Expert Systems With Applications 203 (2022) 117349

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Data driven tools to assess the location of photovoltaic facilities in urban areas

Francisco Rodríguez-Gómez, José del Campo-Ávila [*], Marta Ferrer-Cuesta, Llanos Mora-López

Universidad de Málaga, Departamento de Lenguajes y Ciencias de la Computación, Campus de Teatinos, 29071 Málaga, Spain

ARTICLE INFO

Keywords:
Renewable energy
Photovoltaic systems
LiDAR images
Semantic segmentation
Roof feature extraction
Energy forecast

ABSTRACT

Urban sustainability is a significant factor in combating climate change. Replacing polluting by renewable energies is fundamental to reduce the emission of greenhouse gases. Photovoltaic (PV) facilities harnessing solar energy, and particularly self-consumption PV facilities, can be widely used in cities throughout most countries. Therefore, locating spaces where photovoltaic installations can be integrated into urban areas is essential to reduce climate change and improve urban sustainability. An open-source software (URSUS-PV) to aid decision-making regarding possible optimal locations for photovoltaic panel installations in cities is presented in this paper. URSUS-PV is the result of a data mining process, and it can extract the characteristics of the roofs (orientation, inclination, latitude, longitude, area) in the urban areas of interest. By combining this information with meteorological data and characteristics of the photovoltaic systems, the system can predict both the next-day hourly photovoltaic energy production and the long-term photovoltaic daily average energy production.

1. Introduction

Cities have become a determining factor in climate change, as they are the place where much energy is consumed (64% of global primary energy use) and high levels of greenhouse gases emitted (70% of the global total), due to the use of fossil fuels as energy sources (International Energy Agency, 2016). There is a great opportunity for citizens to reduce these emissions. Replacing polluting energy sources by renewable energies that respect the environment and do not compromise future generations is one of the essential requirements to achieve energy-sustainable cities and favour the fight against climate change. Additionally, switching to renewable energy sources as a detriment to polluting energies could improve health and quality of life. Precisely, one of the goals proposed in the 2030 Agenda for Sustainable Development by UN is *making cities inclusive, safe, resilient and sustainable* (United Nations, 2015).

Solar energy has seen a large increase among renewable energies. According to the International Energy Agency (IEA), there was a 22% growth up to 720 TWh (representing 3% of global electricity generation) in 2019 (International Energy Agency, 2020). Although large photovoltaic infrastructures are away from cities, there has been an exponential rise in distributed installations in buildings, industry and houses in Europe, the United States and Japan (International Energy Agency, 2020). This is very important since local production reduces transportation losses and enhance citizen's responsibility because of inspiration for searching for energy self-sufficiency. In recent years, a new type of building, based on that type of installation, has been proposed as an evolution of *Zero Energy Building* (*ZEB*): the *Positive Energy Building* (*PEB*) (Magrini et al., 2020).

Land availability in urban areas is limited. Roofs are estimated to account for between 20% and 25% of urban surface (Akbari & Rose, 2008). They are therefore an excellent resource to be exploited by installing photovoltaic systems. Having tools that help in decision making to increase an area's energy production (neighbourhood or building complex) would be beneficial in order to pinpoint the most appropriate places for such an intervention. Predictions about the expected energy production in the long-term are also very convenient to assess the suitability of installing these infrastructures. The same occurs with short-term estimations of photovoltaic production. These predictions can help owners of self-consumption installations with better load management (passive or active) and managers of extensive facilities to improve their integration into the power grid.

Collecting data from urban areas and conducting a data mining process are an appropriate way to develop the necessary models and tools. Numerous data sources can currently be used in the design of intelligent decision support systems in the scope of that work. These

[*] Corresponding author.
E-mail addresses: francisco.rdg.gmz@uma.es (F. Rodríguez-Gómez), jcampo@uma.es (J. del Campo-Ávila), martafecu@gmail.com (M. Ferrer-Cuesta), llanos@uma.es (L. Mora-López).

https://doi.org/10.1016/j.eswa.2022.117349
Received 17 May 2021; Received in revised form 6 March 2022; Accepted 25 April 2022
Available online 5 May 2022

include: (a) maps of the cities showing the location of buildings and vegetation; (b) LiDAR images (3D point clouds) that allow height models of urban objects to be determined (Sharma et al., 2021); and (c) information provided by Meteorological Agencies, including radiation, temperature and precipitation values.

In addition to information sources, a data mining process, with algorithms and tools to automate processing data tasks and discover new knowledge, needs to be correctly conducted. Thus, in the geo-computational field, information can be obtained from urban images (in two or three dimensions) on, for example, the location of the available roofs, their orientation and inclination (also known as aspects and slopes) or their sizes. Such information, conveniently combined with other sources, can be used to predict new variables that will affect the produced energy calculation. Moreover, learning patterns can be established to model the problem. Artificial intelligence tools, such as machine learning algorithms, will be needed to carry out those tasks. Extracted knowledge can then be used and channelled into software tools to help the installation potential of an area for photovoltaic systems in urban areas.

This research's main objective is to generate knowledge to evaluate the expected energy production by photovoltaic infrastructures installed in urban areas. This evaluation will allow the potential energy production to be established for the long and short-term, depending on the usable surface in each area and the orientation, inclination and size of the different roofs and rooftops. A methodology based on a data mining process is proposed to achieve this objective. The knowledge from the last phase of the data mining process (deployment phase) can enhance the experts' skills. It has been incorporated into an intelligent decision support system.

The methodology developed and its implementation are of interest to both private users and local administrations, corporations or neighbourhood communities. Thanks to that methodology, zones can be delimited in cities or urban areas and the potential energy production by the photovoltaic system known for the long and short time. This knowledge may be of particular interest to public administrations or large corporations with limited resources to carry out a series of installations and which have to choose the most suitable locations from.

The rest of the paper is organized as follows. Section 2 describes the data mining methodology selected for the development of the tool, along with the tool itself and the technology used for its development. Section 3 describes the results obtained using the tool in an urban area of the city of Málaga (Spain). Section 4 concludes.

2. Methods

Building software that can help experts in their decision making is essential nowadays. Knowledge extracted from different data sources can be used in its development and additional intelligent characteristics added. A data mining process is the most systematic way of achieving worthy results that transfer knowledge to real applications.

Fig. 1 schematically shows the main stages that are defined in CRISP-DM methodology (Chapman et al., 2000), one of the most widely used and extended for data mining projects. Section 2.1 presents how all phases have been developed to implement URSUS-PV with such methodology. Some stages require more focus and they are described in detail in successive points (from 2.1.1 to 2.1.4).

