PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
MESTRADO EM MEIO AMBIENTE E
DESENVOLVIMENTO REGIONAL

MICHELLE TAÍS GARCIA FURUYA

APLICAÇÃO DE APRENDIZAGEM DE MÁQUINA PARA O MAPEAMENTO DE ILHA DE CALOR URBANO SUPERFICIAL E PREDIÇÃO DA TEMPERATURA DE SUPERFÍCIE TERRESTRE A PARTIR DE VARIÁVEIS AMBIENTAIS E SOCIOECONÔMICAS

Presidente Prudente - SP
2022

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
MESTRADO EM MEIO AMBIENTE E
DESENVOLVIMENTO REGIONAL

MICHELLE TAÍS GARCIA FURUYA

APLICAÇÃO DE APRENDIZAGEM DE MÁQUINA PARA O MAPEAMENTO DE ILHA DE CALOR URBANO SUPERFICIAL E PREDIÇÃO DA TEMPERATURA DE SUPERFÍCIE TERRESTRE A PARTIR DE VARIÁVEIS AMBIENTAIS E SOCIOECONÔMICAS

Dissertação de mestrado apresentado à Pró-reitora de Pesquisa e Pós-Graduação como parte dos requisitos para obtenção do título de Mestre em Meio Ambiente e Desenvolvimento Regional (MMADRE).
Área de Concentração: Ciências Ambientais.

Orientadora:
Profª. Drª. Ana Paula Marques Ramos
Co-orientador:
Dr. Danillo R. Pereira
Co-orientador:
Dr. Lucas Prado Osco
Colaborador Externo:
Prof. Dr. José Marcato Júnior (UFMS)

Presidente Prudente - SP
2022

526.982
F992a

Furuya, Michelle Taís Garcia.
Aplicação de aprendizagem de máquina para o mapeamento de ilha de calor urbano superficial e predição da temperatura de superfície terrestre a partir de variáveis ambientais e socioeconômicas / Michelle Taís Garcia Furuya. – Presidente Prudente, 2022.
38 f.:il.

Dissertação (Mestrado em Meio Ambiente e Desenvolvimento Regional) - Universidade do Oeste Paulista – Unoeste, Presidente Prudente, SP, 2022.
Bibliografia.
Orientadora: Ana Paula Marques Ramos

1. Decision tree. 2. Ilha de calor urbano superficial. 3. Sensoriamento remoto. 4. Temperatura de superfície terrestre. I. Título.

Catalogação na Fonte: Ivy Fini Rodrigues – CRB 8/7470

MICHELLE TAÍS GARCIA FURUYA

APLICAÇÃO DE APRENDIZAGEM DE MÁQUINA PARA O MAPEAMENTO DE ILHA DE CALOR URBANO SUPERFICIAL E PREDIÇÃO DA TEMPERATURA DE SUPERFÍCIE TERRESTRE A PARTIR DE VARIÁVEIS AMBIENTAIS E SOCIOECONÔMICAS

Dissertação de mestrado apresentado à Pró-reitora de Pesquisa e Pós-Graduação como parte dos requisitos para obtenção do título de Mestre em Meio Ambiente e Desenvolvimento Regional (MMADRE). Área de Concentração: Ciências Ambientais.

BANCA EXAMINADORA

_____

Prof. Dr. Ana Paula Marques Ramos
Universidade do Oeste Paulista (UNOESTE)
Presidente Prudente – SP

_____

Prof. Dr. Paulo Antônio da Silva
Universidade do Oeste Paulista (UNOESTE)
Presidente Prudente – SP

_____

Prof. Dr. Rejane Ennes Cicerelli
Universidade de Brasília (UnB)
Brasília – DF

DEDICATÓRIA

Aos meus pais, irmãs e toda a família por serem a melhor parte da minha vida.

Ao meu avô Nelson por todo o amor, sabedoria e por sempre estar presente.

AGRADECIMENTOS

Agradeço primeiramente à Deus, pois sem ele nada seria possível.

Agradeço à minha família por todo apoio e por acreditarem em mim. Aos meus pais, Silvana e Osmar, e às minhas irmãs, Isabelle e Danielle. Obrigada por todo o amor, carinho, amizade, ensinamentos e compreensão. É graças a vocês que me tornei o que sou hoje.

Agradeço à minha avó Nízia pelo companheirismo, pelos ensinamentos, pelas alegrias e risadas.

Agradeço ao meu avô Nelson por toda a sabedoria, por todos os ensinamentos, todo o cuidado e amor.

Agradeço a todos os meus familiares por sempre estarem presentes na minha vida.

Agradeço às minhas amigas de infância, Clara e Rafaela, por todos esses anos de amizade.

Agradeço a todos os professores (escola, graduação e pós-graduação) por todos os ensinamentos e por fazerem a diferença.

