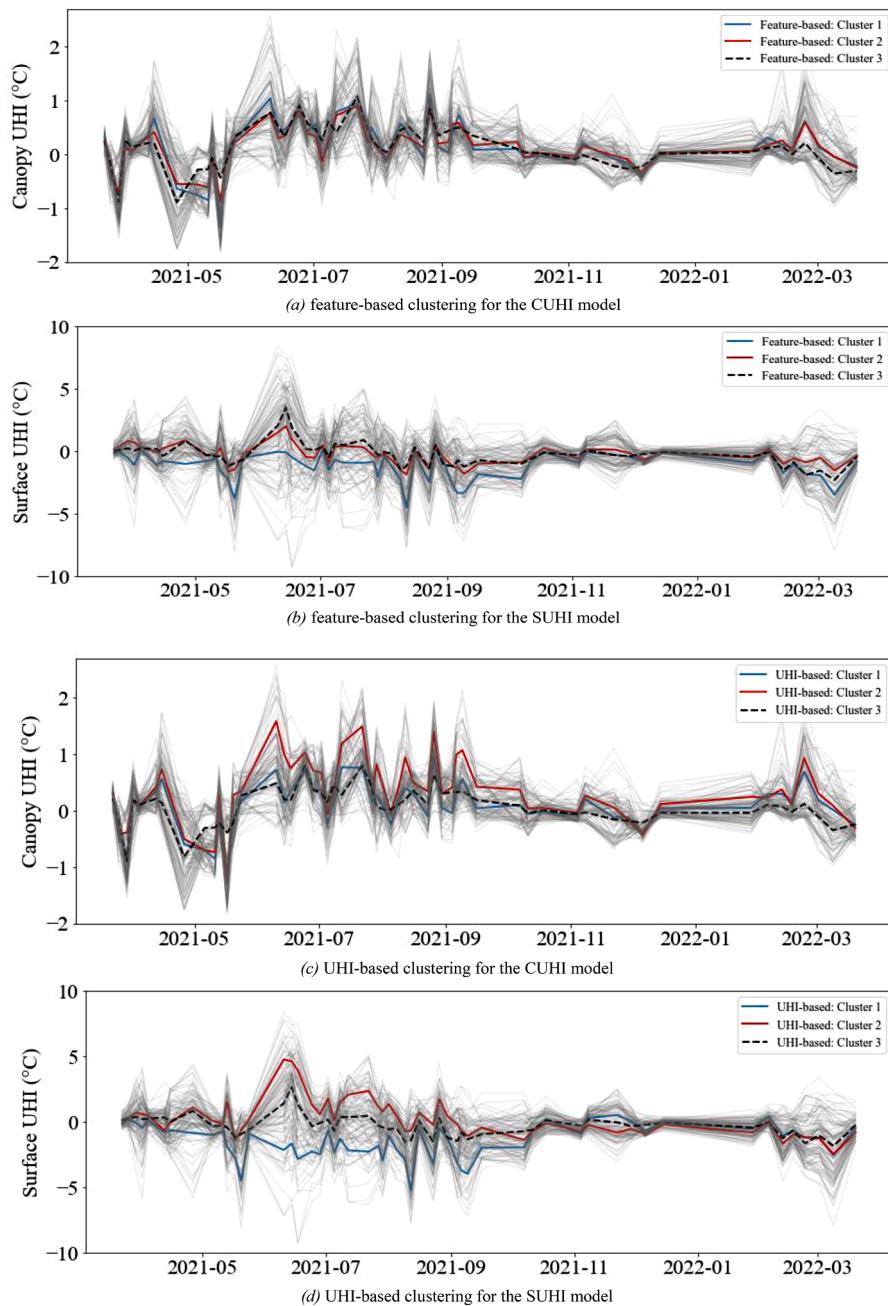


Fig. 9. Time Series plots of Canopy and Surface UHI, comparing the three clusters on the same plot by methodological approach.

0,10 and 0,08, respectively.

4.3.2. UHI-based clustering

A similar pattern emerged when evaluating the clusters derived from the UHI-based clustering method, which highlights the importance of the time of the day, as shown in Fig. 12 (c) and (d). However, for the CUHI model, the most influential feature is the temperature at the reference location, followed by the average surface temperature, and the relative humidity. Among all clusters, the top five features consist of environmental parameters.

When looking at the contribution of urban morphological, and socio-economic parameters for the CUHI clusters, the average building height stands out as a significant factor in Cluster One, with a SHAP value of 0,09. This highlights its role in shaping CUHI in that particular street typology. Conversely, Clusters Two and Three display lower SHAP values for average building height, indicating a comparatively smaller

influence on CUHI in these typologies. The factor of road surface color has a fairly high SHAP value of 0,06 in Cluster Two. However, its SHAP values are lower in Clusters One and Three. In relation to the SUHI models, the geometry of the buildings, represented by average and maximum building height, have a high influence among all clusters. Yet for Cluster Two, the feature representing building density appears to play a more influential role, with a SHAP value of 0,24.

4.3.3. Variation in morphological and socio-economic features

The SHAP values revealed considerable variations in urban morphological and socio-economic features across different clusters and types UHI. These features are under the jurisdiction of urban planners (Rizwan et al., 2008) and therefore are more pivotal for mitigation of UHI. For instance, as shown in Fig. 13(a), the typologies resulting from the feature-based clustering for CUHI show that the average building height had the highest degree of variability among all three clusters.