Finally, Section 2.2 describes different technologies used to implement the software presented in this paper.

2.1. URSUS-PV: Tool for estimating the solar energy produced in areas of interest

The development of URSUS-PV has followed different phases. Some of them are simple to be described, such as Phases 1 or 2, but some others require more detailed explanations that are presented separately. Fig. 2 shows a diagram where the phases are schematically represented.

Table 1
Dataset description.

Meteorological data (long term)	LiDAR image
Radiation and clearness index (hourly)	Resolution (0.5 point/m^2)
Temperature (hourly)	Size (2 × 2 km)
Meteorological data (short term)	
Radiation and clearness index (daily)	
Temperature (daily and 2 short consecutive periods during the day)	
Humidity (daily and 2 short consecutive periods during the day)	
Prediction for temperature (daily and 2 short consecutive periods during the day)	
Prediction for humidity (daily and 2 short consecutive periods during the day)	
Prediction for cloudiness (daily and 2 short consecutive periods during the day)	
Photovoltaic facility data (PV)	
Inclination (slope)	Latitude
Orientation	Longitude
Surface (available)	

Business understanding (Phase 1) established the objective for this tool, namely, to evaluate the potential energy production that can be generated by photovoltaic infrastructures installed in urban areas. Estimations must be performed from a double long and short-term perspective.

During the data understanding step (Phase 2), we worked with both meteorological data and LiDAR aerial images of the urban terrain covering the entire geographic area of interest for the study. Meteorological data are of high enough quality when they provide measurements (such temperature or humidity) with daily updates (even hourly) and predictions for the following day (such as cloudiness). LiDAR images are usually available for most cities and they facilitate the semantic segmentation of urban objects, making it easier to focus on the rooftops. In such segmentation, every point in the image is automatically assigned to a specific class (such as building, vehicles or vegetation). LiDAR images also provide models of the heights of urban objects, which allows the roof orientations and inclinations in areas of interest to be obtained. A description of the data used is given in Table 1.

Data integration and their preparation (Phase 3) mainly consists of two steps that have been automated to great extent: selection of the area of interest (see 2.1.1) and roof segmentation for feature extraction (see 2.1.2).

In the modelling phase (Phase 4), two different models are needed to satisfy the data mining goals established in the first phase: one model for long-term predictions and another for short-term predictions. Although the second one is taken from an existing paper (del Campo-Ávila et al., 2021), both are described in 2.1.3.

The evaluation (Phase 5) was conducted by checking the results obtained by our system for specific points in different areas with the results obtained by other systems. Such other systems use manual processes to calculate the photovoltaic production for a specific system. A correct evaluation allows the models generated in the previous step (Phase 4) to be integrated in the final product (Phase 6).

As regards deployment (Phase 6), the tool can estimate the solar energy produced in areas of interest inside urban zones in the long and short-term. This is the result of integrating a photovoltaic energy estimation model (long-term) with an existing hourly solar radiation predictor system (short-term). This integration is not aimed at improving accuracy, but rather at increasing functionality: long term prediction allows users to assess the suitability of a location to install a new facility, while short term prediction allows users to decide on the expected performance for the next day. In 2.1.4, details about the integration of the results achieved in previous phases are given.

2.1.1. Selection of the area of interest

Studying all the possible locations for photovoltaic systems in a city is unusual because users are mainly interested in specific and

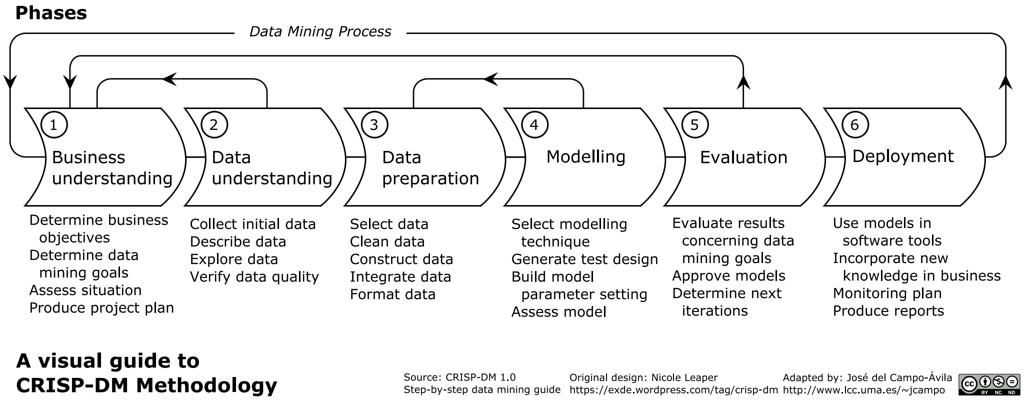

Phases
Data Mining Process
1 Business understanding
2 Data understanding
3 Data preparation
4 Modelling
5 Evaluation
6 Deployment
Determine business objectives
Determine data mining goals
Assess situation
Produce project plan
Collect initial data
Describe data
Explore data
Verify data quality
Select data
Clean data
Construct data
Integrate data
Format data
Select modelling technique
Generate test design
Build model parameter setting
Assess model
Evaluate results concerning data mining goals
Approve models
Determine next iterations
Use models in software tools
Incorporate new knowledge in business
Monitoring plan
Produce reports
A visual guide to CRISP-DM Methodology
Source: CRISP-DM 1.0 Step-by-step data mining guide
Original design: Nicole Leaper https://exde.wordpress.com/tag/crisp-dm
Adapted by: José del Campo-Ávila http://www.lcc.uma.es/~jcampo

Fig. 1. CRISP-DM methodology outline. Six phases are defined, and most relevant generic tasks are enumerated under each phase.