Agradeço a todos do PPGMADRE pelos ensinamentos e pelo apoio.

Agradeço à minha orientadora Dra. Ana Paula Marques Ramos e ao meu co-orientador Dr. Lucas Prado Osco por todo incentivo, por toda ajuda, pelos ensinamentos, pela paciência, mas acima de tudo por acreditarem no meu potencial e me apoiarem sempre.

Agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa concedida e incentivo permanente à ciência.

“ You would rather find purpose than a job or career.
Purpose crosses disciplines.
Purpose is an essential element of you.
It is the reason you are on the planet at this particular time in history.
Your very existence is wrapped up in the things you are here to fulfill.
Whatever you choose for a career path, remember, the struggles along
the way are only meant to shape you for your purpose. ”

(Chadwick Boseman)

LISTA DE SIGLAS

BOA - Bottom Of Atmosphere
CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
DN – Número Digital
DT - Decision Tree
ESA - Agência Espacial Européia
IBGE – Instituto Brasileiro de Geografia e Estatística
ISA - Área de Superfície Impermeável
KNN – K-Nearest Neighbor
LR – Linear Regression
LST – Temperatura de Superfície Terrestre
MAE – Erro Médio Absoluto
MLP – Multilayer Perceptron
NDBI – Índice de Área Construída por Diferença Normalizada
NDVI – Índice de Vegetação por Diferença Normalizada
NIR – Infravermelho Próximo
OLI - Operational Land Imager
PPGMADRE – Programa de Pós-Graduação em Meio Ambiente e Desenvolvimento Regional
PROINTER - Programa de Pesquisa Interdisciplinar
r – Coeficiente de Correlação
RF - Random Forest
RMSE – Raíz Quadrada do Erro Médio
SUHI – Ilha de Calor Urbano Superficial
SVM – Support Vector Machine
SVR - Support Vector Regression
SWIR – Infravermelho de Ondas Curtas
TIR – Infravermelho Termal
TOA - Top-of-Atmosphere
UHI – Ilha de Calor Urbano
USGS - United States Geological Survey
UTM - Universal Transverse Mercator

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 CONSIDERAÇÕES INICIAIS

Este documento está organizado em três seções. A primeira seção consiste no contexto geral da presente pesquisa que é promovida pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e desenvolvida no Programa de Pós-Graduação em Meio Ambiente e Desenvolvimento Regional (PPGMADRE) da Universidade do Oeste Paulista (UNOESTE). A segunda seção consiste em um manuscrito que discute a capacidade de mapear ilha de calor urbano superficial (SUHI) a partir de dados de temperatura de superfície terrestre (LST) estimados por algoritmos de aprendizagem de máquina. A terceira seção apresenta considerações sobre o desenvolvimento deste trabalho.

O PPGMADRE é composto por duas linhas de pesquisa: Avaliação e análise de impactos ambientais; e Planejamento ambiental e desenvolvimento regional, que se enquadra no presente trabalho. O programa é interdisciplinar e conecta as questões ambientais ao desenvolvimento regional. As linhas de pesquisa atuam em um Programa de Pesquisa Interdisciplinar denominado PROINTER.

Este relatório de defesa de mestrado discute a capacidade dos algoritmos de aprendizagem de máquina em otimizar a caracterização da SUHI com base em variáveis ambientais e socioeconômicas extraídas de imagens de satélite Landsat 8, Sentinel-2 e Planet. O estudo contempla o distrito de Presidente Prudente – SP e trata-se de um estudo de um fenômeno climático. Isso atesta a proposta do PPGMADRE de discutir o processo de desenvolvimento regional frente às questões ambientais.

2 MANUSCRITO

APLICAÇÃO DE APRENDIZAGEM DE MÁQUINA PARA O MAPEAMENTO DE ILHA DE CALOR URBANO SUPERFICIAL E PREDIÇÃO DA TEMPERATURA DE SUPERFÍCIE TERRESTRE A PARTIR DE VARIÁVEIS AMBIENTAIS E SOCIOECONÔMICAS

