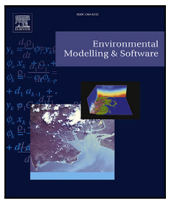
Environmental Modelling and Software 186 (2025) 106364

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft



URSUS_LST: URban SUStainability intelligent system for predicting the impact of urban green infrastructure on land surface temperatures

Francisco Rodríguez-Gómez [a], José del Campo-Ávila [a,*], Luis Pérez-Urrestarazu [b], Domingo López-Rodríguez [c]

[a] Universidad de Málaga, Andalucía Tech, Departamento de Lenguajes y Ciencias de la Computación, Campus de Teatinos, Málaga, 29071, Spain
[b] Universidad de Sevilla, Andalucía Tech, Urban Greening & Biosystems Engineering Research Group, Area of Agro-Forestry Engineering, Ctra. Utrera km. 1, Sevilla, 41013, Spain
[c] Universidad de Málaga, Andalucía Tech, Departamento de Matemática Aplicada, Campus de Teatinos, Málaga, 29071, Spain

ARTICLE INFO

Dataset link: http://earthexplorer.usgs.gov, https://centrodedescargas.cnig.es

Keywords:
Expert system
Urban greening
Urban heat island
Regression models
Open-source

ABSTRACT

Mitigating Urban Heat Island (UHI) effects has become a challenge to improve urban sustainability. The simulation tool URSUS_LST has been developed to allow urban planners to estimate how the addition of different green infrastructure elements would affect temperature. To achieve this, a new methodology was defined based on data mining, geospatial image processing and the knowledge of experts in the domain that predicts the Land Surface Temperature (LST) of any location within a city. It consists of a first data mining phase in which the real LST and the different urban elements of the nearby environment are considered: buildings, vegetation and water bodies. In a second phase, different regression models are induced to predict LST. Additionally, considering the most accurate models, the relevant attributes and their relationships are identified. A real application of the tool in the city of Malaga (Spain) has been used as an example of its usefulness.

1. Introduction

The Urban Heat Island (UHI) effect refers to the difference in temperature between urban constructed areas and rural areas, where there is usually more vegetation cover, including different types of vegetation (forests, agricultural fields or grassland). This effect happens as the result of using construction materials that accumulate and re-emit heat. The consequences are devastating as increased heat is related to diseases and mortality (Goggins et al., 2012; Heaviside et al., 2016), and also affects energy consumption and the quality of life of residents (Lin et al., 2013).

Traditionally, the difference in air temperature between urban and rural areas are used to calculate UHI. However, due to the need of high temporal and special resolution, the land surface temperature (LST) has been also commonly used (Mohamed et al., 2017; Dutta et al., 2021). Therefore, the intensity of the UHI is higher when the difference between the average LST in urban and rural areas (with more vegetation) is greater (El-Hattab et al., 2018).

Urban Green Infrastructure (UGI) elements are a subset of Green Infrastructure (GI) or vegetation elements that are exclusively used in cities like green walls, green roofs or parks, whereas GI elements are used in both urban and non-urban areas (European Commission, 2013). The spatial distribution of temperatures in a city is usually directly related to the presence, amount (low, moderate, or dense) and type of vegetation (Susca et al., 2011; Schlink et al., 2020). Therefore, Urban Green Infrastructures (UGI) play an important role in mitigating the UHI effect (Hart and Sailor, 2009). This problem becomes a target for urban planning as implementing different alternatives to mitigate the UHI effects can make cities more sustainable. Adding new urban green infrastructures and expanding the area of existing ones could be one solution (Herrera-Gomez et al., 2017).

Identifying the relationships between urban elements and temperatures in cities is a complex task, largely due to the vast quantities of data and their diverse nature, involving many variables. Data mining techniques and applications have demonstrated their ability to uncover hidden knowledge in many domains (Liao et al., 2012), including the effect of urban green infrastructures on temperature (Zumwald et al., 2021). Therefore, data mining is a suitable approach to be applied to the problem of determining the Land Surface Temperature (LST) based on the immediate environment.

* Corresponding author.
E-mail addresses: francisco.rdg.gmz@uma.es (F. Rodríguez-Gómez), jcampo@uma.es (J. del Campo-Ávila), lperez@us.es (L. Pérez-Urrestarazu), dominlopez@uma.es (D. López-Rodríguez).

https://doi.org/10.1016/j.envsoft.2025.106364
Received 7 November 2023; Received in revised form 5 December 2024; Accepted 1 February 2025
Available online 11 February 2025

The data mining process definitions are organised in phases sharing main characteristics (Ponsard et al., 2017). The first phase is where the *objectives are defined* and the potential data sources are identified. In the *data preparation phase*, the necessary transformations are carried out to obtain the final dataset to train the algorithms and obtain candidate models. Several techniques are used in the *modelling phase*, and different models are induced. Finally, in the *evaluation phase* the models are assessed by experts, and the acquired knowledge is integrated into real applications in the *deployment phase*.

Given the multidisciplinary nature of research projects, conducting a data mining process without the participation of domain experts would not allow for a proper design or evaluation of the models, resulting in less satisfactory outcomes. Thus, domain experts have a crucial role to play in various phases of the data mining process.

Identifying the most unfavourable areas due to the UHI effect and simulating how that temperature would change by modifying the composition of urban environmental elements are complex and time-consuming tasks. The former task was already approached in Rodríguez-Gómez et al. (2022a), where a methodology and a tool have been developed to determine the most unfavourable areas due to the urban heat island effect. In the case of the simulation task, incorporating the knowledge extracted through the data mining process into a software tool is a significant advantage for urban planners which has not yet been addressed and which is one of the objectives of the present paper. Efforts to combat the effects of climate change can also be enhanced if experts know how increasing the surrounding vegetation would impact temperature.

New tools are necessary in urban planning to effectively resolve urban problems with UGI (Muñoz and Duarte, 2025). They must be designed for the peculiarities of urban settings, with focus on urban ecosystem services and multi-scalability (Oijstaeijen et al., 2020) and must integrate economic, social, and environmental criteria to provide comprehensive and strategic support for urban planners. Hence, they can guide strategic planning and placement of green infrastructure, ensuring that it meets the specific needs of urban areas and enhances the overall urban ecosystem (Tóth et al., 2015; Cherchyk and Khumarova, 2023). For that, it is essential to have information on UGI and their potential effect in order to provide strategic approaches to integrate the planning of green spaces and elements on different scales (Kajosaari et al., 2024).

1.1. Related work

Meteorological data can be used to create models that predict extreme events and identify links between weather variables at extreme temperatures (Herrera et al., 2018). These models, which use meteorological data, can help with a spatial resolution of several kilometres, which is useful for setting alerts.

When a higher resolution is needed, a frequent method for making Land Surface Temperature (LST) predictions involves training machine learning models from LST time series over the years in various urban locations. For instance, Mustafa et al. (2020) classified the urban environment using supervised machine learning algorithms (four soft computing techniques) and arrived at conclusions regarding temperature variations based on the analysed environment in Beijing (China). It is crucial to note that the models were trained with LST time series data alone, without incorporating additional information regarding the characteristics of the urban environment at the points used for training. Other approaches that utilise time series data from different years apply ARIMA models to predict LST values, as demonstrated in the study of Chennai, India by Kesavan et al. (2021).

Examining the features of the urban environment to make LST predictions is a way of incorporating new information that could provide an enhanced understanding of the issue. In this line, Khalil et al. (2021) proposed training machine learning algorithms by analysing environmental features (EVI: Enhanced Vegetation Index, elevation, and road density) along with LST data collected from Lahore, Pakistan over 20 years. Another example employs genetic algorithms to induce models through training with space–time features (such as building density or air pollution level) for LST prediction in Iran (Karimi and Ghajari, 2022).

Employing Landsat images to extract the LST for model training is common, but other sources for images can also be utilised. For example, Kartal and Sekertekin (2022) uses MODIS satellite images from the southern region of Turkey. While Landsat images offer lower temporal resolution, the drawback of MODIS images is their lower spatial resolution.

Research based on the training of models from LST time series that do not consider the urban features of the study areas presents a limitation. Such studies can classify green infrastructure typologies (Bartesaghi-Koc et al., 2019), can assess the impact of change in the vegetation covers (Rahaman et al., 2022), or even can identify best localisation of urban parks (Nesticò et al., 2022), but they do not allow for the simulation of new scenarios to predict temperature by modifying environmental characteristics. Incorporating such features in the training phase is crucial; and LiDAR images can provide that information. Sharma et al. (2021) demonstrate the full potential of this technology in terms of extracting urban characteristics. One benefit of using LiDAR images is the simplification of the data preparation phase, as it includes an urban classification for each point in the image (e.g. building, vegetation or water) (Ramiya et al., 2017). A recent application in a different field is presented by Rodríguez-Gómez et al. (2022b), who use them to identify urban roofs and determine optimal locations for photovoltaic installations.