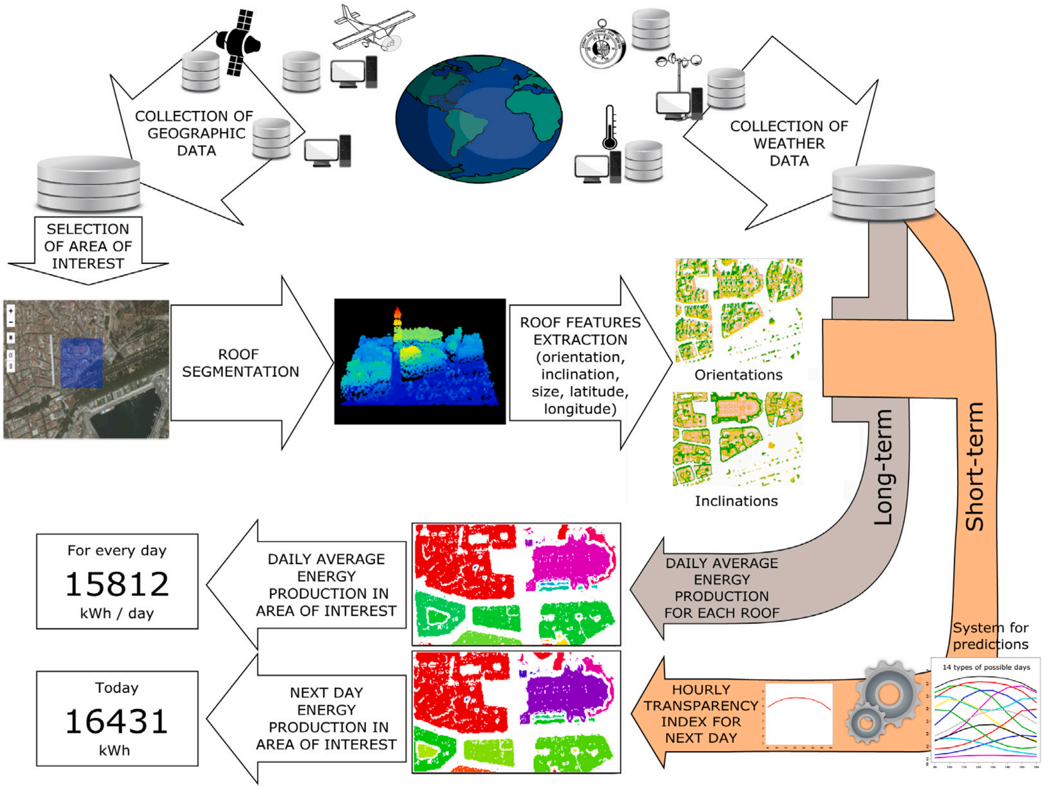

COLLECTION OF GEOGRAPHIC DATA
COLLECTION OF WEATHER DATA
SELECTION OF AREA OF INTEREST
ROOF SEGMENTATION
ROOF FEATURES EXTRACTION (orientation, inclination, size, latitude, longitude)
Orientations
Inclinations
Long-term
Short-term
For every day
15812
kWh / day
DAILY AVERAGE ENERGY PRODUCTION IN AREA OF INTEREST
DAILY AVERAGE ENERGY PRODUCTION FOR EACH ROOF
System for predictions
14 types of possible days
Today
16431
kWh
NEXT DAY ENERGY PRODUCTION IN AREA OF INTEREST
HOURLY TRANSPARENCY INDEX FOR NEXT DAY

Fig. 2. Scheme of the data understanding and preparation phases that uses some of the models generated allow the potential photovoltaic energy produced in an urban area to be estimated.

delimited areas. That behaviour can benefit the system performance as there are fewer computational requirements, while maintaining the same functionality. Therefore this data selection is necessary and must be automated.

The system accesses the geographic information that was previously downloaded and create a map with a grid that delimits the areas available for energy estimation. It allows the user to select the area of interest by defining a polygon on the map. The system then loads the image covering the area delimited by the user's coordinates through the polygon on the map. Subsequent processing steps only take into account the data filtered in this step.

In addition to the geographic information, meteorological data captured in the city or close to the city are needed. However, the proximity to the area of interest is enough and information from the closest weather stations is used.

2.1.2. Image processing for roof segmentation and feature extraction

Two aspects are involved in locating roofs where photovoltaic infrastructures can be installed: first, the detection of roofs in a city and, second, the determination of their characteristics to generate an accurate assessment.

The first step processes a 3D LiDAR image covering the area of interest of the city selected by the user. All urban elements such as vegetation or vehicles are removed, keeping only the ground and rooftops.

In the next step, the normalized height model (CHM or NDSM) is obtained by eliminating the ground height. In this 2D model, the height of the objects is measured at the same level. After this transformation, the system can determine the available roofs and the heights of the different buildings detected. For more details about roof and ground segmentation, and CHM calculation (functions and libraries used), see 3.1.1.

At this point, the system allows the user to select different ranges of interest for different slope and orientation values. For more details see 3.1.2.

A raster layer with the connected components of the roofs can then be created. Each connected component consists of contiguous pixels of roofs that meet the user's criteria for slopes and orientations. For energy calculations, when talking about a roof, it actually means the connected component of the roof. For more details see Section 3.1.3.

The next step is to calculate its mean inclination, orientation, size and latitude for every connected component. All these data will be required to estimate new values that allow the generation of new information to be used in the software proposed. More details about such estimations and the induction of models to make predictions are given below.

2.1.3. Calculation of solar energy potential

The estimation of energy received in each roof depending on the size, orientation, inclination and location, can be modelled by using the expressions proposed by A. Coronas (1983) and Iqbal (1983). In this phase, the actual inclination of roofs has been used. It would be also possible to estimate the optimal inclination, for instance using the proposal made by Chang (2010), but it will require the individual evaluation of the integration of photovoltaic facilities (PV) into the buildings envelope.

The energy that a photovoltaic system could produce in the long-term is estimated using the model proposed by Osterwald (1986), taking into account the information from the previous step. For this estimation, first, the hourly energy produced by each system is calculated using both meteorological data and data from the roof. The meteorological data used are hourly global solar radiation and hourly temperature. The data of the roof are the size, latitude, longitude, inclination, and orientation. The power generated by the system is estimated using the expression:

$$P = P_{STC} \frac{G_{\beta,\alpha}}{1000}(1 + \gamma(T_{mod} - 25))GL$$

where P_{STC} is the power of the system in standard conditions, $G_{\beta,\alpha}$ is the global irradiance on the surface of the modules, β is the inclination of the modules, α is the orientation, γ is the temperature coefficient of the maximum power of the modules, T_{mod} is the module temperature, and GL is the global losses coefficient of the system. This expression has been obtained from the expression proposed by Osterwald (1986) and includes not only the losses produced by temperature but also other losses (soiling, spectral losses, and so on). The daily radiation is calculated as the sum of the estimated hourly values.

Therefore, the daily average energy production can be estimated for every roof, and the estimation for the roofs in a desired area can be calculated by aggregation.