Resumo:
A característica da paisagem é responsável por alterar a dinâmica urbana, causando a formação de Ilhas de Calor Urbano (UHI). A temperatura de superfície (LST) extraída de imagens termais é uma fonte de informação primária para se estudar UHI, caracterizando as Ilhas de Calor Urbano Superficial (SUHI). Além da LST, um conjunto de variáveis ambientais e socioeconômicas tem sido adotado para explicar o fenômeno de SUHI. Embora algoritmos de aprendizagem de máquina demonstrem potencial em várias áreas, ainda se desconhece a aplicação destes na determinação da relação de variáveis socioeconômicas e ambientais e SUHI. Este trabalho propõe a caracterização de SUHI usando variáveis socioeconômicas e ambientais a partir de aprendizagem de máquina. A LST foi extraída de 15 imagens coletadas pelo sensor TIRS, sistema Landsat 8, para o período de 2019 a 2021. Os dados das variáveis socioeconômicas foram obtidos no censo demográfico oficial. As variáveis ambientais, descritas por índices espectrais de vegetação (NDVI), de área construída (NDBI) e de superfícies impermeáveis (ISA), foram extraídas de imagens Sentinel-2 e Planet. Algoritmos de aprendizagem de máquina foram testados para avaliar a capacidade de estimar a LST com base nas variáveis citadas. Os algoritmos utilizados no estudo foram decision tree (DT), k-nearest neighbour (KNN), linear regression (LR), multilayer perceptron (MLP), support vector regression (SVR) e random forest (RF). Os resultados mostraram que o algoritmo DT obteve o melhor desempenho (r= 0.96, MAE= 1.49 °C and RMSE= 1.88 °C), seguido do RF. Além disso a inclusão de todas as estações do ano e de variáveis socioeconômicas mostrou ter relevância nos resultados. A principal contribuição deste trabalho é verificar se os algoritmos podem otimizar o processo de caracterização da SUHI, analisando a influência exercida pelas variáveis estudadas. No âmbito social, as informações produzidas podem auxiliar o planejamento urbano visando a construção de cidades inteligentes.
Palavras-chave: decision tree; ilha de calor urbano superficial; machine learning; sensoriamento remoto; temperatura de superfície terrestre.

Abstract:
The landscape feature is responsible for changing urban dynamics, causing the formation of Urban Heat Islands (UHI). The Land surface Temperature (LST) extracted from thermal images is a primary source of information for exploiting UHI, characterizing the Surface Urban Heat Islands (SUHI). In addition to LST, a set of environmental and socioeconomic variables has been adopted to explain the SUHI phenomenon. Although machine learning algorithms have demonstrated great potential in several areas of applications, their use to investigate the relationship between socioeconomic and environmental variables and SUHI is still unknown. This work proposes the characterization of SUHI using socioeconomic and environmental variables from machine learning approach. The LST was extracted from 15 images collected by the TIRS sensor, Landsat 8 system, from 2019 to 2021. Data on socioeconomic variables were obtained from the official demographic census. The environmental variables, described by spectral indices of vegetation (NDVI), built-up area (NDBI) and impervious surfaces (ISA), were extracted from Sentinel-2 and Planet images. Machine learning algorithms were tested to evaluate the ability to estimate LST based on the aforementioned variables. The algorithms used in the study were decision tree (DT), k-nearest neighbor (KNN), linear regression (LR), multilayer perceptron (MLP), support vector regression (SVR) and random forest (RF). The results showed that DT algorithm obtained the best performance (r= 0.96, MAE= 1.49 °C and RMSE= 1.88 °C), followed by the RF. Furthermore, the inclusion of all seasons and socioeconomic variables proved to be relevant in the results. The main contribution of this work is to verify if the algorithms can optimize the SUHI characterization process, analyzing the influence exerted by the studied variables. In the social sphere, the information produced can help urban planning in the construction of smart cities.

Keywords: decision tree; surface urban heat island; machine learning; remote sensing; land surface temperature.

Introduction

The urbanization process is responsible for transforming the natural landscape, which can cause an increase in impervious surfaces. Changes in land use and land cover result in an increase in land surface temperature (LST), as the higher the impervious surface index, the higher the temperature. Consequently, there is the formation of urban heat island (UHI), characterized by higher temperatures in urban areas when compared to rural surroundings. This phenomenon can generate negative impacts on the ecosystem, the local climate, the hydrological system, biodiversity and human health (Gomez-Martinez et al. 2021; Li et al., 2018; Mathew et al., 2015).

As UHI is related to temperature changes, it is possible to state that LST is a key parameter to represent UHI (Mathew et al., 2015; Guha et al., 2018). In related works, the LST variation was characterized based on a single date (Weng et al., 2004; Jenerette et al. 2006; Mathew et al., 2015; Yoo, 2018), two (Zhang et al., 2009; Li et al. 2011), three (Guha et al. 2018) or four dates (Yuan and Bauer, 2007; Buyantuyev and Wu, 2009; Weng et al., 2011) being often dates from the same year. However, using a single moment may not represent its real variation in the region under analysis, since temperature is a dynamic variable with continuous distribution in space that can undergo variations due to different factors such as seasonality and humidity. To solve this, it is possible to work with a larger amount of satellite images in order to obtain more temperature information.