1.2. Contributions and organisation of the paper

Prediction can be improved by incorporating urban features in addition to LST values to create a more accurate model. However, the urban characteristics at different distance intervals from the study points should also be taken into account. Conducting an analysis of the urban environment at different distance ranges for a variety of study points allows the impact of adding urban green infrastructure at varying distances to be simulated. Calibrating the impact of the distance factor on LST is another benefit for domain experts, providing them with novel and relevant knowledge.

This paper presents the advances achieved through a data mining process to develop models and tools for predicting the LST and simulating how it is influenced by changes in urban environment features. The proposed methodology is extensible to any city and minimises initial pre-processing steps. It is based on the information collected from Landsat images (for calculating LST values), Sentinel images (for segmenting urban water masses) and LiDAR images (for extracting urban features).

Specifically, the main objective of this research is to develop a tool for predicting the impact of adding new urban green infrastructure on temperatures in the areas most affected by the urban heat island effect. The system allows for tuning low, moderate, or dense vegetation in increments of 250 m distance intervals from 0 to 1 km. This will enable users to determine the optimal distance, quantity, and type of urban green infrastructure.

The most relevant contributions are the following:

- Identification of the most important variables for predicting LST based on the urban elements of the environment. As experts have selected a broad set of urban features relevant for model training, the resulting models take into account the environmental features with the greatest impact on temperature.
- Design of an expert system for predicting LST values in a region by analysing the urban environment features.

Table 1
Urban elements and NDVI ranges.

	No vegetation		Vegetation		
	Water	Other	Low	Moderate	Dense
NDVI range	[−1.0, 0.0]	(0.0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 1.0]

- Development of a software tool (implemented as open-source) that enables urban planners to simulate the impact of increasing urban green infrastructures on the temperature. This increase can be defined as a percentage, and its effect can be measured at different distance ranges.

The rest of the paper is organised as follows. The prior knowledge necessary to follow the article is described in Section 2. The proposed methodology is explained in Section 3: the data preparation phase is described in Section 3.1, including the data sources, their selection and processing, while Section 3.2 describes the algorithms and configurations that have been tested and evaluated. Results, including the model, the web application, and the most important variables to predict the LST, are described in Section 4 and Section 5. Finally, Section 6 summarises the main conclusions of the work.

2. Preliminaries

This section aims to review a series of crucial concepts for a thorough understanding of the proposed methodology. These include: calculation of the Normalised Difference Vegetation Index (NDVI), calculation of Land Surface Temperature (LST), supervised learning algorithms for regression, and metrics for evaluating the accuracy of the induced models. All references to temperatures from now on correspond to LST.

2.1. Calculating normalised difference vegetation index from sentinel-2

The Normalised Difference Vegetation Index (NDVI) is a numerical index used to assess the presence of vegetation in urban areas. It ranges from −1 to 1, and Table 1 illustrates the correspondence between NDVI ranges and various urban elements (Fusami et al., 2020). This same classification was used for defining the quantity of vegetation observed in the different pixels or areas.

The NDVI value for each pixel of a selected Sentinel-2 satellite image is calculated using the red ($b4$) and the near-infrared ($b8$) bands according to the following formula:

$$\mathrm{NDVI} = \frac{b8 - b4}{b8 + b4} \tag{1}$$

The vegetation amount (density or strata) refers to the fraction of the image pixel covered by vegetation, and it is related to the NDVI value calculated in it (Jiang et al., 2006).

2.2. Calculating land surface temperature from landsat −8

Land Surface Temperature (LST) is often utilised as a metric for identifying the areas most affected by the Urban Heat Island (UHI) effect. Images from the Landsat −8 satellite are employed to calculate LST. To generate a raster layer with LST values for each pixel of a selected image of a city, the first step is to calculate the Top of Atmospheric (TOA) spectral radiance (L_λ) using the rescaling factors provided in the metadata file (either Band 10 or Band 11 may be used) (Mejbel Salih et al., 2018):

$$\mathrm{L}_\lambda = M_L \cdot Q_{cal} + A_L \tag{2}$$

where M_L and A_L represent the band-specific multiplicative and additive rescaling factors from the metadata, respectively; and Q_{cal} is the quantised and calibrated standard product digital number pixel value.

The TOA spectral radiance L_λ is then converted to brightness temperature (BT) using the following formula:

$$\mathrm{BT} = \frac{K_2}{\ln\left(\frac{K_1}{L_\lambda} + 1\right)} \tag{3}$$

where K_1 and K_2 are band-specific thermal conversion constants from the metadata, corresponding to the same band used to calculate L_λ.

Intermediate calculations are necessary to obtain the LST value, as detailed in Mejbel Salih et al. (2018). The main term is the Land Surface Emissivity (LSE or ε_λ), which is influenced by certain emissivity constants and the proportion of vegetation (computed from NDVI values) according to:

$$\mathrm{LST} = \frac{BT}{1 + \frac{\lambda \cdot BT}{\rho} \cdot \ln \varepsilon_\lambda} \tag{4}$$

where λ is the mean wavelength of the band used and ρ is a constant (dependent on universal constants such as Boltzmann's or Planck's).

2.3. Supervised learning algorithms for regression problems

In this subsection, the regression algorithms to be used in the modelling phase are listed and briefly described.

Machine learning classification algorithms involve identifying relationships between input data and output data. The input data may be discrete or continuous, while the output data (class) is typically a discrete variable (Liu and Wu, 2012). Regression methods are employed when the class variable is numerical. In supervised learning techniques, the class to which observations belong is known and the models learn to predict the dependent variable (class) from labelled examples. When the goal is to predict LST values, we can use supervised learning regression techniques, as the class (LST at a given study point) and the composition of the urban environment within a radius of 1 km for each observation are known.

There is a wide array of classification algorithms available. In this work, we will focus on regression algorithms that are suitable for addressing the problem at hand.

Artificial Neural Networks (ANNs) (Murtagh, 1991) consist of interconnected neurons that are organised into input, hidden, and output layers. The connections between neurons (weights) are adjusted to alter the relationship between inputs and outputs and filter input values and generate a desired output with the help of activation functions. ANNs often employ the backpropagation technique to adjust the weights of the network to fit the desired output to the input variables from labelled examples.

Support Vector Machines (SVMs) (Cortes et al., 1995) are a type of classification algorithm that utilise linear classifiers to identify the hyperplane that divides data into different categories. Support Vector Regression (SVR) (Awad and Khanna, 2015) is a variant of SVM specifically designed for regression problems. The algorithm searches for a curve or line (hyperplane) with maximum symmetric margin on both sides that best captures the trend of the data.

Decision trees are supervised machine learning algorithms that can generate interpretable and explainable predictive models for classification or regression problems. Each branch in the tree represents a rule expressed as a conditional statement that can be understood without expert knowledge. C4.5 (Quinlan, 2014) is a widely used decision tree algorithm. When the target variable to predict is numerical, regression trees (RTs) are utilised (Breiman et al., 2017).

Ensemble methods, such as bagging or boosting, are often used in conjunction with decision trees and are often used in the environmental field as reference algorithms (Zumwald et al., 2021).

Bagging (Breiman, 1996) is a learning method in which multiple weak models are trained with different sets of randomly selected observations from the dataset. The training data for each new model may include data used to train other models. The model's output is the mean

of the output of the weak models. Inspired by this approach, Random Forest (RFs) (Breiman, 2001) is a decision tree-based technique that has demonstrated great success (Wyner et al., 2017). It involves generating multiple decision trees using different subsets of attributes. If regression trees are employed instead of decision trees, a random regression forest is created.

XGBoost (XGB) (Bentéjac et al., 2020), or eXtreme Gradient Boosting, is a supervised machine learning predictive algorithm that uses the boosting principle to generate multiple weak prediction models sequentially. Each model is built upon the results of the previous model to create a new model with improved predictive power and greater stability in its results. An optimisation algorithm such as Gradient Descent is used to combine these weak models into a more robust model.