On the other hand, in a short-term scenario, the photovoltaic energy produced can be estimated one-day-ahead. Using the model proposed by del Campo-Ávila et al. (2021) the next-day prediction of hourly solar radiation is calculated. The RMSE of this model is 97 $\mathrm{Wh/m^2}$ and 63 $\mathrm{Wh/m} * 2$ that are similar to the errors estimated using other data mining models such as in Cannizzaro et al. (2021) where are 107 and 58 respectively. This model uses as independent variables most of the significant input variables selected in the proposal by Castangia et al. (2021). This model takes as input meteorological data registered during the current day and certain meteorological forecasts, and predicts the type of radiation expected for the following day (sunny, cloudy, partially sunny, etc.). The model selects from 14 types of radiation, that were identified in the induction phase of the model, and provides the hourly transparency index expected for each hour between 9:00 and 16:00 for the following day. The prediction process is summarized on the right side of Fig. 2. As it is explained in del Campo-Ávila et al. (2021) using the meteorological data available for a day, the developed system responds with the type of day estimated for the next day. The hourly global solar radiation is estimated using the centroid that contains the hourly values of clearness index for that day, and the values of hourly extraterrestrial radiation. This conveniently transformed estimation gives the potential solar energy calculated for next day for every roof, and, by aggregation, for the area of interest.

2.1.4. Integration and operation

In addition to the image processing model and extraction of the characteristics of the roofs, photovoltaic energy calculation models, and the previously presented predictor system are needed for the system to operate correctly, along with a series of scripts for automating daily downloads of radiation, meteorological and forecast data.

Although the processing of images can be applied to all the elements, the user could select the characteristics of interest (like a specific orientation or the maximum or minimum inclination) to reduce the computational effort. The estimated energy is calculated for each roof that meets the user's requirements (daily average for long-term prediction or hourly energy for short-term prediction). Their aggregated calculation will then constitute the estimated photovoltaic energy in the urban area of interest.

The tool's general operating scheme with the integration is shown in Fig. 3.

The general operation of the system is now presented. The user interacts with the system by selecting the preferences, identified with upper-case letters in this description:

A. Selection of the area of interest
    A.1. System performs the roof segmentation.

B. Selection of roofs with desired orientation and inclination
    B.1. System extracts roof's features (size, latitude, longitude, mean orientation, mean inclination).

C. Selection between long-term or short-term solar energy calculation:
    C.1. Long-term:
        C.1.1. Calculate daily average solar energy for each roof.
        C.1.2. Calculate total daily average solar energy in the area.
    C.2. Short-term:
        C.2.1. Obtain data from meteorological observations, radiation and hourly forecasts that are downloaded daily. If meteorological information system offers an API, automatization is easier.
        C.2.2. Calculate the radiation expected for one-day-ahead using a prediction system that analyse previous data. For every roof, short-term photovoltaic energy estimation is performed.
        C.2.3. Calculate next-day solar energy expected to be produced in the area of interest.

URSUS-PV is offered as a free web tool available to individuals, municipalities, neighbourhood communities or companies in the photovoltaic sector (see *Code Availability* Section). It is also supplied as a package that can be altered to be adapted to specific scenarios or improved with new features.

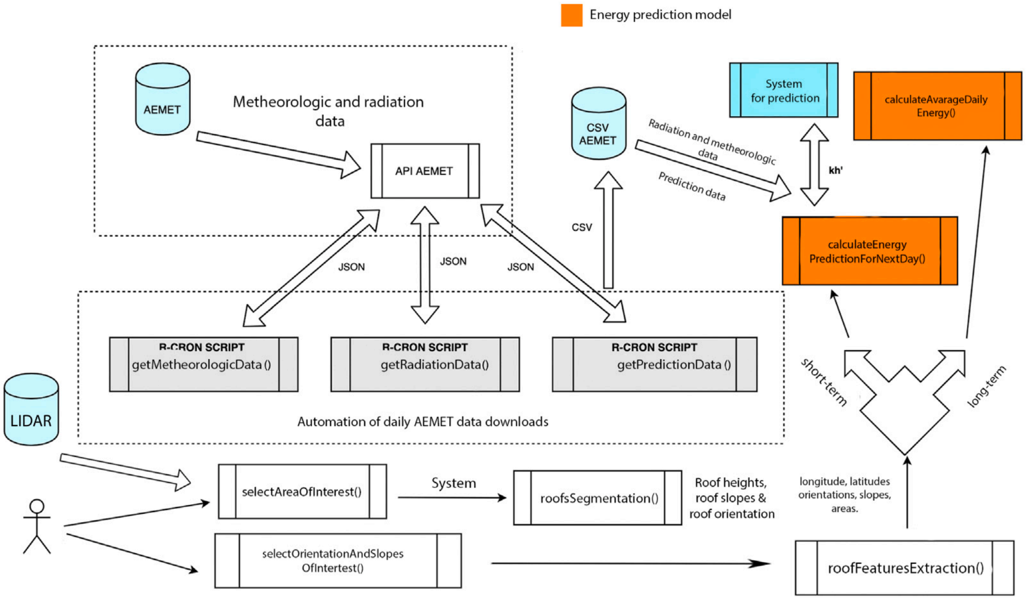

Energy prediction model
AEMET
Metheorologic and radiation data
API AEMET
CSV AEMET
Radiation and metheorologic data
Prediction data
System for prediction
kh'
calculateAvarageDaily Energy()
calculateEnergy PredictionForNextDay()
CSV
JSON
JSON
JSON
R-CRON SCRIPT
getMetheorologicData ()
R-CRON SCRIPT
getRadiationData()
R-CRON SCRIPT
getPredictionData ()
Automation of daily AEMET data downloads
short-term
long-term
LIDAR
selectAreaOfInterest()
System
roofsSegmentation()
Roof heights, roof slopes & roof orientation
longitude, latitudes orientations, slopes, areas.
selectOrientationAndSlopes OfIntertest()
roofFeaturesExtraction()

Fig. 3. General diagram of the tool with its integrations.

2.2. Technologies for implementation

Some of the tools now most commonly used to analyse and process urban images are the Geographic Information System (GIS) (Li et al., 2019). Some of the most popular tools used for this purpose are QGIS and ArcGIS. Image processing of this type with data science languages such as R (R Core Team, 2020) has recently been gaining momentum thanks to the development of libraries such as `lidR`, `rspatial` or `raster`. These include a large number of functions that greatly facilitate working with 2D and 3D urban images to obtain the necessary information by processing them.

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

In first phases, like data acquisition and transformation, `dplyr` package (Wickham et al., 2020) was used to obtain core knowledge. `cronR` package (Wijffels, 2020) has been used to automate the daily download of meteorological and radiation data.