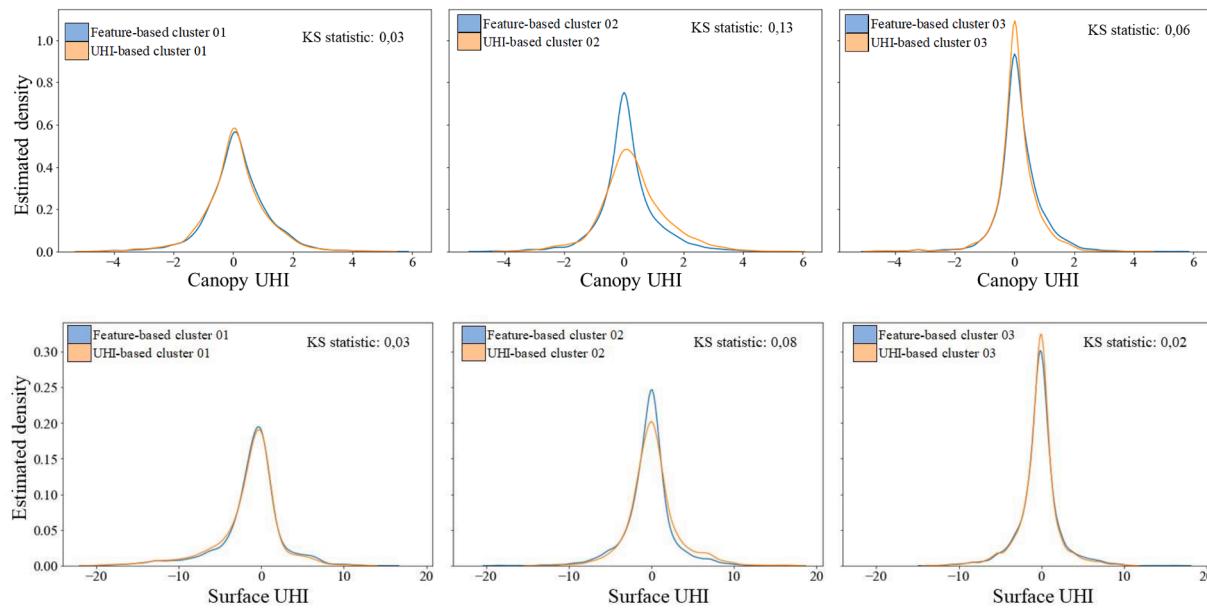


Fig. 10. Comparison of Kernel Density Estimates (KDE) for Canopy UHI (CUHI) and Surface UHI (SUHI) based on the two clustering approaches implementing in this research. Each subplot indicates a specific cluster pairing, showcasing the KDE of either CUHI or SUHI. The probability density is represented on the y-axis, while the respective UHI values are plotted on the x-axis. The KS statistic value for each pair of clusters is also provided.

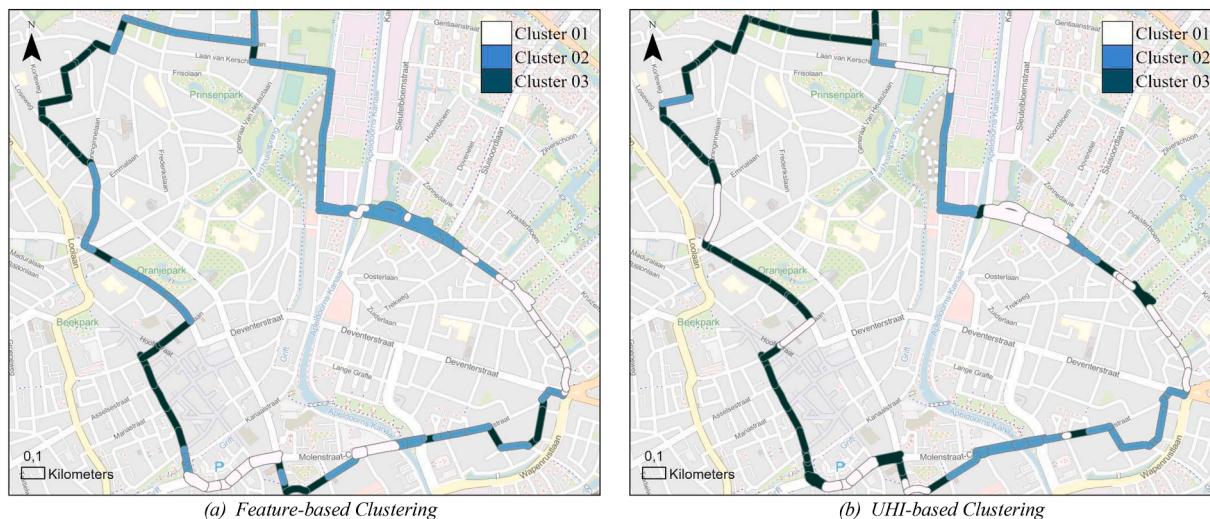


Fig. 11. Spatial distribution of resulting clusters.

Table 5

Model performances for Canopy UHI (CUHI) and Surface UHI (UHI) using R^2 and MAE metrics. The models are evaluated on a complete dataset and the clusters resulting from both the feature-based and UHI-based approaches.

	CUHI R^2	MAE (°C)	SUHI R^2	MAE (°C)
Complete dataset	0,83	0,23	0,8	0,87
Feature-based: Cluster 01	0,89	0,14	0,83	0,49
Feature-based: Cluster 02	0,86	0,3	0,74	1,33
Feature-based: Cluster 03	0,87	0,17	0,82	0,74
UHI-based: Cluster 01	0,89	0,22	0,87	0,85
UHI-based: Cluster 02	0,85	0,28	0,82	0,85
UHI-based: Cluster 03	0,77	0,20	0,72	0,80

Conversely, for SUHI, as shown in Fig. 13 (b), the H/W ratio has the highest fluctuation. Other parameters such as road width and vegetation density also show varied influences across the clusters. For instance,

vegetation density has a larger influence in Cluster One for both CUHI and SUHI. Road surface color, on the other hand, has a low influence on CUHI clusters but moderate importance in the SUHI clusters.