2.4. Evaluation metrics

To assess the accuracy of the models produced by supervised learning algorithms for regression problems, we will utilise the following metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R-Squared (R^2). All of which are well-known and we describe them in this section for the sake of self-completeness of the article.

The coefficient of determination, denoted by R^2, is a statistical measure that provides information about the fit of a regression model. Specifically, it represents the proportion of the variance in the dependent variable (response) that is predictable from the independent variable (predictor). The formula for the coefficient of determination R^2 is

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\hat{y}_i$ represents the predicted value of the dependent variable for observation i, $\bar{y}$ is the mean value of the dependent variable, y_i is the observed value of the dependent variable for observation i, and n is the number of observations. The formula can also be expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where the numerator represents the sum of squared errors (SSE) and the denominator represents the total sum of squares (TSS). Therefore, R^2 can be interpreted as the proportion of the TSS that is explained by the SSE. The value of R^2 ranges from 0 to 1, where 0 indicates that the model does not explain any of the variability of the response data around its mean, and 1 indicates that the model perfectly explains it all.

R^2 is often used as a goodness-of-fit metric to determine how well the model fits the data. One advantage of R^2 is that it is easy to interpret, as it provides a simple indication of the proportion of variability in the response variable that is explained by the model. However, a disadvantage is that R^2 does not provide information about the validity of the model's assumptions or the statistical significance of the model. Therefore, it should be used in conjunction with other diagnostic measures to assess the overall quality of the model. In this work, we use the MAE and RMSE, defined as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Both MAE and RMSE provide useful insights into the performance of a regression model, since they measure the error incurred when approximating y_i through $\hat{y}_i$, but they have different interpretations and applications. MAE is often used to compare the prediction errors of different models, while RMSE is often used to assess the overall fit of a model to the data. In general, a lower value for either metric indicates a better fit of the model to the data.

3. Methodology

The methodology followed to conduct the data mining process was the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000). It defines six phases: (1) business understanding, (2) data understanding, (3) data preparation, (4) modelling, (5) evaluation and (6) deployment.

The problem and data study phases (phases 1 and 2) have already been described in Section 1. Given the importance of the data preparation (phase 3) and the modelling (phase 4), two subsections have been dedicated to their description in this Section. The final phases, evaluation and deployment (phases 5 and 6), where the methodology checks and apply the results attending to the initial objectives, are set out at Section 4 and Section 5 respectively.

In the project in which the presented methodology was developed, three experts in urban greening participated. Their involvement was important for the entire data mining process: searching for relevant data and sources, validating images, and checking tools in real urban use cases.

The rest of this section outlines the procedure to obtain the required data to train the models, and a study of the processing conducted to select the most appropriate model for predicting LST.

3.1. Data preparation

The data inputs needed to extract knowledge about the influence of the near-environment on the land surface temperature are compound by three sources:

- *Sentinel-2 images* are required to calculate the NDVI values of the study city using Eq. (1). Considering the scale presented in Table 1, this information is used mainly to locate water bodies and identify the non-vegetated areas in the region. The surroundings of points inside water bodies (such as seas or oceans) are not of interest as they can be considered non-urban points (and their environment is irrelevant). However, the presence of a water body in the near-environment of other urban points is indeed relevant, hence the importance of identifying water bodies in the image. On the other hand, areas without vegetation are of particular interest because they will be the best candidates for action to lower the temperature. It is, therefore, crucial to better understand how the environment affects the temperature in these non-vegetated areas. The resolution of Sentinel-2 images is 10 m/pixel.
- *LiDAR maps* are used to determine the use of urban soil (with or without vegetation). Different labels are assigned to points taking the land use into account (American Society for Photogrammetry & Remote Sensing, 2011). The information relevant to this research is marked with LiDAR labels 3 to 6. Values 3 to 5 are used for low, moderate, and dense vegetation, respectively, and value 6 indicates the presence of buildings. In addition, the LiDAR label 2 stands for "ground", which can be used to determine the height of such buildings. The resolution of the LiDAR images is 0.5 point/m^2 and they have been rasterised to 1 m/pixel.
- *Landsat −8 images* are needed to calculate the LST values at different points of the city using Eq. (4). As mentioned above, the LST value will be the target variable to predict using regression models. The resolution of Landsat −8 images is 30 m/pixel.

Fig. 1 summarises these data inputs and outlines the transformations performed to obtain the final dataset. The details of these transformations are described below.

For each point in the city under examination, we have its classification within one of the following categories: *low*, *moderate* and *dense* vegetation, *building*, *water*, or *other*. This results in an aggregate of 6 categories into which we divide the urban soil. Note that the category "*other*" accounts for roads, bridges, cables, and other urban elements not included in the previous classes.

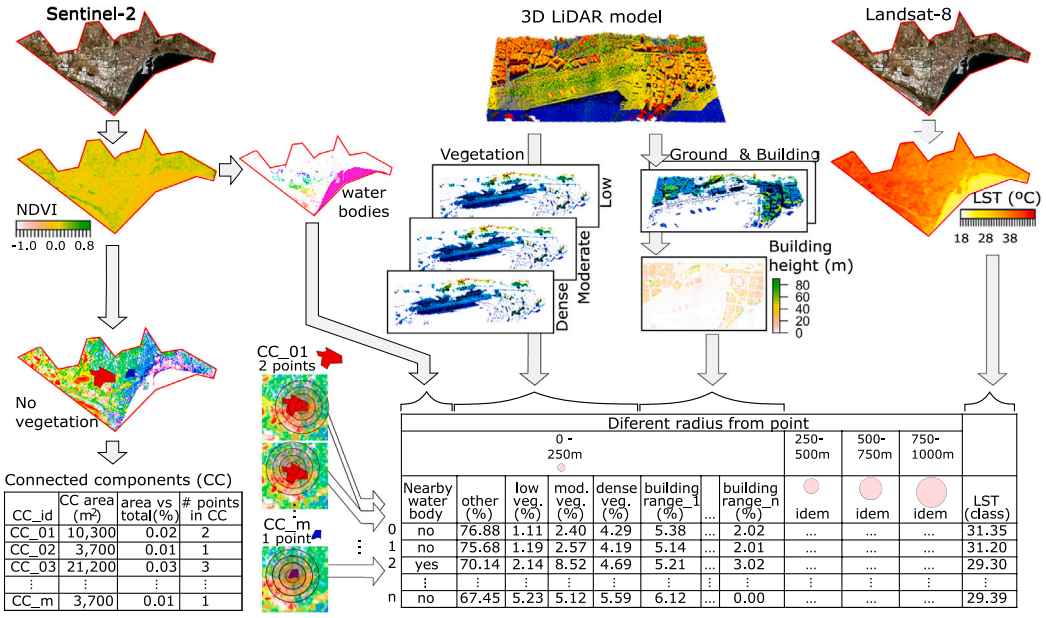


Fig. 1. Proposed methodology to obtain the required data and create the dataset.

As regards the pixels categorised as *building*, the proposed methodology includes a step to calculate new subcategories taking their heights into account. The height is calculated using a normalised digital surface model (NDSM) (Rao et al., 2022), obtained from the LiDAR image filtered with classification points (2: Ground, 6: Building).

The NDSM measures the difference between the Digital Terrain Model (DTM) and the Digital Surface Model (DSM), displaying the heights of objects relative to the ground. If necessary, the domain experts can discretise this numerical attribute in multiple ranges, in line with their knowledge about the studied city.

The points of interest whose near-environment needs to be studied are a determining factor for a successful data mining process. Domain experts are interested in how the environment affects the temperature of non-vegetated areas. For that reason, this methodology follows an adaptive approach to choose points on the map representing every non-vegetated area. Connected Components Labelling (CCL) is the key to segmenting maps and finding these areas. CCL is an image processing technique that assigns a unique label to all the connected pixels due to neighbourhood and similarity (He et al., 2017). This way, separated non-vegetated areas are identified and random points are selected from them. The number of random points chosen from each area depends on its extension with respect to the total area. Thus, few points are selected for small areas, while as the area increases, more points are selected. Choosing the same number points for every area independently of its size is inefficient because the environment of many points in a small area is the same and can be represented with few points (or even only one point).

For each of the randomly selected points, some variables that describe their near-environment are calculated:

- the presence (or not) of a nearby water body, and
- the percentages of land (with respect to the total urban area) occupied by:
  - low, moderate, and dense vegetation,
  - buildings (that can be optionally discretised in ranges), and
  - other urban elements.

These variables are calculated at different scales of distance: less than 250 m, from 250 m to 500 m, from 500 m to 750 m, and finally, from 750 m to 1 km. The whole preparation process is depicted in Fig. 1.