For data preparation some other R packages were used like `rspatial` (Hijmans, 2018), `raster` (Hijmans, 2020) and `lidR` (Roussel & Auty, 2021; Roussel et al., 2020). `lidR` processes 3D LiDAR models and classifies the urban objects in it (buildings, ground, or vegetation) while `raster` allows the calculation of height, orientation, or inclination.

R libraries for Machine Learning such as `RWeka` package (Hornik et al., 2009), that integrates algorithms implemented in Weka (Witten et al., 2016), or `caret` package (Kuhn, 2020) have been used to automatize part of the process that creates predictive models.

Finally, the `shiny` package (Chang et al., 2021) offers a framework for the development of the DashBoard or web application for calculating photovoltaic energy in urban areas of interest. A Linux server has been configured for the app deployment.

3. Results

This section shows an example of an actual use case of the tool developed to estimate photovoltaic energy in an area of urban interest. The estimation was performed for the long-term (daily average production). The short-term (one-day-ahead) estimation would follow an identical process.

Comparisons to previous methods and tools are enumerated too at the end of this section. There, the advantages presented in the proposed system (URSUS-PV) are highlighted.

3.1. Validation example

Once URSUS-PV was implemented, it was used in a real scenario to test its capacities. We selected the city of Málaga, in Spain, and data were collected: (a) meteorological data for 10 years (from the Spanish Meteorological Agency, AEMET[1]), and (b) LiDAR aerial images of the urban terrain covering the entire geographic area of interest for the study (from National Geographic Information Center, CNIG[2]).

The usage of the application is described below.

3.1.1. Selection of the area of interest and roof segmentation

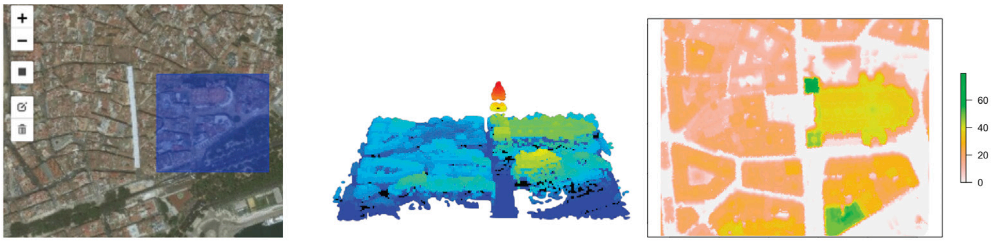
The first step was to select the urban area of interest for photovoltaic energy estimations. A polygon draw tool integrated into the map tool area was used to define that input. The map shows the LiDAR images available for that city as a overlapped grid. The user could zoom and scroll the city map to locate the area of interest (AOI). The system processed the LiDAR 3D image of the selected area, filtered out urban objects that were not of interest (water, vegetation), leaving only the ground and roofs or terraces. The `lidR` library includes some functions that allow the extraction of different information. The roofs and ground were calculated using `lasfilter`. The `lasNormalize` function was used to remove the ground. To get the 2D normalized roof height model `gridCanopy` was used. Finally, the system showed the roofs of the area of interest and the height of every roof pixel. Figs. 4(a), 4(b), 4(c) shows information calculated at different moments during the selection of AOI and its segmentation.

3.1.2. Filtering roofs with desired orientation and inclination

The user could select the orientation and the inclination (or slopes) of the potential roofs and rooftops for installing photovoltaic systems. The system allowed the user to select ranges of interest for slopes and orientations grouped from 10 to 10 degrees. Once the slope ranges of interest were selected, the system could display two raster layers; one with the slope value of each pixel of the roof, and another with the slope range of interest to which each pixel belongs. Fig. 5 shows UI for selected slope ranges. In this example, the user is interested in slope ranges: 1.[0,10), 2.[10,20), …, 7.[60–70]. Fig. 5(a) shows slope values

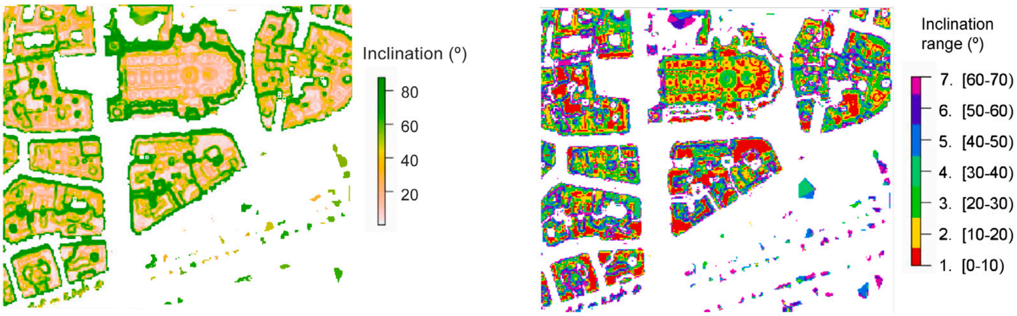
[1] http://www.aemet.es.

[2] http://centrodedescargas.cnig.es.



(a) (b) (c)

Fig. 4. Roof Segmentation: (a) Select the urban area of interest, (b) LiDAR 3D model with roof segmentation, (c) Normalized roof heights model.



(a) (b)

Fig. 5. Inclination selection: (a) Inclination values for each pixel, (b) User selected inclination range for each pixel.

for each pixel. Fig. 5(b) shows the discretized values corresponding to the slope ranges in which the user is interested. The process for orientation is the same.

The function `terrain` in the `raster` library calculates the orientations and slopes of the roofs of a normalized model.

3.1.3. *Photovoltaic energy estimations*

At this point, the user, depending on their objectives, could select whether to make short- or long-term estimations. In our example, presented in Fig. 6 the system made a long-term estimation. The first step was to calculate the connected components of the roofs. The connected components (CC) of the roofs are pieces composed of contiguous pixels of roofs that satisfy the orientation and inclination criteria. Each CC has a unique ID. A raster layer of connected components was created. The `raster` library function for generating the connected component raster layer is `clump(formask, directions=8)`, where `formask` is a raster layer with value 1 for pixels that satisfy the orientation and slope conditions (pixels that for the discretized layer of orientations and slopes have a value other than null). That means the user selected that range for the orientation or slope. The rest of the pixels of `formask` are identified as null values. From this point on, when we refer to a roof, we mean a connected roof component.