Remote sensing data, through satellite images, correspond to highly relevant information for UHI mapping. With these data, it is possible to study the relationship between UHI and surface biophysical parameters in wide geographic areas and on a large cartographic scale, even though LST is an instantaneous measurement of surface temperature. Multiple studies (Li et al., 2011; Voogt & Oke, 2003; Yang et al., 2017) adopt the term Surface Urban Heat Island (SUHI) to name the UHI characterized by LST, whose data are collected remotely. The Landsat 8 system, for example, provides thermal images at a spatial resolution of 100 meters, but resampled to 30 m, which enable the diurnal mapping of the LST and, therefore, the study of SUHI, at a regional scale (Li et al., 2011; Guha et al., 2018).

To understand the occurrence of SUHI, it is essential to understand the configuration of the urban landscape. For this reason, studies (Guha et al., 2018; Mathew et al., 2015; Weng et al., 2004; Yoo, 2018) seek to analyze the relationship of LST with environmental variables. These variables include the percentage of vegetation cover and the percentage of impervious surface, which are generally described by the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Built-Up Index (NDBI), respectively, which can be extracted, for example, from Landsat images.

Although still few, there are studies (Buyantuyev & Wu, 2010; Tang et al., 2017; Yoo, 2018) that include socioeconomic variables in the SUHI study. These variables are described by the data from the demographic census, representing a complement to the analysis with environmental variables. As changes in the urban landscape do not occur in isolation, incorporating socioeconomic variables is fundamental. For the city of Baltimore, in the United States, Tang et al. (2017) performed the SUHI analysis with seven socioeconomic variables: population density, median income, number of households, average age of the population, age of construction, family size and unemployment rate. Population density was the variable that most influenced the SUHI value. Although the impact of this variable was smaller compared to environmental variables, it interferes, for example, in the presence of paving and constructions that, therefore, influence the impervious surface area, which is strongly related to the occurrence of SUHI (Tang et al. 2017).

Buyantuyev and Wu (2010) also incorporated socioeconomic variables in the work carried out in the metropolitan region of Phoenix, United States. They concluded that median family income was the second most influential variable on daytime temperature (daytime temperature decreased by 0.36 °C in June and 0.23 °C in October for every $10,000 increase in family income) denoting the importance of combining socioeconomic and environmental variables in the analysis.

SUHI characterization studies have adopted traditional methods of analysis such as Pearson's correlation (Li et al., 2011; Tang et al., 2017), path analysis, multiple ordinary least squares regression (OLS) and geographically weighted regression (GWR) (Buyantuyev and Wu, 2010). Although these approaches bring promising results, methods that are more robust should be investigated given the complexity of the interaction between SUHI and environmental and socioeconomic variables. In recent years, the application of machine learning techniques appears as a new

approach for studies in different areas. In the context of SUHI, there are the works by Sun et al. (2019), Osborne and Sanches (2019) and Kafy et al. (2020). However, these studies use machine learning to infer the value of LST, without establishing any relationship between environmental and socioeconomic variables and SUHI, which demonstrates a gap to be filled.

In this study, we proposed a machine learning approach to predict LST and characterize the influence of socioeconomic and environmental variables on SUHI formation. As a complement, the study aims to verify whether there is any change in the spatial distribution of SUHI throughout the year, whether there is an association between the spatial distribution of SUHI and socioeconomic and environmental variables, and to measure the contribution of these variables in the formation of SUHI. The study area was divided into 309 census sectors provided by the Brazilian Institute of Geography and Statistics (IBGE). The scientific contribution of this study consists of verifying how machine learning algorithms can optimize the SUHI characterization process based on a set of environmental and socioeconomic variables. Whereas the social contribution consists of producing information that helps in urban planning and contributes to the production of smart cities.

Materials and Method

The method (Fig. 1) was divided into four main phases: (1) data collection and preparation; collected from different satellites to generate environmental and socioeconomic variables; (2) data organization; submitted to statistical calculations and subsequent definition of attributes for each model; (3) machine learning regression; applied to indicate which algorithm has the best performance in predicting the LST according to the data provided; (4) Definition of SUHI areas; preparation of thematic maps and influence of variables.