Table 2
Hyperparameter tuning.

Algorithm	Hyperparameters	Analysed values
RT	Max depth	{1, 2, 3, 4, 5}
ANN	Number of neurons in hidden layer	{2, 4, 8, 16, 24}
SVR	Cost	{0.25, 0.5, 1, 2, 4}
XGB	Depth × Number of rounds	{3, 7} × {250, 500, 750}
Bagging	Number of iterations	{10, 20, 25, 35, 40}
RF	Number of trees	{50, 100, 150, 200}

3.2. Modelling

The process carried out for the selection of the best model to predict LST, which is one of the main contributions of this work, follows a model training with hyperparameter optimisation (Bischl et al., 2023).

Several regression algorithms are chosen: Artificial Neural Network (ANN), Regression Trees (RT), Random Forest (RF), Bagging (with regression trees as base classifier), Support Vector Regression (SVR) and XGBoost (XGB), all of them already introduced in Section 2. Since all these algorithms have different hyperparameters that control their execution (and therefore can impact their performance), it is necessary to determine which is the optimal configuration for each.

Thus, fine-tuning of the hyperparameters is performed, conducting a grid search to obtain the configuration that would provide the best results for each algorithm, in terms of minimising the MAE and RMSE errors obtained. In Table 2, we can observe which hyperparameters are analysed for each algorithm, as well as the values that were examined. A 10-fold cross-validation strategy is repeated for 3 times to ensure the stability of the results for each generated model.

4. Evaluation results

The data preparation and modelling phases – part of the data mining methodology presented in the previous section – are generic and can be used whenever data are available. In this section, we show the results evaluated by experts (researchers specialised in urban green areas) in a specific case and applying the methodology in a real city as an example. Therefore, their process to validate the knowledge before being incorporated into the expert system to support the decision-making process is shown. In particular, the Urban Atlas Copernicus project (European Commission, 2014) has been used as a complementary source of information in this validation process.

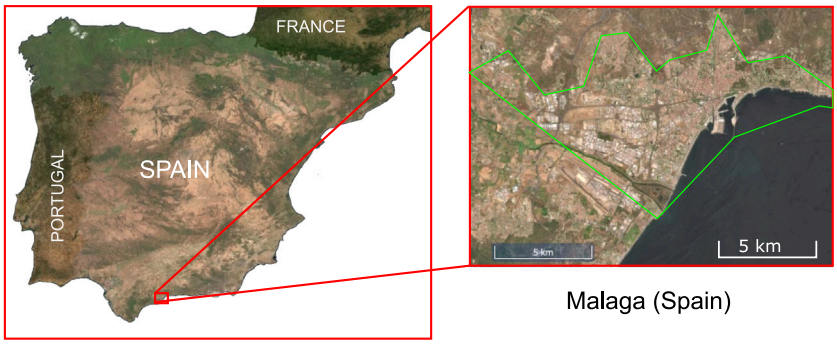


Fig. 2. Location of Malaga in Spain (Europe) and urban area of interest. RGB image downloaded from the Spanish Geographic Institute (IGN) with the license CC-BY 4.0. http://www.ign.es.

4.1. Data preparation: validation example

Data from the city of Malaga, in southern Spain, have been collected and processed following the description given in Section 3.1. Malaga is located in southern Spain (see Fig. 2) and has a population of 578 000 inhabitants. It is on the coast, between two river valleys, with a mountain range to the northeast and the Mediterranean Sea to the south. Temperatures are milder (although with higher relative humidity). Even so, they can reach up to 35 °C.

Every point in the area of interest was defined by the class assigned to it; that is, each point was labelled as water, vegetation (low, moderate, or dense), building (including its height) or other. Each point corresponded to a pixel in the LST image, and its dimension was defined as a square of 10 m × 10 m.

Then the information was segmented using a Connected Components Labelling (CCL) algorithm, and 8 754 connected components were identified. Points were selected to ensure that every connected component was represented. The smallest components, with an area of 100 m^2, only needed one point for their study, while the greatest components, with an area of 1 865 300 m^2 were inspected using 308 points. In the end, a total of 16 678 points were calculated in the city of Malaga.

The observed Land Surface Temperature (LST) was known for every point from those 16 678 selected points, and 28 additional variables were calculated to describe its near-environment. Four concentric regions of 250 m width were defined in the range of 1 km, and seven variables were calculated for every region. Those variables indicated the presence of a water body, the percentage of uncategorised elements (other), the percentage of every kind of vegetation (low, moderate, or dense), and the percentage of two types of buildings. In this case, the experts decided to discretise the height of buildings in two ranges: smaller and bigger than 24 m. Some other discretisation options were considered. For example, avoiding discretisation produced more errors in the model. In contrast, a more fine-grained discretisation did not improve the result but made the model more complex.

4.2. Modelling: validation example

Once the dataset with the information of the near-environment characteristics of relevant points in the city of Malaga (defined by 16 678 examples described by 29 attributes) was created, it was used to train different models. The algorithms and configurations used were enumerated in Table 2 and the best results achieved are described in Table 3. Every algorithm (and configuration) was evaluated by repeating 3 times a 10-fold cross validation, and therefore 30 experiments have been performed. A statistical validation was conducted using the Wilcoxon test over those 30 experiments in order to detect significant differences.

The configurations that obtain the best evaluation metrics (average and standard deviation values) for every algorithm are shown in Table 3. There is one exception with Random Forest as the selected model uses 50 trees instead of 200. The results achieved by 200 trees are slightly better than those achieved by the version with 50 trees, but the differences in accuracy are not statistically significant. We therefore opted for the version with 50 trees, since its computational complexity is much lower.

As is presented in Table 3, the most accurate model is the one induced by Random Forest with 50 trees. In order to check this statement, we have conducted statistical tests comparing this model with the others, showing that the error obtained by the rest of models is significantly greater (p-value $\ll 0.05$). This is depicted in Table 3 with the symbol $\ominus$.

4.3. Evaluating and tuning a model: validation example

Considering the results achieved in the modelling phase, the most accurate models are selected and the experts try to learn something about them. They can also tune the model by making a feature selection to keep the most informative attributes while discarding those that are not so relevant. Thus, the model can be adapted to be less complex, while keeping the explanatory capacity.

The overall importance of every variable is calculated (see Fig. 3). It can be observed that the most relevant variables are those that affect the closest nearby point, in the range between 0 and 250 m. The percentage of high buildings (over 24 m) is particularly relevant. On the other hand, the presence of water bodies does not seem to affect the differences in LST observed between different locations within the studied city.

The final model selected to be incorporated into the expert system implemented during the deployment phase has considered these findings regarding feature importance. After performing a search-based feature selection using a wrapper approach (Li et al., 2017), only the 10 most relevant attributes have been considered. Differences in accuracy between models with 10 or more than 10 attributes are negligible for the experts' requirements.

5. Expert system including knowledge

This section describes the application implemented to use the knowledge gained during the application of the data mining process. The model might change from one city to another, and it might be necessary to repeat the modelling phase to better adjust it to a new city. However, the implemented expert system uses the model in a transparent way. New models can update the preloaded model and continue to work. The steps followed by the application to predict the land surface temperature (LST) at a specific point or to simulate the new LST value when changes in the near-environment are the same regardless of the loaded model.

Table 3
Algorithms, configuration selected, and accuracy values error.