For each connected component of the roofs, the system calculated the mean latitude, mean orientation, mean inclination, and size (area in m^2). With the roof features extracted for every available roof, the necessary meteorological and radiation data of the municipality and the weather station closest to it was added. Using the photovoltaic system characteristics, the system determined the average daily energy (kWh) for each roof. Finally, the sum of the daily energy estimated for each roof would be the daily energy estimated for every day in the urban area. In the energy estimation panel (see Fig. 6), the tool displayed a map with the rooftops that met the orientation and inclination user's requirements. Additionally, two labels displayed the number of roofs that satisfied the requirements and the photovoltaic energy estimation for the urban area initially selected. The system also showed detailed information about the roofs processed.

3.2. *Comparison to previous methods*

In this section, the main models, technologies, and tools identified for estimating solar energy in cities are described. Finally, a comparative of these tools is shown in Table 2, in which the differences provided by URSUS-PV are revealed.

In Freitas et al. (2015), a literature review of solar estimation models in cities is carried out. Different alternatives are analysed, from integrating numerical radiation algorithms into GIS tools to Web-based solar maps or from simple 2D visualizations to 3D representation. One of its objectives, as in this paper, is the communication of the benefits achieved by using solar energy.

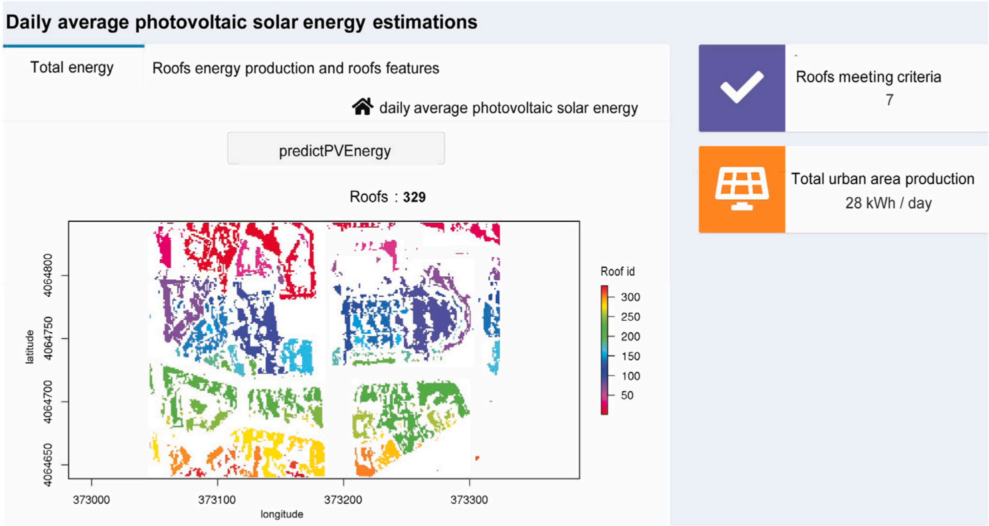
In Liang et al. (2015), a framework is proposed to be used with applications that integrate 3D models with geographic data for solar estimations tasks in cities. It highlights its potential for calculating projected shadow models taking advantage of the GPU computation and displaying the radiation data reproduced in buildings interactively. Another alternative, like (Radosevic et al., 2020), combines KNIME (open-source scientific workflow management system) with Solar Analyst (proprietary tool for assessing solar energy potential) to show how a data science workflow can be used to improve the reproducibility and verifiability of modelling. However, one of its handicaps is that the system cannot be utterly open-source as one component is closed-source.

As noted earlier, the proposed tool will work with the roofs in an area of interest. Semantic segmentation techniques of urban images are usually used to identify those roofs. The main technologies available are: (a) supervised learning using artificial neural networks (Khoshboresh-Masouleh et al., 2020; Pan et al., 2020), and (b) unsupervised learning with clustering techniques using algorithms such as k-means (El Joumani et al., 2017; Gavankar & Ghosh, 2019).

The main problem encountered with supervised learning techniques and artificial neural networks is related to the overfitting problem. Models are trained with aerial images of specific cities, with particular characteristics (resolution of images, shooting angle, types of roofs or kind of vegetation) and those models cannot be extended to other cities.

Table 2
Comparative of solar estimation tools.

	Estimation (long-term) Prediction (short-term)	Updatable data source	Domain	Access
Deep photovoltaic nowcasting	Short-term (one-minute-ahead)	Not available	Research installation in Japan	Not available
DeepRoof	Long-term (daily peak sun hours for the year)	Not available	Few cities in USA	Web (not referenced in paper)
Google Sun Roof	Long-term (daily average for the year)	Not available	Many cities in USA/any location	Web
PVWatts Calculator	Long-term (daily average for the month or the year)	Not needed	World	Web
Huella Solar	Long-term (daily average for the month or the year)	Yes, but not simple, neither intuitive	Few cities in Spain	Web (requires registration)
URSUS-PV	Short-term (hourly for one-day-ahead) long-term (daily average for the year)	Yes, only needs LiDAR images and meteorological data	Any city loaded in the system (for now Spain)	Web


Daily average photovoltaic solar energy estimations

Roofs : 329

Roofs meeting criteria
7

Total urban area production
28 kWh / day

Fig. 6. Energy estimation panel showing the number of roofs that meet user's requirements and daily average photovoltaic solar energy estimation.

This is mainly as the context is different and the quality of images to retrain the models are available only for very few cities in the world. The problem encountered with unsupervised learning techniques is the lack of precision in the segmentation of buildings.

These problems do not allow the automatic segmentation of buildings in urban areas and other information sources are therefore needed. LiDAR technology offers 3D images (taken by drones or aeroplanes) with information on the height of urban objects and can be used to segment objects in urban areas (Awrangjeb et al., 2013). Additionally to the segmentation, other features such as orientation or inclination can be easily calculated for every roof in the area.