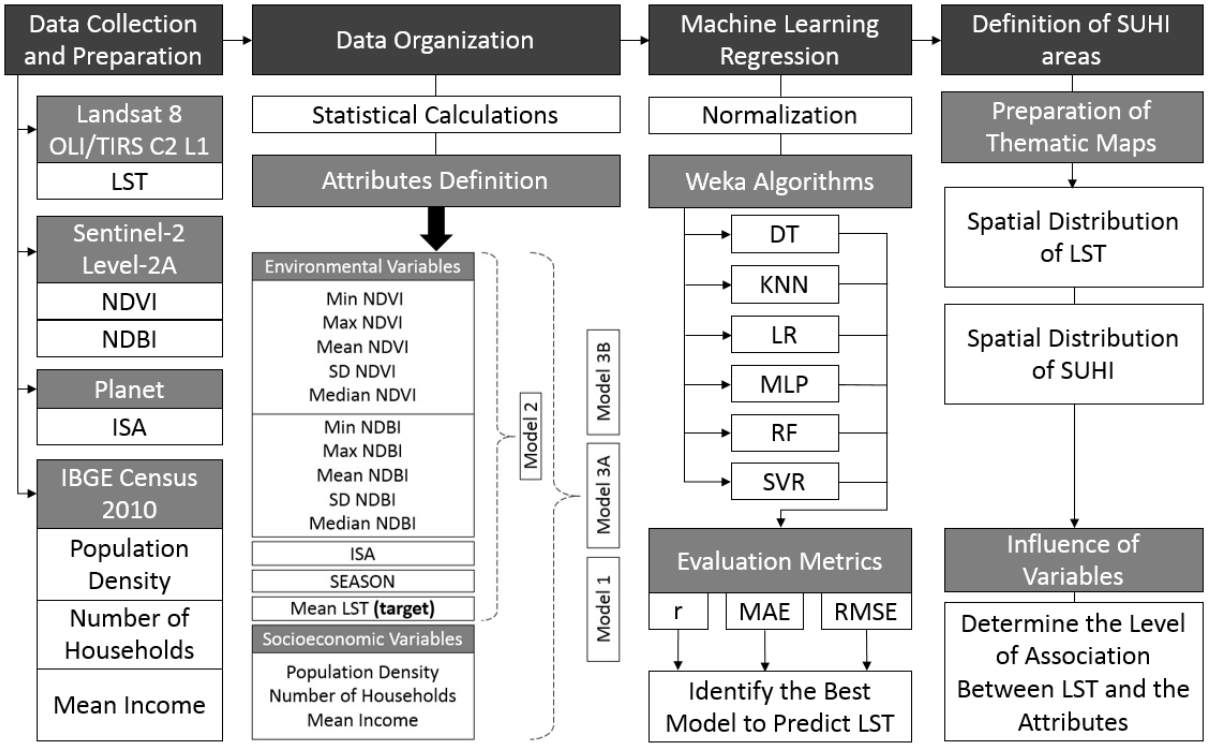

Figure 1. Workflow of the proposed method.

Study Area

The study area corresponds to the district of Presidente Prudente, with an approximate area of 230.05 km², including the urban and rural perimeter. The municipality is located in the western of São Paulo state, Brazil and has an area of 560.637 km². Its estimated population is 231,953 inhabitants in 2021 (IBGE, 2021). The city has a high concentration of commerce and services. Therefore, it presents a considerable expansion of the urban fabric over the last few years. According to the Köppen climate classification, the climate corresponding to the study area is tropical humid (Aw) characterized by dry winters (Beck et al, 2018). Figure 2 shows the location of São Paulo state, the location of Presidente Prudente city and the delimitation of the Presidente Prudente district related to the extension of the municipality. The district image corresponds to bands 4, 3 and 2 of the Sentinel -2 Level 2A satellite at April 22, 2021.