Similar patterns emerge when looking at UHI-based clusters where average building height has the highest variability, as shown in Fig. 13 (c) and (d) (for CUHI, and SUHI respectively). Vegetation density shows a high variability when it comes to CUHI, but consistent importance in the SUHI typologies. Yet, features such as land use, road material, and the street designated use (e.g., vehicle, bicycle, or pedestrian) seems to be consistent among different street types for both CUHI and SUHI.

4.3.4. Representative streets per cluster

As discussed in Section 1, one of the goals of this research was to determine the extent to which street typology can help reduce the need for large-scale data collection. In this regard, it was expected that a small sample of streets from each street type could effectively explain the majority of streets in that cluster. For this reason, different portions of

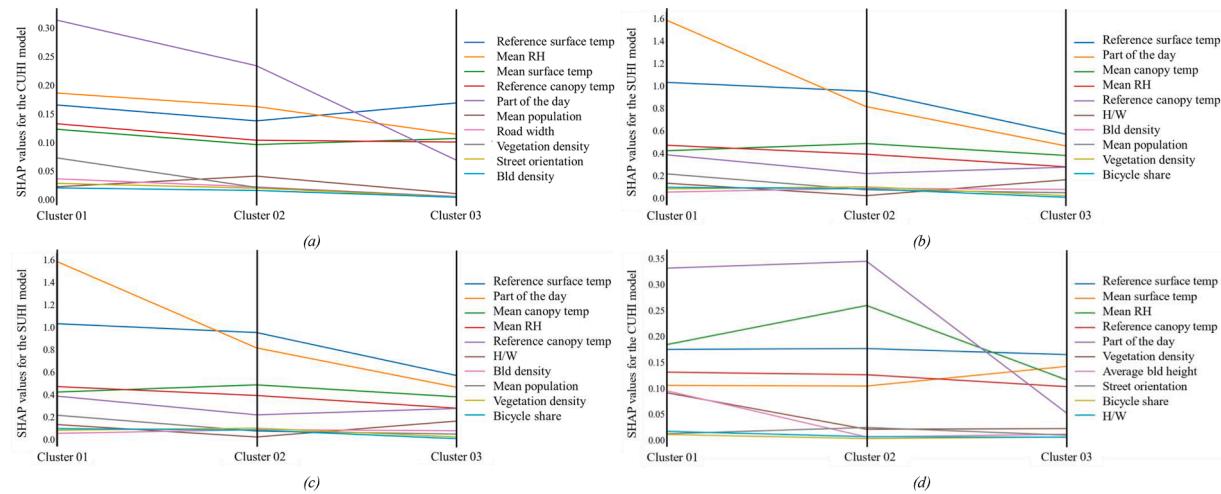


Fig. 12. Parallel plot of (a) feature-based clustering for the CUHI model, (b) feature-based clustering for the SUHI model, (c) UHI-based clustering for the CUHI model, and (d) UHI-based clustering for the SUHI model, where the x-axis represents three different clusters and the y-axis represents the feature importance for each cluster.