The main tools related to prediction of potential solar energy to be generated on roofs are as follows:

- *Deep photovoltaic nowcasting* (Zhang et al., 2018). Tool that makes predictions of energy produced in a photovoltaic system for a very immediate term (one-minute-ahead). The technology used is based on artificial neural networks trained with aerial images of the sky and associated energy values. The system works with the photovoltaic panels and cameras previously installed that take photos of the sky. Learning is based on the energy obtained and the photos of the state of the sky, so it does not help to establish optimal location of facilities, as is the aim of our system.
- *DeepRoof* (Lee et al., 2019). This tool has been developed at the University of Massachusetts and uses data from six different USA cities, focusing on one city in Framingham (Massachusetts). It uses satellite images and real estate data to estimate the size and geometry (orientation and inclination) of the roofs. In addition, these images need to be labelled by experts to create the training set. The system uses a deep learning approach to estimate the solar potential of the roof. It takes 5 s to process each image, so the authors propose to use a cluster of servers that can speed up the process. Buildings are selected individually through a web interface or can be enumerated for batch processing, but cannot be automatically identified in a region of interest. The main problem is that public access is currently not available.
- *Google Sun Roof* (Google, 2021). Tool developed by Google that allows the average daily solar energy production that would be achieved in the roof of a specific building manually selected in the map of a city to be estimated for the long-term. It currently is available for cities in the United States of America. The system uses Google satellite images to calculate the characteristics of the roofs and meteorological information to make energy predictions. Our proposal would differ in that we provide the concept of working with areas, with the system being the one that determines the available roofs, and estimates the energy that can be produced in the long-term (daily for every day in the year) and short-term (hourly for next day).
- *PVWatts Calculator* (Dobos et al., 2019; NREL, 2022). PVWatts calculator estimates the energy production of grid-connected PV energy systems. It has been developed by NREL laboratory of the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy. It uses hourly typical meteorological year (TMY) data from the NREL National Solar Radiation Database. In

this case, the user defines the size of the installation and manually enter the orientation and inclination of the roof; it is not possible to extract information about the available surface in the buildings. As for the Google Sun Roof, the main difference is that it does not work with working areas.
- *Huella Solar* (HuellaSolar, 2021). It allows areas in a city to be selected and performs the monthly and yearly energy estimation. Calculation is made for the whole area depending on radiation. Building segmentation is therefore not automatic and needs to be performed manually if only roofs are desired. A positive point is that it considers shadowing and provided a tool to estimate energy produced by facades. Cities offered are limited in the free version and new ones can be included in the registered version but it requires advanced knowledge about maps and images to be provided.

The contributions that our tool offers with respect to others, such as those described above, are summarized in Table 2 and described in the following paragraphs:

- URSUS-PV uses LiDAR images that allow the urban elements of an area to be determined precisely. Those images are easy to be downloaded for many countries and the availability is increasing. Meteorological data needed in the system are also common (such as temperature or humidity) and they are commonly registered by national agencies. Therefore, URSUS-PV is very flexible in the incorporation of new urban areas. The process to include new cities is easy (and it is documented in the software distribution).
- The tool is open-source and free. It can be used by anyone: public administrations, cooperatives, photovoltaic systems distribution companies or individuals. The system can be easily personalized for any set of cities. Users can work in local mode, but they also can make results available via a web application (as showed in Section 3.1).
- Roofs are automatically segmented in the area of interest and they are aggregated in the global estimation attending to the user's orientation and inclination requirements.
- Short-term photovoltaic energy predictions (one-day-ahead) of a urban area of interest is calculated in addition to a long-term estimation.

The differentiating element that characterizes the developed system is that it allows this information to be obtained for larger areas, and more easily as it has an automated process.

4. Conclusions

An open-source tool, URSUS-PV, has been built to estimate potential electricity that can be generated in photovoltaic (PV) facilities in the short-term (one-day-ahead) and in the long-term (daily average) in an urban area of interest (neighbourhoods, streets, complex buildings). It could be potentially useful for multiple types of users, including municipalities, public administrations, companies in the photovoltaic sector, cooperatives or neighbourhood communities.

One of the tools' most significant benefits is the automation of the complex process. It initially had to be performed manually to obtain global results in urban areas of interest to produce short-term or long-term photovoltaic energy estimations. After conducting a data mining process, such manual processing has now been computerized. CRISP-DM methodology has supported the process. We have executed all the steps from the business understanding to the final deployment, which includes models discovered in intermediate phases. Therefore, following such methodology, once the data sources (meteorological and geographical data) have been identified, its acquisition and integration are automated. Transformations done in the preprocessing stage are also programmed, meaning that segmentation of the roofs available in urban areas of interest and extraction of their features can be easily computed. The calculating photovoltaic energy potential phase has also been fully automated using meteorological data of the city and the configuration of PV facility that could be integrated into each roof.

URSUS-PV can be easily extended to include as many cities as needed. Only LiDAR images and meteorological data from the city are needed and such data are commonly available from National Agencies. The use of this software could improve urban energy sustainability and help in the fight against climate change. It provides long-term production information (daily average), which is essential when determining whether or not it is worth integrating PV facilities. Therefore, prior to embarking on installing any PV unit, it could estimate whether daily energy requirements demanded can be satisfied in an area. Short-term information will help managers of large facilities to improve their integration into the power grid. It can also help owners of self-consumption facilities to determine when they will have energy generated by their facilities and shift their consumption to these hours.

This tool offers excellent opportunities because it can be easily updated with new features made by the same developers or new contributors. After all, the source code is released. Those new features could include focusing on new types of elements in a city or performing a more precise estimation of the energy potential. Many other surfaces apart from roofs could be considered in cities, mainly by the local administration, to install PV facilities. Plots without buildings or main streets with proper PV installations can exploit solar energy while simultaneously generating shades for pedestrians.

Shadowing between buildings is another point not considered in this version. Even though it is not essential because shadowing occurs in extreme hours of the day and PV facilities get maximum performance in central hours of the day, there is some margin for improvement.

Code availability

The URSUS-PV software is available in two different ways: (1) A free web application can be accessed at http://ursus-shiny.uma.es where the system is loaded with the maps of Málaga, the city used in the validation example (all steps described in Section 3.1 can be executed in this web version), and (2) the source code of the tool released under the GNU General Public License v3.0 is available at https://github.com/ursusdm/ursusdm_pv.

CRediT authorship contribution statement

Francisco Rodríguez-Gómez: Methodology, Software, Validation, Formal analysis, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **José del Campo-Ávila:** Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Marta Ferrer-Cuesta:** Methodology, Software. **Llanos Mora-López:** Conceptualization, Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project administration.

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, Spain.

References

A. Coronas, M. V. (1983). Radiación solar total y directa sobre superficies de cualquier inclinación y orientación en Barcelona. *Quaderns D'Enginyeria*, *4*(1), 127–147.

Akbari, H., & Rose, L. S. (2008). Urban surfaces and heat island mitigation potentials. *Journal of the Human-Environment System*, *11*(2), 85–101. http://dx.doi.org/10.1618/jhes.11.85.

Awrangjeb, M., Zhang, C., & Fraser, C. S. (2013). Automatic extraction of building roofs using LIDAR data and multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *83*, 1–18. http://dx.doi.org/10.1016/j.isprsjprs.2013.05.006.

del Campo-Ávila, J., Takilalte, A., Bifet, A., & Mora-López, L. (2021). Binding data mining and expert knowledge for one-day-ahead prediction of hourly global solar radiation. *Expert Systems with Applications*, *167*, Article 114147. http://dx.doi.org/10.1016/j.eswa.2020.114147.