Algorithm	Selected hyperparam.	MAE	RMSE	R^2	
RT	5	1.02 ± 0.020	1.33 ± 0.030	0.55 ± 0.024	⊖
ANN	24	0.86 ± 0.030	1.14 ± 0.050	0.68 ± 0.006	⊖
SVR	4	0.69 ± 0.020	0.98 ± 0.040	0.76 ± 0.017	⊖
XGB	7 × 5	0.47 ± 0.010	0.67 ± 0.030	0.89 ± 0.011	⊖
Bagging	25	1.00 ± 0.020	1.30 ± 0.030	0.58 ± 0.018	⊖
RF	*50*	***0.44 ± 0.010***	***0.63 ± 0.020***	***0.9 ± 0.007***	

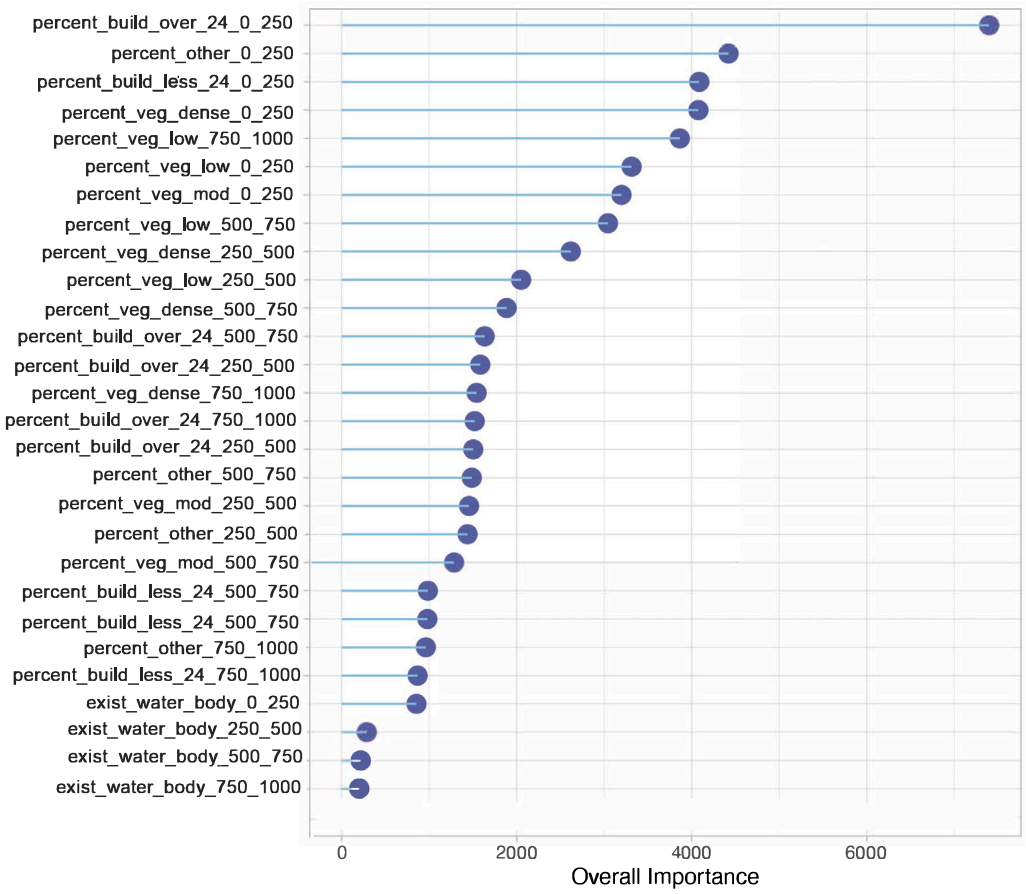


Fig. 3. Importance of variables used during a feature selection process. The model used is a Random Forest with 50 trees. Attributes refer to the percentage of different land uses in four increasing ranges. The name of variables include information such as the kind of variable (*percentage* as a float number or *existence* as a boolean), the kind of land use (*vegetation*, *building* and *other*) and the range (0–250, 250–500, 500–750 and 750-1 000 m).

5.1. Technologies for implementation

Currently, R (R Core Team, 2020) and python are the most widely used languages for data mining, machine learning or artificial intelligence processes. R has been selected for this research as it has the necessary libraries to process LiDAR, Landsat and Sentinel images, and as it offers a framework that facilitates the conversion of the results obtained from data mining processes into web tools.

The main R technologies used to develop the tool can be grouped according to the phase of the data mining process where they have been intensively used.

R packages were used, such as `raster` (Hijmans, 2025) and `lidR` (Roussel et al., 2020; Roussel and Auty, 2021) for data preparation. `lidR` allows the extraction of buildings and vegetation from the near environment while `raster` facilitates the calculation of building height, NDVI, or LST, and the extraction of near environment features in 250 m rings. `dplyr` package (Wickham et al., 2020) was used to obtain core knowledge for training models and for models predictions.

R libraries for machine learning such as the `caret` package (Kuhn, 2008) have been used to train different models using different algorithms and configurations, as well as to evaluate the models.

Finally, the `shiny` package (Chang et al., 2021) has been used for the development of the dashboard for LST predictions and simulations. A Linux server has been configured for the web app deployment.

5.2. Prediction and simulation

Screenshots are presented in Fig. 4 to illustrate the processing of the information to obtain LST predictions. Continuing with the validation example of the previous section, they correspond to the selection of a point in the centre of the city of Malaga, in Spain, and specifically, the Cathedral.

The usage of the application is described below.

Selecting the study point. Cropping area of interest (a). The map shows the user blue squares representing the area covered by the LiDAR imagery, where the study point can be selected. This is the first step:

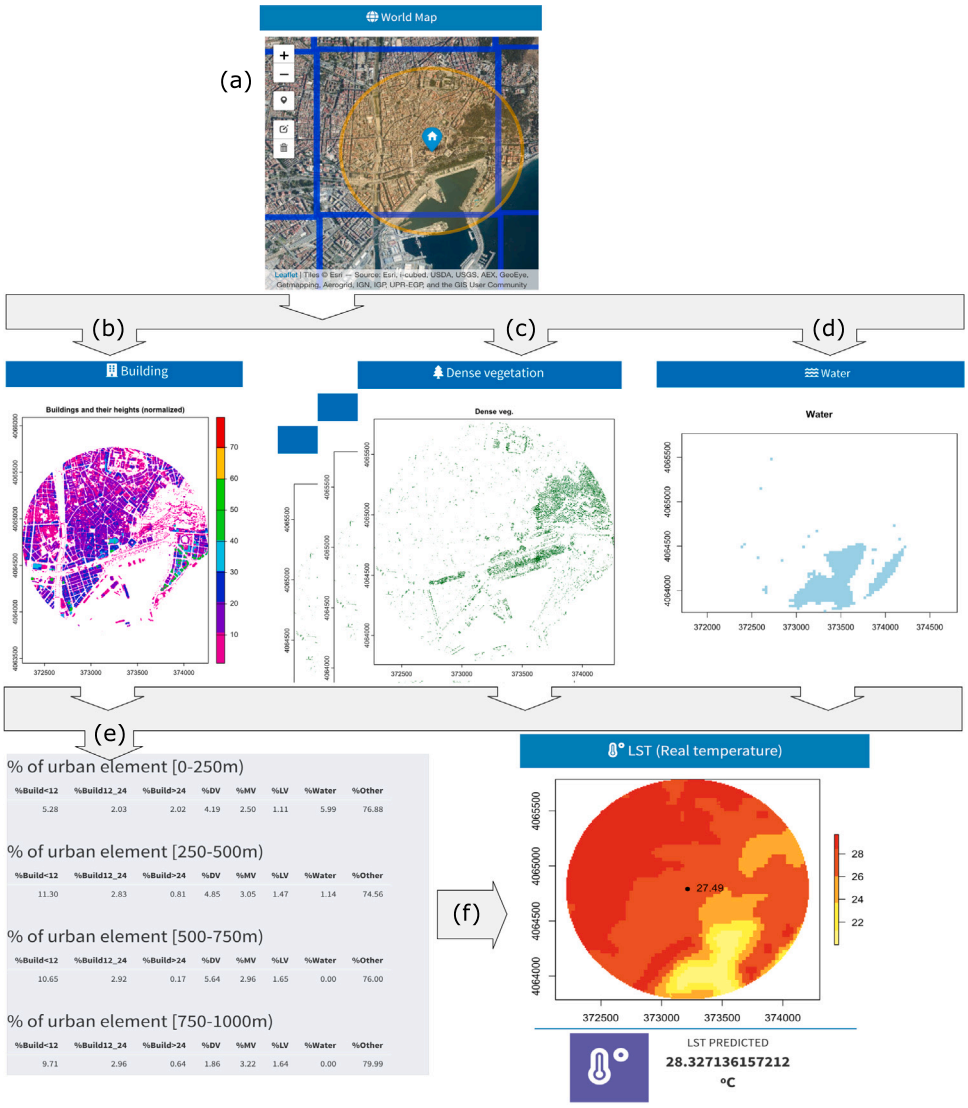


Fig. 4. Screenshots with part of the information generated until a prediction of LST is given: (a) Selecting the study point. Cropping area of interest (3,1416 km^2 – 1 km radius circle–). (b) Building extraction. (c) Vegetation extraction. (d) Existence of water bodies nearby. (e) Calculating the area percent of urban features. (f) Model predictions.

selecting the point of interest on a map using interactive tools. The system then draws an orange circle of radius 1 km and starts extracting urban features. From the catalogue of LiDAR images of the city, the application performs the cropping of the area surrounding the study point for 1 km around it.