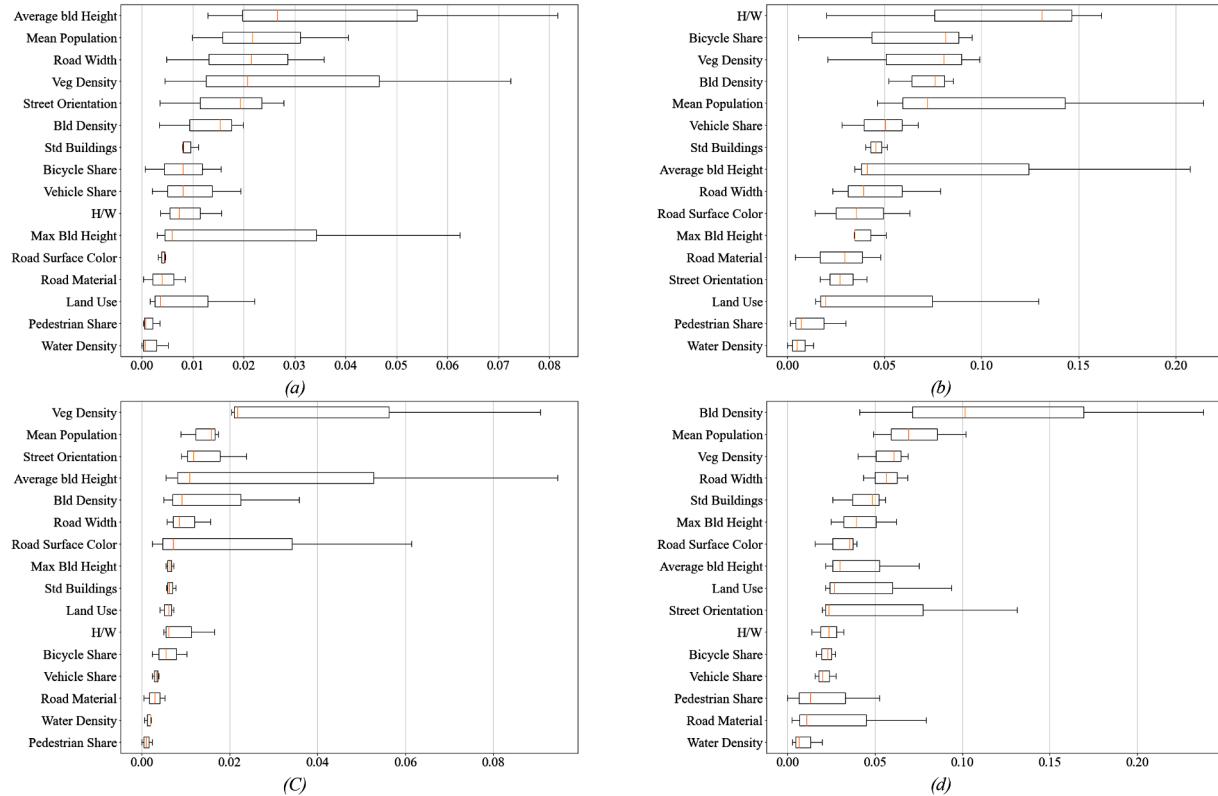


Fig. 13. Boxplots of the distribution of feature importance the urban morphological and socio-economic features among the three clusters resulting from (a) feature-based clustering of CUHI, (b) feature-based clustering of SUHI, (c) UHI-based clustering of CUHI, and (d) UHI-based clustering of SUHI.

streets from each cluster were used to train various models and assess their performances. These portions (e.g., 30 % of streets) were selected randomly. To account for the stochasticity arising from the random selection of streets, each model was trained using 30 iterations, each time with a different random set of streets. Fig. 14 shows the results of this analysis.

To clarify, the street portioning used in this analysis is different from the common data partitioning applied in data-driven modeling. This is because data partitioning of the combined datasets involves selecting a partition of the entire datasets for training, which may include all the

streets but not necessarily all measurements throughout the year. In street portioning, a certain number of streets are completely excluded from the training. For example, if 70 % of the entire dataset is used for training, assuming the data includes 105 streets and 165 measurements per street, there is a very good chance that all streets make it to the training dataset, albeit with only a few measurements. However, if only 70 % of streets are used for training, 30 % of streets (and their associated measurements) are not included in the training.

As shown in Fig. 14, not all street types exhibit similar generalizability trends. For instance, in the case of Cluster One of featured-based

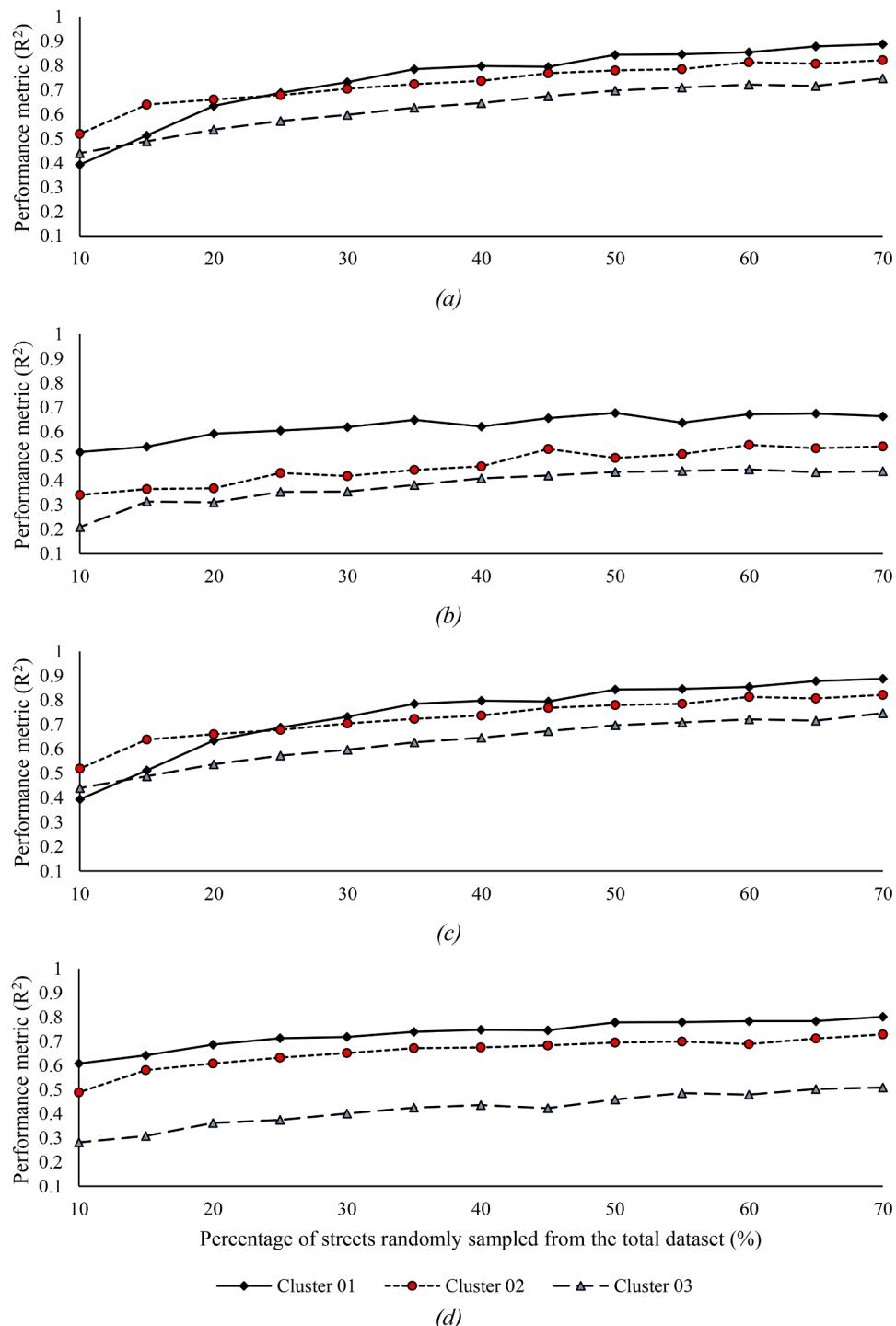


Fig. 14. Model performances of (a) Feature-based CUHI, (b) Feature-based SUHI model, (c) UHI-based CUHI, and (d) UHI-based SUHI using subsets of streets from each cluster.