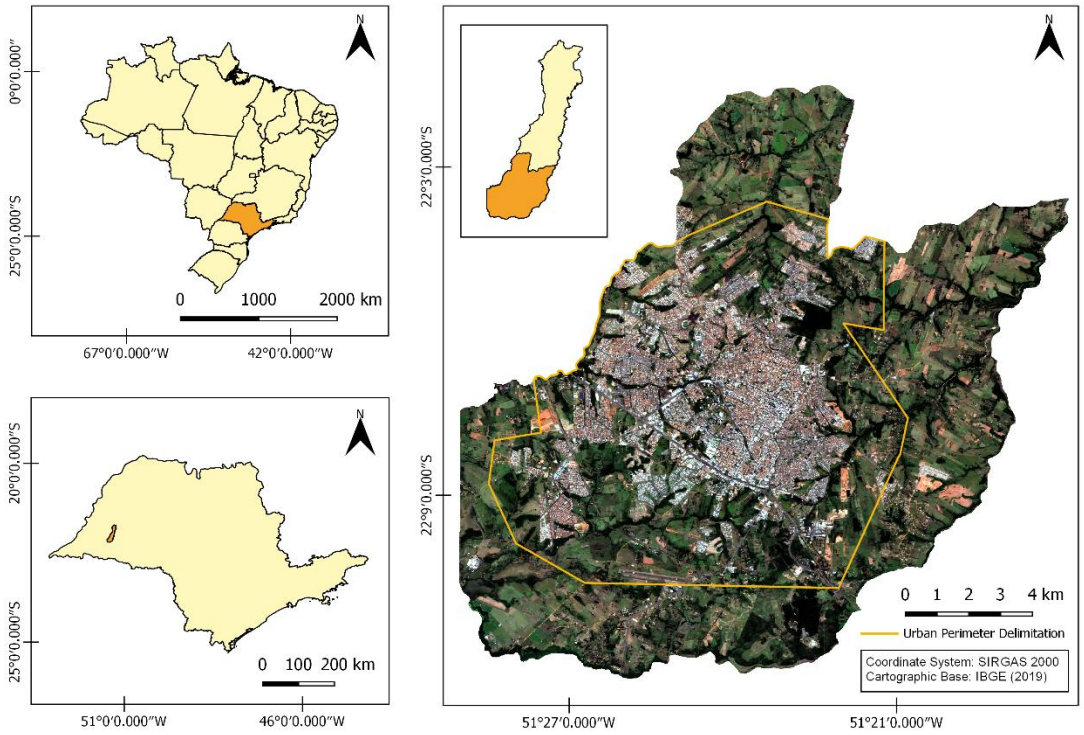

Figure 2. Study area

Data Collection and Preparation

Our study considered all cloudless images available from 2019 to 2021. This strategy ensures a more accurate representation of the LST, as the temperature has continuous spatial variability, in addition to varying from one season to another. Therefore, analyzing the temperature based on isolated dates may not represent its true variation. For LST, 15 dates were used considering all seasons. For each LST date, we combine the nearest NDVI/NDBI date available. The LST and NDVI/NDBI dates are not always the same due to the difference in the temporal resolution of the Landsat 8 - 16 days - and Sentinel-2 - 5 days - satellites. The LST dates, the corresponding NDVI/NDBI dates and their respective seasons are described in table 1.

Table 1. Information regarding the 15 Landsat 8 and Sentinel-2 images used in this study.

Date (LST) Landsat 8	**Date (NDVI/NDBI) Sentinel-2**	**Season in the South Hemisphere**
22 August 2019	16 August 2019	Winter
07 September 2019	15 September 2019	Winter
25 October 2019	25 October 2019	Spring
29 January 2020	28 January 2020	Summer
18 April 2020	17 April 2020	Autumn
04 May 2020	17 May 2020	Autumn
21 June 2020	21 June 2020	Winter
07 July 2020	11 July 2020	Winter
23 July 2020	21 July 2020	Winter
08 August 2020	10 August 2020	Winter
24 August 2020	25 August 2020	Winter
09 September 2020	09 September 2020	Winter
25 September 2020	24 September 2020	Spring
05 April 2021	02 April 2021	Autumn
26 July 2021	26 July 2021	Winter

Land Surface Temperature Retrieval from Landsat 8

The temperature data was retrieved from the Landsat 8 OLI/TIRS satellite of collection 2, level 1, in order to get the primary data in digital number. To retrieve the LST we use the Land Surface Temperature (LST) Plugin, developed by Ndossi and Avdan (2016), with QGIS 2.18. This plugin allows the LST to be extracted from a thermal image. For this case, we used band 10 (Thermal Infrared - TIRS sensor) of the Landsat 8 satellite, whose spatial resolution is 30 m. Although Landsat 8 has two Thermal Infrared (TIR) bands, it is recommended to only use band 10 to extract LST due to the calibration uncertainties presented by band 11(USGS, 2019). First, the plugin converts the digital number (DN) to radiance and then converts it to brightness temperature. Subsequently, the plugin uses the Red and Near Infrared bands (4 and 5 respectively) to obtain the NDVI in order to perform the surface emissivity calculation (Ndossi e Avdan, 2016). The OLI (Operational Land Imager) sensor provides bands 4 and 5, both with a spatial resolution of 30m. Finally the plugin estimates the surface temperature allowing the data to be displayed in degree Celsius.

The conversion from digital number (DN) to radiance is done based on Equation 1:

$$L_\lambda = M_L Q_{cal} + A_L - O_i \quad (1)$$

where $L\lambda$ is the radiance corresponding to the energy of the Top-of-Atmosphere (TOA) in Watts/(m2 srad μm), M_L and A_L are band-specific scale factors (multiplicative and additive, respectively) available in the metadata, Q_{cal} are the pixel values of the image in DN and O_i is the calibration in band 10 recommended by the USGS (2019) whose value is -0.29.

After radiance, it is necessary to convert to brightness temperature (Equation 2) so that the emitted radiation is equal to that of the real body:

$$T_{sen} = \frac{K_2}{ln\left(\frac{K_1}{L_\lambda} + 1\right)} \quad (2)$$

where T_{sen} is the brightness temperature, $L\lambda$ is the radiance and K_1 and K_2 are thermal band specific conversion constants.