Building extraction (b). An image containing only the LiDAR points classified as buildings is obtained from the LiDAR model of the area of interest. The normalised height model is then obtained and a Raster layer of the buildings categorised by height is calculated. For the case of Málaga, the application obtains 2 images by filtering the heights in the image with all buildings to determine the buildings below and above 24 m.

Building height thresholds, potentially more than one, can be configured according to the morphology of each city. In historic cities, of intermediate size, with compact morphology and modern expansion, such as Malaga, 24 m can be an appropriate value because it allows separating two of the most common types of buildings. Those in the historic centre, with a predominance of low-rise buildings, and those in the expansion area, with high-rise buildings (above 24 m) (Wu et al., 2023).

Vegetation extraction (c). To calculate the areas with low, moderate and dense vegetation, a process similar to the one described above for the segmentation of buildings is followed. In this case, the application calculates 3 LiDAR images, filtering by points belonging to low, moderate and dense vegetation, and their 2D raster models are then obtained.

In this step, as in the whole process, experts in the domain have validated the results achieved. They have compared the real vegetation identified in the Urban Atlas (Copernicus Project) with the estimation of vegetation automatically calculated (URSUS_LST). Fig. 5 shows these two maps where correspondence is clear.

Existence of water bodies nearby (d). To obtain the areas covered by water, a raster image with the terrain classified according to the NDVI of the city is used. This raster image is cropped to a radius 1 km from the study point. The next step is to perform a Connected Component Labelling on the raster water map, as it is necessary to determine the extent and shape of the different water bodies.

Calculating the area percent of urban features (e). The division of each raster layers (water, buildings or vegetation) into 250 m intervals is easily performed using the `distanceFromPoints(point,raster)` function of the `raster` library. The tool can create a mask that represents the cutting ring, removing points that are at distances that

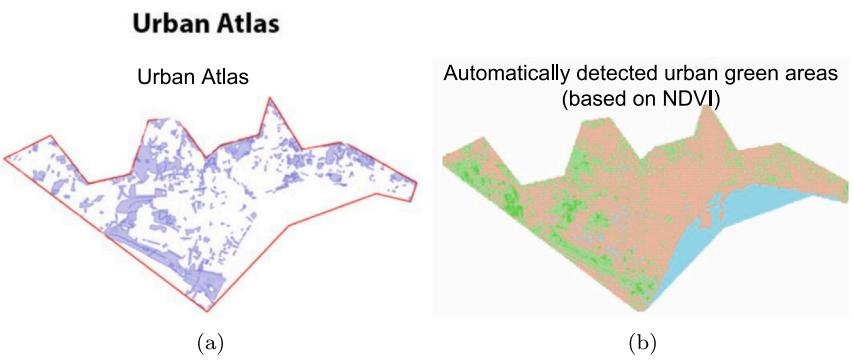


Fig. 5. Green areas in Malaga: (a) Urban Atlas, (b) URSUS_LST.

are not of interest. The distance mask (cutting ring) can be applied on different raster layers and is immediate to calculate the area of every feature.

Model predictions (f). Once the values of the variables that the proposed model needs to make the predictions (presence of water bodies, %vegetation, %buildings, and %other features) have been calculated for each interval of 250 m, the model is fed to obtain the predicted LST at the selected point, and the tool displays the predicted LST. The selected model is Random Forest with 50 trees trained with the 10 most relevant variables as described in Section 4.2. Fig. 4 shows a complete example of a real case in the city of Malaga for the prediction of LST at a selected point.

Simulating new environment scenarios (g). The system is a means to simulate how a modification of the vegetation percentages in the intervals could affect the land surface temperature (LST). This simulation can be used by urban planners and designers, landscape architects and even administrations to decide on the development of new UGI installations.

Although several variables are identified as relevant, the simulator only takes into account the percentage of different vegetation densities (low, moderate and dense). The main reason is that this potential changes are easy to configure and implement in reality by an urban planner. Changing the percentage of buildings or their height would not be so feasible to change unless new constructions are made. In addition, the percentage considered refers to the ring closest to the point (less than 250 m), as these nearest features were proved to have the highest influence on the LST obtained (see Fig. 3).

In this new situation, the tool shows the temperature reduction over the real scenario. Fig. 6 shows the actual information describing the environment of a point and presents the controls that a user can configure to set up a new scenario. The user can define new values for the percentages corresponding to the three categories of vegetation density in the close range. The temperature in the new scenario after changing the vegetation percentages is shown at the bottom.

6. Conclusions

This paper presents a novel methodology for predicting Land Surface Temperature (LST) and simulating the effects of adding new urban green infrastructure on temperatures in areas potentially affected by the urban heat island effect. The incorporation of urban features, in addition to LST values, results in a more accurate model for predicting temperatures. This methodology is extensible to any city and minimises initial pre-processing steps for retraining, making it efficient and effective for urban planners.

This methodology has allowed the optimal model to be determined and trained in the aforementioned task of estimating the LST values from urban features and how they will be affected if there are new urban elements introduced.

The analysis of urban environment features at different distance intervals for different study points has shown that the distance factor's impact on LST can be calibrated to provide domain experts with novel and relevant knowledge. Furthermore, the proposed tool allows for simulating the existence of different quantities of vegetation in a certain radius from the study point (in increments of 250 m distance intervals from 0 to 1 km). This will enable users to determine the optimal distance, quantity, and type of urban green infrastructure to reduce temperatures.

In the proposed methodology, the identification of the most important variables for predicting LST based on the urban elements of the environment ensures that the resulting models take into account those features with the greatest impact on temperature. We have found that the amount of vegetation and buildings in less than 250 m are the most influential variables to determine temperature.

Furthermore, the significant contributions of this work are (a) the design of an expert system for predicting LST in a region by analysing the urban features; and (b) the development of an open-source web application for urban planners to simulate the impact on temperature of increasing urban green infrastructures.

According to experts in the domain, the use of the tool in different cities would have a global impact on mitigating the UHI effect in cities, and therefore on improving urban sustainability and on the fight against climate change. The main limitation is the dependency of the model accuracy on the geospatial images used to train it. For example, the dates of the images are a key factor, given that some features (i.e., quantity of vegetation or its state) are dependent on the season and climatic conditions. If the intent is studying the impact on extreme high temperatures, the model should be trained with images acquired in the warm season. Also, the availability of certain type of images should be considered. For instance, LiDAR information is often limited to one or two specific dates for a certain city, so all the rest of images used should correspond to similar dates.

In conclusion, the proposed methodology and tools have the potential to facilitate urban planners' work by providing more accurate predictions of LST and simulating the impact of adding urban green infrastructure. As such, it represents a valuable contribution to the development of sustainable urban environments.

Future work could explore the integration of other factors, such as air pollution, to further improve the accuracy of temperature predictions and the impact of urban green infrastructure on urban environmental quality.

Some other parameters, like cost, available space, look and feel of the place or time to implement the project could be included in multi-criteria decision-making. Another improvement that experts consider important, and on which we are working, is the automatic identification of the most promising areas in which to apply the desired changes. This will be achieved by using more information, from new sources, concerning the viability of installing urban green infrastructure. In addition, visualisations will be generated which will enrich the

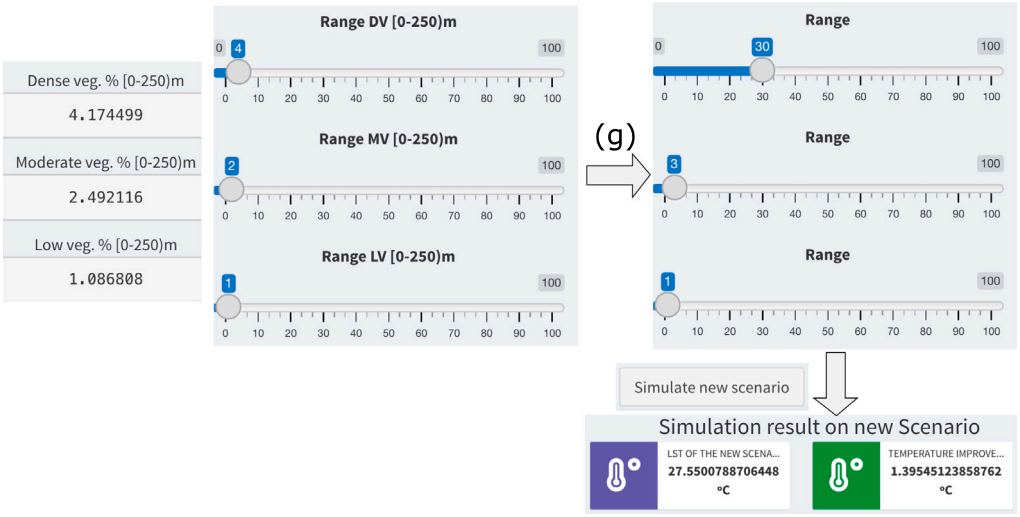


Fig. 6. Screenshots with part of the process to simulate new scenarios (g). The new LST value is estimated by changing some variables in the near-environment of the study point.

application and allow the final users to gain a deeper understanding of the forecast results and of the influence of the different factors in a given scenario.