CUHI, R^2 only improves by 11 % when using 70 % of streets instead of 40 %. On the other hand, Cluster Three has relatively low performance in all cases, indicating that streets with high-rise buildings and a more dominant presence of cars tend to be less generalizable. Nevertheless, all cases demonstrate a flat performance trend concerning the increase in the size of the training street. This demonstrates a relatively high homogeneity of streets in each cluster, in such a way that increasing the number of streets in the training set would not considerably improve performance. It should be noted that in almost all cases, a very small number of streets in the training dataset results in poor performance, but

the model reaches the equilibrium point with approximately 40 % of streets. In almost all cases, CUHI demonstrates better predictability than SUHI.

5. Discussion

The main contribution of this research is to shed light on the extent to which street typologies can be used to predict UHI effectively and potentially be leveraged to coordinate a more efficient data collection campaign for UHI studies. In general, street typologies show promising

potential in effectively capturing the UHI mechanism. Previous research has demonstrated the usefulness of clustering techniques (Liu et al., 2018; Luo et al., 2010). However, little attention has been given to comparing the clustering of features versus clustering the response variable directly (this is CUHI and SUHI). Besides, it indicates that despite the complex mechanism of UHI, there is a certain level of intelligibility to the models because one can expect that similar streets (in terms of socioeconomic and morphological features) demonstrate similar UHI patterns.

While two clustering approaches bundle streets differently, their performance is comparable. The authors interpret this observation as an indication that while each of these clustering can capture part of the complex UHI mechanism none of them supersede the other. This is because the entire premise of the research is that studied features are sufficient to accurately model UHI. However, if there are additional features that are not considered in this research that may have an impact on UHI, they can considerably change the performance of the model. This can be both through unmasking hidden similarities in features that are not captured when they are excluded or through improved performance when streets are clustered based on their temperature profile. Therefore, it can be concluded that within the scope of features considered in this research, and many other similar research, both clustering approaches can perform sufficiently well. Having said that, from the practicality standpoint, feature-based clustering can be more advantageous because urban planners can perform the clustering prior to the collection of the data. This basically indicates that street typologies can be built prior to the collection of high-resolution data based on socioeconomic and morphological features and then use sample street sets per cluster as the target of the data collection.

On the other hand, while ULCZs provides a standardized way of classifying urban areas that can be used to understand the mechanisms of UHI (Demuzere et al., 2021), this research goes further by developing street typologies that considered both CUHI and SUHI at the same spatiotemporal resolution. This approach differs from other approaches such as Maharoof et al. (2020), which primarily focused on parameters related to the classification of ULCZs, such as openness, surface properties, and façade geometry. Their research aimed to categorize different urban areas based on these parameters, and the data collected was then used to validate a physics-based model implemented. In contrast, this research provides a more comprehensive understanding of the factors contributing to UHI patterns at the street-level by means of utilizing a wider range of urban parameters that can be collected and processed from open data sources, along with mapping surface and air temperatures at a high spatiotemporal resolution. From here, the street typologies can be used to explain the mechanisms of both SUHI and CUHI.

The combination of the results of the present research and the earlier work of authors (Pena Acosta et al., 2023a) suggest that the assumption of ULCZs which presupposes universal classification does not hold. UHI is highly context-specific and need to be studied independently within each urban context. However, microclimate demonstrates classifiable and predictable pattern within each urban context. In other words, this study posits that ULCZs should be defined per urban context and not universally. These findings carry significant implications for sustainable urban development, as they can inform targeted strategies for UHI mitigation at a local level, while keeping a holistic approach to the problem.

Furthermore, the diurnal and nocturnal patterns of UHI have received significant attention in the existing literature (Oke, 1982; Zhou et al., 2017). However, relatively less focus has been given to exploring the distinct relationships between urban features and the varying time periods throughout the day. Through the feature importance analysis, it was consistently observed across all clusters that the time of the day holds significant relevance. The impact of urban elements on UHI varies depending on the specific time of day. While it may not be feasible to dynamically adapt urban infrastructure to various moments of the day, it can be possible to gain a deeper understanding of how specific types of

urban elements can potentially impact temperature in different street typologies across various geographical locations. For example reflective pavements (Doulos et al., 2004) that are commonly used to reduce heat absorption during the day might negatively impact the surrounding infrastructure if not studied in conjunction with it. Overall, these surfaces will still release heat at night, particularly after long days, contributing to the overall UHI locally. Hence, examining how the built environment responds to both CUHI and SUHI at various times of the day can provide valuable insights for developing targeted, heat-resilient infrastructure. This is particularly important in light of heatwaves in countries that are not accustomed to facing such annual disasters.

When implementing data-driven approaches, there is often a question about the generalizability of the models (Pena Acosta et al., 2023a), as the required data may not always be available. Yet, the question of the level of granularity needed to achieve the best performance (and therefore the accuracy of the results) in these models is not frequently raised. This research evaluates the level of depth represented in the number of groups of streets needed to best explain the patterns of CUHI and SUHI concurrently. The results suggest that analyzing the UHI at the street-level by typologies increases the performance of the models and that a good representation of the behavior of the models can be achieved by taking a representative sample of the streets within each cluster.