Cannizzaro, D., Aliberti, A., Bottaccioli, L., Macii, E., Acquaviva, A., & Patti, E. (2021). Solar radiation forecasting based on convolutional neural network and ensemble learning. *Expert Systems with Applications*, *181*, Article 115167. http://dx.doi.org/10.1016/j.eswa.2021.115167.

Castangia, M., Aliberti, A., Bottaccioli, L., Macii, E., & Patti, E. (2021). A compound of feature selection techniques to improve solar radiation forecasting. *Expert Systems with Applications*, *178*, Article 114979. http://dx.doi.org/10.1016/j.eswa.2021.114979.

Chang, Y.-P. (2010). An ant direction hybrid differential evolution algorithm in determining the tilt angle for photovoltaic modules. *Expert Systems with Applications*, *37*(7), 5415–5422. http://dx.doi.org/10.1016/j.eswa.2010.01.015.

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).

Dobos, A. P., Freeman, J. M., & Blair, N. J. (2019). Improvements to pvwatts for fixed and one-axis tracking systems. In *2019 IEEE 46th photovoltaic specialists conference (PVSC)* (pp. 1249–1254). IEEE, http://dx.doi.org/10.1109/PVSC40753.2019.8981312.

El Joumani, S., Mechkouri, S. E., Zennouhi, R., El Kadmiri, O., & Masmoudi, L. (2017). Segmentation method based on multiobjective optimization for very high spatial resolution satellite images. *EURASIP Journal on Image and Video Processing*, *2017*(1), 26. http://dx.doi.org/10.1186/s13640-016-0161-2.

Freitas, S., Catita, C., Redweik, P., & Brito, M. (2015). Modelling solar potential in the urban environment: State-of-the-art review. *Renewable and Sustainable Energy Reviews*, *41*, 915–931. http://dx.doi.org/10.1016/j.rser.2014.08.060.

Gavankar, N. L., & Ghosh, S. K. (2019). Object based building footprint detection from high resolution multispectral satellite image using k -means clustering algorithm and shape parameters. *Geocarto International*, *34*(6), 626–643. http://dx.doi.org/10.1080/10106049.2018.1425736.

Google (2021). Google project sunroof. URL: https://www.google.com/get/sunroof.

Hijmans, R. J. (2018). Rspatial: rspatial.org data. URL: https://rspatial.org/.

Hijmans, R. J. (2020). Raster: Geographic data analysis and modeling. URL: https://cran.r-project.org/package=raster.

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets weka. *Computational Statistics*, *24*(2), 225–232. http://dx.doi.org/10.1007/s00180-008-0119-7.

HuellaSolar (2021). Huella solar. URL: http://www.huellasolar.com/.

International Energy Agency (2016). *Energy technology perspectives*, *Energy technology perspectives 2016*. Paris: OECD, http://dx.doi.org/10.1787/energy_tech-2016-en.

International Energy Agency (2020). *Solar pv*: *Technical report*, Paris: International Energy Agency (IEA), URL: https://www.iea.org/reports/solar-pv.

Iqbal, M. (1983). *An introduction to solar radiation*. New York. London: Academic Press, Inc.

Khoshboresh-Masouleh, M., Alidoost, F., & Arefi, H. (2020). Multiscale building segmentation based on deep learning for remote sensing RGB images from different sensors. *Journal of Applied Remote Sensing*, [ISSN: 1931-3195] *14*(03), 1. http://dx.doi.org/10.1117/1.JRS.14.034503.

Kuhn, M. (2020). Caret: Classification and regression training. URL: https://cran.r-project.org/package=caret.

Lee, S., Iyengar, S., Feng, M., Shenoy, P., & Maji, S. (2019). DeepRoof: A Data-driven approach for solar potential estimation using rooftop imagery. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2105–2113). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3292500.3330741.

Li, W., He, C., Fang, J., Zheng, J., Fu, H., & Yu, L. (2019). Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. *Remote Sensing*, *11*(4), 403. http://dx.doi.org/10.3390/rs11040403.

Liang, J., Gong, J., Zhou, J., Ibrahim, A. N., & Li, M. (2015). An open-source 3D solar radiation model integrated with a 3D geographic information system. *Environmental Modelling & Software*, *64*, 94–101. http://dx.doi.org/10.1016/j.envsoft.2014.11.019.

Magrini, A., Lentini, G., Cuman, S., Bodrato, A., & Marenco, L. (2020). From nearly zero energy buildings (NZEB) to positive energy buildings (PEB): The next challenge - the most recent European trends with some notes on the energy analysis of a forerunner PEB example. *Developments in the Built Environment*, *3*(June), Article 100019. http://dx.doi.org/10.1016/j.dibe.2020.100019.

NREL (2022). *PVWatts calculator*. NREL National Laboratory of the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, URL: https://pvwatts.nrel.gov/index.php.

Osterwald, C. (1986). Translation of device performance measurements to reference conditions. *Solar Cells*, *18*(3–4), 269–279. http://dx.doi.org/10.1016/0379-6787(86)90126-2.

Pan, Z., Xu, J., Guo, Y., Hu, Y., & Wang, G. (2020). Deep learning segmentation and classification for urban village using a worldview satellite image based on U-net. *Remote Sensing*, *12*(10), 1574. http://dx.doi.org/10.3390/rs12101574.

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, URL: https://www.r-project.org/.

Radosevic, N., Duckham, M., Liu, G.-J., & Sun, Q. (2020). Solar radiation modeling with KNIME and solar analyst: Increasing environmental model reproducibility using scientific workflows. *Environmental Modelling & Software*, *132*, Article 104780. http://dx.doi.org/10.1016/j.envsoft.2020.104780.

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 Sensing of Environment*, *251*, Article 112061. http://dx.doi.org/10.1016/j.rse.2020.112061.

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

United Nations (2015). The 2030 agenda for sustainable development (a/RES/70/1).

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

Wijffels, J. (2020). CronR: Schedule r scripts and processes with the 'cron' job scheduler. URL: https://cran.r-project.org/package=cronR.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

Zhang, J., Verschae, R., Nobuhara, S., & Lalonde, J.-F. (2018). Deep photovoltaic nowcasting. *Solar Energy*, *176*, 267–276. http://dx.doi.org/10.1016/j.solener.2018.10.024.