CRediT authorship contribution statement

Francisco Rodríguez-Gómez: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Data curation. **José del Campo-Ávila:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Conceptualization. **Luis Pérez-Urrestarazu:** Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Funding acquisition, Conceptualization. **Domingo López-Rodríguez:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization.

Software and data availability

The main technologies used to develop the proposed tool are based in the R language (R Core Team, 2020) and some R libraries which are free under public licenses. The URSUS-LST is available in three different ways: (1) script for feature extraction of near urban environment of study points, (2) script for model training, and (3) web app for LST predictions and simulations

Software name: URSUS-LST

First year available: 2024

Software requirements: R and `RStudio`

Source code availability: https://github.com/ursusdm/URSUS_LST_prediction

License: GNU General Public License v3.0

The analysed remote sensing data (Landsat −8 and Sentinel-2 satellite images) can be downloaded free of charge from Earth Explorer (USGS, Department of the Interior, U.S.A. http://earthexplorer.usgs.gov). LiDAR images are freely available from Centro Nacional de Información Geográfica (CNIG, Spain, https://centrodedescargas.cnig.es).

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.

Acknowledgements

This work has been supported by the project RTI2018-095097-B-I00 at the 2018 call for I+D+i Project of the Ministerio de Ciencia, Innovación *y* Universidades, Spain. Funding for open access charge: Universidad de Málaga/CBUA.

Data availability

Landsat-8 and Sentinel-2 satellite images are freely available from http://earthexplorer.usgs.gov. LiDAR images are freely available from https://centrodedescargas.cnig.es.

References

American Society for Photogrammetry & Remote Sensing, 2011. LAS Specification version 1.4 - R13. URL https://www.asprs.org/wp-content/uploads/2010/12/LAS_1_4_r13.pdf.

Awad, M., Khanna, R., 2015. Support vector regression. In: Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. A Press, Berkeley, CA, pp. 67–80. http://dx.doi.org/10.1007/978-1-4302-5990-9_4.

Bartesaghi-Koc, C., Osmond, P., Peters, A., 2019. Mapping and classifying green infrastructure typologies for climate-related studies based on remote sensing data. Urban For. Urban Green. 37, 154–167. http://dx.doi.org/10.1016/j.ufug.2018.11.008.

Bentéjac, C., Csörgő, A., Martínez-Muñoz, G., 2020. A comparative analysis of gradient boosting algorithms. Artif. Intell. Rev. 54 (3), 1937–1967. http://dx.doi.org/10.1007/s10462-020-09896-5.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A., Deng, D., Lindauer, M., 2023. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. WIREs Data Min. Knowl. Discov. e1484. http://dx.doi.org/10.1002/widm.1484.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140. http://dx.doi.org/10.1007/BF00058655.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 2017. Classification and Regression Trees. Routledge, Monterey, CA, http://dx.doi.org/10.1201/9781315139470.

Chang, W., Cheng, J., Allaire, J.J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., Borges, B., 2021. shiny: Web Application Framework for R. URL https://cran.r-project.org/package=shiny.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. CRISP-DM 1.0. pp. 1–76.

Cherchyk, L., Khumarova, N., 2023. Green infrastructure management of urban ecosystems. Econ. Innov. 25, 142–151. http://dx.doi.org/10.31520/ei.2023.25.1(86).142-151.

Cortes, C., Vapnik, V., Saitta, L., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. http://dx.doi.org/10.1007/BF00994018.

Dutta, K., Basu, D., Agrawal, S., 2021. Synergetic interaction between spatial land cover dynamics and expanding urban heat islands. Environ. Monit. Assess. 193, 1–22. http://dx.doi.org/10.1007/S10661-021-08969-4.

El-Hattab, M., Amany, S.M., Lamia, G.E., 2018. Monitoring and assessment of urban heat islands over the Southern region of Cairo Governorate, Egypt. Egypt. J. Remote. Sens. Space Sci. 21, 311–323. http://dx.doi.org/10.1016/j.ejrs.2017.08.008.

European Commission, 2013. Building a Green Infraestructure for Europe. p. 24. http://dx.doi.org/10.2779/54125.

European Commission, 2014. E.U. copernicus. URL https://www.copernicus.eu. (date Accessed: 01 December 2024).

Fusami, A.A., Nweze, O.C., Hassan, R., 2020. Comparing the Effect of Deforestation Result by NDVI and SAVI. Int. J. Sci. Res. Publ. (IJSRP) 10 (06), 918–925. http://dx.doi.org/10.29322/ijsrp.10.06.2020.p102110.

Goggins, W.B., Chan, E.Y., Ng, E., Ren, C., Chen, L., 2012. Effect modification of the association between short-term meteorological factors and mortality by urban heat islands in Hong Kong. PLoS One 7 (6), 9–14. http://dx.doi.org/10.1371/journal.pone.0038551.

Hart, M.a., Sailor, D.J., 2009. Quantifying the influence of land-use and surface characteristics on spatial variability in the urban heat island. Theor. Appl. Climatol. 95 (3–4), 397–406. http://dx.doi.org/10.1007/s00704-008-0017-5.

He, L., Ren, X., Gao, Q., Zhao, X., Yao, B., Chao, Y., 2017. The connected-component labeling problem: A review of state-of-the-art algorithms. Pattern Recognit. 70, 25–43. http://dx.doi.org/10.1016/j.patcog.2017.04.018.

Heaviside, C., Vardoulakis, S., Cai, X.M., 2016. Attribution of mortality to the urban heat island during heatwaves in the West Midlands, UK. Environ. Health: A Glob. Access Sci. Source 15 (Suppl 1), http://dx.doi.org/10.1186/s12940-016-0100-9.

Herrera, M., Ramallo-González, A.P., Eames, M., Ferreira, A.A., Coley, D.A., 2018. Creating extreme weather time series through a quantile regression ensemble. Environ. Model. Softw. 110, 28–37. http://dx.doi.org/10.1016/j.envsoft.2018.03.007.

Herrera-Gomez, S.S., Quevedo-Nolasco, A., Pérez-Urrestarazu, L., 2017. The role of green roofs in climate change mitigation. A case study in Seville (Spain). Build. Environ. 123, 575–584. http://dx.doi.org/10.1016/j.buildenv.2017.07.036.

Hijmans, R.J., 2025. raster: Geographic Data Analysis and Modeling. In: R package version 3.6-31. https://rspatial.org/raster.

Jiang, Z., Huete, A.R., Chen, J., Chen, Y., Li, J., Yan, G., Zhang, X., 2006. Analysis of NDVI and scaled difference vegetation index retrievals of vegetation fraction. Remote Sens. Environ. 101, 366–378. http://dx.doi.org/10.1016/J.RSE.2006.01.003.

Kajosaari, A., Hasanzadeh, K., Fagerholm, N., Nummi, P., Kuusisto-Hjort, P., Kyttä, M., 2024. Predicting context-sensitive urban green space quality to support urban green infrastructure planning. Landsc. Urban Plan. 242, 104952. http://dx.doi.org/10.1016/j.landurbplan.2023.104952.

Karimi, A., Ghajari, Y.E., 2022. Improving land surface temperature prediction using spatiotemporal factors through a genetic-based selection procedure (Case Study: Tehran, Iran). Adv. Space Res. 69 (9), 3258–3267. http://dx.doi.org/10.1016/j.asr.2022.02.004.

Kartal, S., Sekertekin, A., 2022. Prediction of MODIS land surface temperature using new hybrid models based on spatial interpolation techniques and deep learning models. Env. Sci. Pollut. Res. Int. 29 (44), 67115–67134. http://dx.doi.org/10.1007/s11356-022-20572-9.