Importantly, the approach employed in this research shed light on clear patterns of feature importance when individual street typologies were examined. Both clustering methods generated similar results in terms of feature importance. This is particularly valuable for urban planners and decision-makers seeking a more comprehensive understanding of the importance of various features, and mitigation strategies as it highlights the context-specific nature of UHI. However, it should be noted that this study was based on a sample of 105 streets, and while it shows the applicability of the methods, it is worthwhile to expand the study to more streets, this could result in a different number of typologies. However, the results of this research open the discussion on how UHI phenomena mitigation strategies can be studied in terms of streets rather than cities (this is a bottom-up approach, rather than top-down). Although the data campaign yielded valuable insights into the correlation between UHI intensity and the generated typologies, a limitation exists concerning temperature variation at the reference location. Over the course of an hour-long bicycle route, it is probable that the temperature at the reference location experienced fluctuations that were not accounted for in the methodology presented in this research. For future studies, it is advisable to implement continuous measurements of both canopy and surface temperatures at the reference location to better capture these variations.

6. Conclusions

This research analyzed different typologies of urban streets based on socio-economic and morphological factors and evaluates how these parameters play a different role in creating this temperature variation of CUHI and SUHI at the same spatiotemporal resolution at the street-level. The results demonstrate how data-driven approaches can provide a more nuanced understanding of the interplay between the built environment and UHI. Furthermore, analyzing UHI at the street-level by typologies can increase the performance of the models, highlighting the context-specific nature of UHI.

The findings suggest that a bottom-up approach to studying UHI, based on street-level typologies can provide a more accurate and context-specific understanding of the phenomena, which is critical for developing effective mitigation strategies. Furthermore, urban morphological and socio-economic features have varying levels of importance in different typologies and parts of the day for both types of UHI. Urban geometries such as building height and road width are important for CUHI and SUHI, respectively, while features such as relative humidity, temperature, and building density consistently remain important in determining both types of UHI. The results

highlight the importance of considering the time of day when developing mitigation strategies, as the drivers of UHI can vary depending on the time of day.

The outcomes of this study contribute to the understanding of practical applications and could have a significant impact on the design and planning of climate-conscious urban environments, particularly in the face of increasing heatwaves and the intensification of urban temperatures. Consequently, the authors advocate against a one-size-fits-all approach to mitigation strategies. Instead the finding of this research should direct urban planners to (1) look at the characteristics of streets in their jurisdiction and apply clustering to identify specific street types, (2) develop a strategy for high-resolution data collection from a set of representative streets from each street types, and (3) develop their own context-specific UHI decision support tool as shown in the earlier work of the author (Pena Acosta et al., 2021).