As stated, the emissivity calculation (Equation 3) can be done using the NDVI (Sobrino et al. 2004):

$$\varepsilon_\lambda = \varepsilon_{v\lambda} P_v + \varepsilon_{s\lambda} (1 - P_v) + C_\lambda \quad (3)$$

where ε_v and ε_s is the vegetation emissivity and the soil emissivity respectively, P_v is the proportion of vegetation and C is the value referring to the effect of the surface geometry, assuming a null value for flat surfaces.

Finally, the calculation of the LST (Equation 4) is done, in Kelvin, using the Planck function:

$$T_s = \frac{BT}{\left\{1 + \left[\frac{\lambda \cdot BT}{\rho}\right] \cdot ln\varepsilon\right\}} \quad (4)$$

where T_s is the land surface temperature, BT is the brightness temperature, λ is the wavelength of the emitted radiation, ρ is a constant equal to 1.438 ×10−2 mK and ε is the emissivity. To convert the temperature from Kelvin to degrees Celsius it is necessary to subtract the value 273.15 from the result of the equation (Carrasco et al., 2020).

As the division of the territory was based on the census sectors, the number of pixels for the Landsat images in each sector varies from 3 to 25390 and the area in pixels varies from 2700 m² to 22851000 m².

Acquisition of Environmental and Socioeconomic Variables

The data to generate the NDVI and NDBI indices were extracted from the Sentinel-2 Level-2A satellite. The images were downloaded from the Copernicus Open Access Hub platform. For Sentinel-2 Level-2A images, there is no need to perform atmospheric correction, as this step is already part of image processing. The algorithm developed by DLR/Telespazio uses Top Of Atmosphere (TOA) reflectance present in Level-1C products to calculate Bottom Of Atmosphere (BOA) reflectance (ESA, 2015). Sentinel-2 images with MSI sensor have spatial resolution of 10, 20, and 60 m, 13 spectral bands and 12-bit radiometric resolution.

NDVI is obtained from the Near Infrared (NIR) and Red bands, bands 4 and 8 of the Sentinel-2 satellite respectively. Therefore, to calculate the NDVI we use the following equation (Sobrino et al., 2004):

$$NDVI = \frac{NIR - R}{NIR + R} \quad (5)$$

NDBI is obtained from the Short Wave Infrared (SWIR) and Near Infrared (NIR) bands, bands 11 and 8 of the Sentinel-2 satellite respectively. Therefore, to calculate the NDBI we use the following equation (Zha., 2003):

$$NDBI = \frac{SWIR1 - NIR}{SWIR1 + NIR} \quad (6)$$

To calculate the impervious surface area, a Planet image from April 22, 2021 was used. The spatial resolution of the image is 3m and it was projected to the WGS-84 UTM 22 S zone system. The image only covers the urban perimeter of the study area. We performed the image segmentation process and based on training samples containing impervious and pervious polygons we did the classification process using the Support Vector Machine (SVM) algorithm.

The socioeconomic variables were obtained using data from the last census in Brazil, in 2010, provided by the Brazilian Institute of Geography and Statistics (IBGE). Based on the literature (Buyantuyev and Wu, 2010; Tang et al., 2017), the chosen variables were population density, number of households and median income.

Data Organization

After extracting the necessary environmental, socioeconomic and temperature parameters, statistical information from the data was calculated. Among the statistics obtained, the minimum, maximum, mean, standard deviation and median values of NDVI and NDBI were selected, in addition to the mean of LST. It is noteworthy that the statistical values, as well as the variables, were calculated following the division of the 309 census sectors.

A total of 16 attributes were defined as input file for the algorithms: 5 for NDVI (minimum, maximum, mean, standard deviation and median), the same 5 for NDBI, the Impervious Surface Area - ISA, seasons, 3 socioeconomic variables (population density, number of households and median income) and the mean LST that represents the target attribute of the model.

To better understand the influence of attributes, amount of data and seasons, given as input file for the algorithms, we created four models with specific characteristics: model 1, model 2, model 3A and model 3B. Model 1 has 16 attributes, 12 training dates, 3 testing dates and considers all seasons. Model 2 does not consider socioeconomic variables as attributes, thus presenting 13 attributes in total. As for training and testing dates and seasons, model 2 follows the same pattern as model 1. Models 3A and 3B are complementary as they consider two seasons for each model. Whereas model 3A contains the autumn and winter dates, with 10 dates for training and 2 for testing, model 3B is composed of spring and summer dates, with 2 dates for training and 1 for testing. For models 3A and 3B the 16 attributes are present. The details of the attributes used for each model can be seen in the workflow (Figure 1) and the information regarding each model can be seen in table 2.