Kesavan, R., Muthian, M., Sudalaimuthu, K., Sundarsingh, S., Krishnan, S., 2021. ARIMA modeling for forecasting land surface temperature and determination of urban heat island using remote sensing techniques for Chennai city, India. Arab. J. Geosci. 14 (11), 1016. http://dx.doi.org/10.1007/s12517-021-07351-5.

Khalil, U., Aslam, B., Azam, U., Khalid, H.M.D., 2021. Time series analysis of land surface temperature and drivers of urban heat island effect based on remotely sensed data to develop a prediction model. Appl. Artif. Intell. 35 (15), 1803–1828. http://dx.doi.org/10.1080/08839514.2021.1993633.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. Artic. 28 (5), 1–26. http://dx.doi.org/10.18637/jss.v028.i05.

Li, Y., Li, T., Liu, H., 2017. Recent advances in feature selection and its applications. Knowl. Inf. Syst. 53 (3), 551–577. http://dx.doi.org/10.1007/s10115-017-1059-8.

Liao, S.H., Chu, P.H., Hsiao, P.Y., 2012. Data mining techniques and applications – A decade review from 2000 to 2011. Expert Syst. Appl. 39 (12), 11303–11311. http://dx.doi.org/10.1016/j.eswa.2012.02.063.

Lin, B.S., Yu, C.C., Su, A.T., Lin, Y.J., 2013. Impact of climatic conditions on the thermal effectiveness of an extensive green roof. Build. Environ. 67, 26–33. http://dx.doi.org/10.1016/j.buildenv.2013.04.026.

Liu, Q., Wu, Y., 2012. Supervised learning. In: Encyclopedia of the Sciences of Learning. Springer US, Boston, MA, pp. 3243–3245. http://dx.doi.org/10.1007/978-1-4419-1428-6_451.

Mejbel Salih, M., Zakariya Jasim, O., I. Hassoon, K., Jameel Abdalkadhum, A., 2018. Land surface temperature retrieval from LANDSAT-8 thermal infrared sensor data and validation with infrared thermometer camera. Int. J. Eng. Technol. 7 (4.20), 608. http://dx.doi.org/10.14419/ijet.v7i4.20.27402.

Mohamed, A.A., Odindi, J., Mutanga, O., 2017. Land surface temperature and emissivity estimation for Urban Heat Island assessment using medium- and low-resolution space-borne sensors: A review. Geocarto Int. 32, 455–470. http://dx.doi.org/10.1080/10106049.2016.1155657.

Muñoz, L.S., Duarte, D.H.S., 2025. Green infrastructure as a planning tool: A comprehensive systematization of urban redesign strategies to increase vegetation within public places. Cities 156, 105551. http://dx.doi.org/10.1016/j.cities.2024.105551.

Murtagh, F., 1991. Multilayer perceptrons for classification and regression. Neurocomputing 2 (5–6), 183–197.

Mustafa, E.K., Co, Y., Liu, G., Kaloop, M.R., Beshr, A.A., Zarzoura, F., Sadek, M., 2020. Study for predicting land surface temperature (LST) using landsat data: a comparison of four algorithms. Adv. Civ. Eng. 2020, 7363546. http://dx.doi.org/10.1155/2020/7363546.

Nesticò, A., Passaro, R., Maselli, G., Somma, P., 2022. Multi-criteria methods for the optimal localization of urban green areas. J. Clean. Prod. 374, 133690. http://dx.doi.org/10.1016/J.JCLEPRO.2022.133690.

Oijstaeijen, W.V., Passel, S.V., Cools, J., 2020. Urban green infrastructure: A review on valuation toolkits from an urban planning perspective. J. Environ. Manag. 267, 110603. http://dx.doi.org/10.1016/j.jenvman.2020.110603.

Ponsard, C., Touzani, M., Majchrowski, A., 2017. Combining process guidance and industrial feedback for successfully deploying big data projects. Open J. Big Data 3 (1), 26–41.

Quinlan, J.R., 2014. C4. 5: Programs for Machine Learning. Elsevier.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.r-project.org/.

Rahaman, Z.A., Kafy, A.A., Saha, M., Rahim, A.A., Almulhim, A.I., Rahaman, S.N., Fattah, M.A., Rahman, M.T., S, K., Faisal, A.A., Al Rakib, A., 2022. Assessing the impacts of vegetation cover loss on surface temperature, urban heat island and carbon emission in Penang city, Malaysia. Build. Environ. 222, 109335. http://dx.doi.org/10.1016/J.BUILDENV.2022.109335.

Ramiya, A.M., Nidamanuri, R.R., Krishnan, R., 2017. Segmentation based building detection approach from LiDAR point cloud. Egypt. J. Remote. Sens. Space Sci. 20 (1), 71–77. http://dx.doi.org/10.1016/j.ejrs.2016.04.001.

Rao, B.S., Kumar, G.A., Runjhun, C., Rao, C.V.K.V.P.J., Babu, G.V., 2022. Improvement of airborne LiDAR intensity image content with shaded nDSM and assessment of its utility in geospatial data generation. J. Indian Soc. Remote. Sens. 50 (3), 507–521. http://dx.doi.org/10.1007/s12524-021-01468-6.

Rodríguez-Gómez, F., del Campo-Ávila, J., Ferrer-Cuesta, M., Mora-López, L., 2022b. Data driven tools to assess the location of photovoltaic facilities in urban areas. Expert Syst. Appl. 203, 117349. http://dx.doi.org/10.1016/j.eswa.2022.117349.

Rodríguez-Gómez, F., Fernández-Cañero, R., Pérez, G., del Campo-Ávila, J., López-Rodríguez, D., Pérez-Urrestarazu, L., 2022a. Detection of unfavourable urban areas with higher temperatures and lack of green spaces using satellite imagery in sixteen Spanish cities. Urban For. Urban Green. 78, 127783. http://dx.doi.org/10.1016/J.UFUG.2022.127783.

Roussel, J.R., Auty, D., 2021. Airborne LiDAR data manipulation and visualization for forestry applications. URL https://cran.r-project.org/package=lidR.

Roussel, J.R., Auty, D., Coops, N.C., Tompalski, P., Goodbody, T.R., Meador, A.S., Bourdon, J.F., de Boissieu, F., Achim, A., 2020. lidR: An R package for analysis of Airborne Laser Scanning (ALS) data. Remote Sens. Environ. 251, 112061. http://dx.doi.org/10.1016/j.rse.2020.112061.

Schlink, U., Mohamdeen, A., Raabe, A., 2020. Temporal modes and spatial patterns of urban air temperatures and limitations of heat adaptation. Environ. Model. Softw. 132, 104773. http://dx.doi.org/10.1016/j.envsoft.2020.104773.

Sharma, M., Garg, R.D., Badenko, V., Fedotov, A., Min, L., Yao, A., 2021. Potential of airborne LiDAR data for terrain parameters extraction. Quat. Int. 575–576, 317–327. http://dx.doi.org/10.1016/j.quaint.2020.07.039.

Susca, T., Gaffin, S.R., Dell'osso, G.R., 2011. Positive effects of vegetation: urban heat island and green roofs. Environ. Pollut. 159 (8–9), 2119–2126. http://dx.doi.org/10.1016/j.envpol.2011.03.007.

Tóth, A., Halajová, D., Halaj, P., 2015. Green infrastructure: a strategic tool for climate change mitigation in urban environments. Ecol. Saf. 9, 132–138.

Wickham, H., François, R., Henry, L., Müller, K., 2020. dplyr: A grammar of data manipulation. URL https://cran.r-project.org/package=dplyr.

Wu, W.B., Ma, J., Banzhaf, E., Meadows, M.E., Yu, Z.W., Guo, F.X., Sengupta, D., Cai, X.X., Zhao, B., 2023. A first Chinese building height estimate at 10 m resolution (CNBH-10 m) using multi-source earth observations and machine learning. Remote Sens. Environ. 291, 113578. http://dx.doi.org/10.1016/J.RSE.2023.113578.

Wyner, A.J., Olson, M., Bleich, J., Mease, D., 2017. Explaining the success of adaboost and random forests as interpolating classifiers. J. Mach. Learn. Res. 18 (1), 1558–1590, URL http://jmlr.org/papers/v18/15-240.html.

Zumwald, M., Baumberger, C., Bresch, D.N., Knutti, R., 2021. Assessing the representational accuracy of data-driven models: The case of the effect of urban green infrastructure on temperature. Environ. Model. Softw. 141, 105048. http://dx.doi.org/10.1016/J.ENVSOFT.2021.105048.