However, it is crucial to note that this research relied on a sample of 105 streets. Expanding the study to include more streets may lead to different typologies. An intriguing avenue for future research would be the implementation of a surrogate modeling approach to concurrently explore both CUHI and SUHI at a high spatio-temporal resolution. This approach would build on the street typologies method to pinpoint pivotal streets in each city. For these essential streets, a physics-based modeling technique could be employed, with high-resolution data gathered specifically for those streets. Based on the detailed results from this surrogate model, a data-driven model for the entire city can then be developed, resulting in a model that is both highly context-sensitive and accurate.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Adilkhanova, I., Ngarambe, J., & Yun, G. Y. (2022). Recent advances in black box and white-box models for urban heat island prediction: Implications of fusing the two methods. *Renewable and Sustainable Energy Reviews*, 165, Article 112520. <https://doi.org/10.1016/J.RSER.2022.112520>, 112520.
- Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering—a decade review. *Information Systems*, 53, 16–38.
- Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, 63(2), 503–527.
- Alkbari, H., & Kolokotsa, D. (2016). Three decades of urban heat islands and mitigation technologies research. *Energy and Buildings*, 133, 834–842.
- Aleksandrowicz, O., Vuckovic, M., Kiesel, K., & Mahdavi, A. (2017). Current trends in urban heat island mitigation research: Observations based on a comprehensive research repository. *Urban Climate*, 21, 1–26. <https://doi.org/10.1016/j.ulclim.2017.04.002>
- Ali-Toudert, F., & Mayer, H. (2007). Effects of asymmetry, galleries, overhanging facades and vegetation on thermal comfort in urban street canyons. *Solar Energy*, 81(6), 742–754.
- Chakraborty, T., Hsu, A., Manya, D., & Sheriff, G. (2020). A spatially explicit surface urban heat island database for the United States: Characterization, uncertainties, and possible applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 168, 74–88. <https://doi.org/10.1016/j.isprsjprs.2020.07.021>
- Chakraborty, T., & Lee, X. (2019). A simplified urban-extent algorithm to characterize surface urban heat islands on a global scale and examine vegetation control on their spatiotemporal variability. *International Journal of Applied Earth Observation and Geoinformation*, 74, 269–280. <https://doi.org/10.1016/j.jag.2018.09.015>
- Chao, Z., Wang, L., Che, M., & Hou, S. (2020). Effects of different urbanization levels on land surface temperature change: Taking Tokyo and Shanghai for example. *Remote Sensing*, 12(12), 2022.
- Ching, J., Mills, G., Bechtel, B., See, L., Feddeema, J., Wang, X., & Neophytou, M. (2018). WUDAPT: An urban weather, climate, and environmental modeling infrastructure for the anthropocene. *Bulletin of the American Meteorological Society*, 99(9), 1907–1924.
- Creutzig, F., Lohrey, S., Bai, X., Baklanov, A., Dawson, R., Dhakal, S., & Munoz, E. (2019). Upscaling urban data science for global climate solutions. *Global Sustainability*, 2, e2.
- Dang, L., & Kim, S. (2023). An analysis of the spatial and temporal evolution of the urban heat island in the city of Zhengzhou using MODIS data. *Applied Sciences*, 13(12), 7013.
- Demuzere, M., Kittner, J., & Bechtel, B. (2021). LCZ generator: A web application to create local climate zone maps. *Frontiers in Environmental Science*, 9. <https://doi.org/10.3389/fenvs.2021.637455>
- Dewan, A., Kiselev, G., & Botje, D. (2021). Diurnal and seasonal trends and associated determinants of surface urban heat islands in large Bangladesh cities. *Applied Geography*, 135, Article 102533.
- Doulos, L., Santamouris, M., & Livada, I. (2004). Passive cooling of outdoor urban spaces. The role of materials. *Solar Energy*, 77(2), 231–249.
- Du, H., Zhan, W., Liu, Z., Li, J., Li, L., Lai, J., & Jiang, S. (2021). Simultaneous investigation of surface and canopy urban heat islands over global cities. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181, 67–83. <https://doi.org/10.1016/j.isprsjprs.2021.09.003>
- Feng, W., & Liu, J. (2022). A literature survey of local climate zone classification: Status, application, and prospect. *Buildings*, 12(10), 1693.
- Geng, X., Zhang, D., Li, C., Yuan, Y., Yu, Z., & Wang, X. (2023). Impacts of climatic zones on urban heat island: Spatiotemporal variations, trends, and drivers in China from 2001 to 2020. *Sustainable Cities and Society*, 89, Article 104303.
- Grimmond, C. S. B., Blackett, M., Best, M., Barlow, J., Baik, J., Belcher, S., & Dandou, A. (2010). The international urban energy balance models comparison project: First results from phase 1. *Journal of Applied Meteorology and Climatology*, 49(6), 1268–1292.
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17, 107–145.
- Howard, L. (1833). The climate of London. <http://www.jstor.org/stable/1793062>.
- Huang, F., Zhan, W., Wang, Z. H., Voogt, J., Hu, L., Quan, J., & Lai, J. (2020). Satellite identification of atmospheric-surface-subsurface urban heat islands under clear sky. *Remote Sensing of Environment*, 250, Article 112039.
- Kim, S. W., & Brown, R. D. (2021). Urban heat island (UHI) variations within a city boundary: A systematic literature review. *Renewable and Sustainable Energy Reviews*, 148, Article 111256.
- Ding, L., Petersen, H., Craft, W., Haddad, S., Yenneti, K., Paolini, R., & Glackin, S. (2023). *Urban heat island mitigation decision-support tool*. Retrieved 17 November from <http://uhimitigationindex.fbe.unsw.edu.au/uhitool/index.html>.
- Li, H., Zhou, Y., Wang, X., Zhou, X., Zhang, H., & Sodoudi, S. (2019). Quantifying urban heat island intensity and its physical mechanism using WRF/UCM. *Science of the Total Environment*, 650, 3110–3119.
- Litardo, J., Palme, M., Borbor-Cordova, M., Caiza, R., Macias, J., Hidalgo-Leon, R., & Soriano, G. (2020). Urban heat island intensity and buildings' energy needs in Duran, Ecuador: Simulation studies and proposal of mitigation strategies. *Sustainable Cities and Society*, 62, Article 102387 (2020). *Sustainable Cities and Society*, 62.
- Liu, H., Zhan, Q., Yang, C., & Wang, J. (2018). Characterizing the spatio-temporal pattern of land surface temperature through time series clustering: Based on the latent pattern and morphology. *Remote Sensing*, 10(4), 654.
- Liu, Z., Zhan, W., Lai, J., Bechtel, B., Lee, X., Hong, F., & Li, J. (2022). Taxonomy of seasonal and diurnal clear-sky climatology of surface urban heat island dynamics across global cities. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187, 14–33.
- Luo, Y., Jiang, Y., Khan, S., Peng, S., Feng, Y., & Han, B. (2010). Analysis of urban heat island effect using k-means clustering. In *Proceedings of the 2nd international conference on information science and engineering*.
- Lyu, F., Wang, S., Han, S. Y., Catlett, C., & Wang, S. (2022). An integrated cyberGIS and machine learning framework for fine-scale prediction of urban heat island using satellite remote sensing and urban sensor network data. *Urban Informatics* 2022 1:1, 1(1), 1–15. <https://doi.org/10.1007/S44212-022-00002-4>
- Maharoof, N., Emmanuel, R., & Thomson, C. (2020). Compatibility of local climate zone parameters for climate sensitive street design: Influence of openness and surface properties on local climate. *Urban Climate*, 33, Article 100642.
- Manoli, G., Fatichi, S., Schlapfer, M., Yu, K., Crowther, T. W., Meili, N., & Bou-Zeid, E. (2019). Magnitude of urban heat islands largely explained by climate and population. *Nature*, 573(7772), 55–60. <https://doi.org/10.1038/s41586-019-1512-9>
- Mirzaei, P. A. (2015). Recent challenges in modeling of urban heat island. *Sustainable Cities and Society*, 19, 200–206. <https://doi.org/10.1016/J.JSCS.2015.04.001>
- Mohammad, P., & Goswami, A. (2021). Quantifying diurnal and seasonal variation of surface urban heat island intensity and its associated determinants across different climatic zones over Indian cities. *GIScience & Remote Sensing*, 58(7), 955–981.
- Oke, T. R. (1982). The energetic basis of the urban heat island. *Quarterly Journal of the Royal Meteorological Society*, 108(455), 1–24.
- Oke, T. R. (1988a). Street design and urban canopy layer climate. *Energy and Buildings*, 11(1–3), 103–113.
- Oke, T. R. (1988b). The urban energy balance. *Progress in Physical geography*, 12(4), 471–508.
- Oukawa, G. Y., Krecl, P., & Targino, A. C. (2022). Fine-scale modeling of the urban heat island: A comparison of multiple linear regression and random forest approaches. *Science of the Total Environment*, 815, Article 152836.
- PDOK Datasets. 2023 <https://www.pdoek.nl/datasets>.
- Pena Acosta, M., Dikkens, M., Vahdatikhaki, F., Santos, J., & Dorée, A. (2023a). A comprehensive generalizability assessment of data-driven urban heat island (UHI) models. *Sustainable Cities and Society*. <https://doi.org/10.1016/j.scs.2023.104701>
- Pena Acosta, M., Vahdatikhaki, F., Santos, J., Hammad, A., & Doree, A. (2021). How to bring UHI to the urban planning table? A data-driven modeling approach. *Sustainable Cities and Society*, 71. <https://doi.org/10.1016/J.JSCS.2021.102948>
- Pena Acosta, M., Vahdatikhaki, F., Santos, J., Hammad, A., & Doree, A. (2022). A framework for a comprehensive mobile data acquisition setting for the assessment of urban heat island phenomenon. In *Proceedings of the 39th international symposium on automation and robotics in construction*. ISARC, 2022.

- Pena Acosta, M., Vahdatikhaki, F., Santos, J., & Dorée, A. (2023). A comparative analysis of surface and canopy layer urban heat island at the micro level using a data-driven approach. *Submitted for publication*.
- Peng, W., Wang, R., Duan, J., Gao, W., & Fan, Z. (2022). Surface and canopy urban heat islands: Does urban morphology result in the spatiotemporal differences? *Urban Climate*, 42. <https://doi.org/10.1016/j.ul clim.2022.101136>
- Qi, J., Ding, L., & Lim, S. (2020). Planning for cooler cities: A framework to support the selection of urban heat mitigation techniques. *Journal of cleaner production*, 275. <https://doi.org/10.1016/j.jclepro.2020.122903>
- Rizwan, A. M., Dennis, L. Y., & Chunho, L. (2008). A review on the generation, determination and mitigation of urban heat island. *Journal of Environmental Sciences*, 20(1), 120–128.
- Salvati, A., Kolokotroni, M., Kotopouleas, A., Watkins, R., Giridharan, R., & Niklopoulou, M. (2022). Impact of reflective materials on urban canyon albedo, outdoor and indoor microclimates. *Building and Environment*, 207, Article 108459. <https://doi.org/10.1016/j.buildenv.2021.108459>
- See, L., Mills, G., & Ching, J. (2015). Community initiative tackles urban heat. *Nature*, 526(7571), 43, 43.
- Shen, C., Hou, H., Zheng, Y., Murayama, Y., Wang, R., & Hu, T. (2022). Prediction of the future urban heat island intensity and distribution based on landscape composition and configuration: A case study in Hangzhou. *Sustainable Cities and Society*, 83, Article 103992.
- Shen, Y., Zeng, C., Cheng, Q., & Shen, H. (2021). Opposite spatiotemporal patterns for surface urban heat island of two "Stove cities" in China: Wuhan and Nanchang. *Remote Sensing*, 13(21), 4447.
- Stewart, I. D., & Oke, T. R. (2012). Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, 93(12), 1879–1900. <https://doi.org/10.1175/bams-D-11-00019.1>
- Sun, T., Sun, R., & Chen, L. (2020). The trend inconsistency between land surface temperature and near surface air temperature in assessing urban heat island effects. *Remote Sensing*, 12(8), 1271.
- Ullah, N., Siddique, M. A., Ding, M., Grigoryan, S., Khan, I. A., Kang, Z., & Zhang, Y. (2023). The impact of urbanization on urban heat island: Predictive approach using Google Earth engine and CA-Markov modelling (2005–2050) of Tianjin City, China. *International Journal of Environmental Research and Public Health*, 20(3), 2642.
- Unger, J. (2004). Intra-urban relationship between surface geometry and urban heat island: Review and new approach. *Climate Research*, 27(3), 253–264.
- Van Hove, L., Steeneveld, G., Jacobs, C., Heusinkveld, B., Elbers, J., Moors, E., & Holtslag, A. (2011). Exploring the urban heat island intensity of Dutch cities: Assessment based on a literature review, recent meteorological observation and datasets provide by hobby meteorologists.
- Venter, Z. S., Chakraborty, T., & Lee, X. (2021). Crowdsourced air temperatures contrast satellite measures of the urban heat island and its mechanisms. *Science Advances*, 7 (22), eabb9569.
- Wang, R., Gao, W., & Peng, W. (2020). Downscale MODIS land surface temperature based on three different models to analyze surface urban heat island: A case study of Hangzhou. *Remote Sensing*, 12(13), 2134.
- Wong, N. H., Tan, C. L., Kolokotsa, D. D., & Takebayashi, H. (2021). Greenery as a mitigation and adaptation strategy to urban heat. *Nature Reviews Earth & Environment*, 2(3), 166–181.
- Yang, C., Yan, F., & Zhang, S. (2020). Comparison of land surface and air temperatures for quantifying summer and winter urban heat island in a snow climate city. *Journal of Environmental Management*, 265, Article 110563.
- Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. *J. 2(2)*, 226–235.
- Zhou, B., Rybski, D., & Kropp, J. P. (2017). The role of city size and urban form in the surface urban heat island. *Scientific Reports*, 7(1), 4791.