Table 2. Description of the study models

Model Description	Model 1	Model 2	Model 3A	Model 3B
Number of Attributes	16	13	16	16
Number of Training Dates	12	12	10	2
Number of Testing Dates	3	3	2	1
Number of Training Instances	3708	3708	3090	618
Number of Testing Instances	927	927	618	309
Seasons	All	All	Autumn/ Winter	Spring/ Summer

Machine Learning Regression

Before submitting the four models to machine learning algorithms, data normalization was necessary so that the numerical scale of the data was common. Thus, the ranges of values are maintained, but the data scale varies from 0 to 1. This step was necessary because the data have different scales from each other, which may interfere in the performance of the algorithms (Singh and Singh, 2019). To evaluate the performance in estimating land surface temperature, our study considered the six most used algorithms in regression: decision tree (Dos Santos, 2020), k-nearest neighbour (Ali et al., 2019), linear regression (Štepanovský et al., 2017), multilayer perceptron (Linares-Rodriguez et al., 2013), support vector regression (Ebrahimy e Azadbakht, 2019) and random forest (Sun et al., 2019). This stage was performed in the software Weka 3.9.5.

Based on the amount of data available in each model, the training and testing datasets were divided following the proportion of 80% and 20%, respectively. All hyperparameters follow Weka 3.9.5 default. We also performed a 10-fold cross-validation with 10 repetitions for each set. The metrics used to evaluate the performance of the algorithms were the correlation coefficient (r), the mean absolute error (MAE) and the root mean square error (RMSE). For regression situations, the MAE and RMSE metrics have been widely used (Dos Santos, 2020; Marques Ramos et al.,2020).

Definition of SUHI Areas

The SUHI mapping (Equations 7 and 8) was based on the LST variation (Ma et al. 2010; Guha et al. 2017; Guha et al. 2018):

$$LST > \mu + 0{,}5 * \text{SD} \quad (7)$$

$$0 < LST \leq \mu + 0{,}5 * \text{SD} \quad (8)$$

where μ and SD are, respectively, the mean and standard deviation of LST in the study area.

The calculation of the SUHI spatial distribution is based on the LST extracted from the satellite image, but for this study the SUHI mapping was made for both the real LST and the predicted LST.

Results

As stated, model 1 considers all environmental and socioeconomic variables and all seasons, model 2 differs from the first in not considering socioeconomic variables and models 3A and 3B consider all variables, but with two seasons in each model (autumn and winter for 3A and spring and summer for 3B). As the study has 4 models applied to 6 algorithms, we obtained results for 24 regression models. Table 3 shows the values of each metric used (r, MAE, RMSE) to evaluate the performance of the algorithms in the four proposed models.

Table 3. Performance evaluation applying the trained models into testing models

Model	Algorithm	r	MAE	RMSE
1	DT	0.96	1.49	1.88
	KNN	0.92	1.69	2.15
	LR	0.96	1.74	2.04
	MLP	0.94	1.78	2.16
	RF	0.98	1.54	1.91
	SVR	0.96	1.72	2.04
2	DT	0.95	1.55	1.92
	KNN	0.91	1.79	2.25
	LR	0.96	1.75	2.07
	MLP	0.92	1.93	2.31
	RF	0.97	1.60	1.96
	SVR	0.96	1.75	2.07
3A	DT	0.40	1.04	1.40
	KNN	0.40	1.38	1.98
	LR	0.55	1.33	1.60
	MLP	0.69	1.05	1.37
	RF	0.67	0.90	1.16
	SVR	0.58	1.35	1.62
3B	DT	0.72	2.55	2.66
	KNN	0.69	2.47	2.59
	LR	0.53	1.86	2.06
	MLP	0.66	3.47	3.58
	RF	0.86	2.82	2.87
	SVR	0.57	1.92	2.11

The correlation coefficient (r) range was between 0.40 and 0.98. The MAE ranged from 0.90 °C to 3.47 °C. The RMSE ranged from 1.16 °C to 3.58 °C. An analysis considering all metrics shows that the DT algorithm of model 1 achieved better results with r= 0.96, MAE= 1.49 °C and RMSE= 1.88 °C. The RF algorithm in model 1 obtained a correlation coefficient 2% higher than the DT, however the MAE (1.54 °C) and RMSE (1.91 °C) values were higher. Since RMSE squares errors, it penalizes large errors (Chai e Draxler, 2014), which can be useful for improving the model performance. Therefore, MAE and RMSE values had greater weight in this study compared to r.

Conversely, MAE and RMSE values were the lowest in the 3A model, which only considers autumn and winter dates. However, the model cannot be considered the one with the best performance due to the low r values, ranging from 0.40 to 0.69.

Low r values for models 3A and 3B indicate the relevance of seasons in predicting LST. In addition, r values up to 2% higher for model 1 compared to model 2 demonstrate the influence of socioeconomic variables on the landscape configuration, which indicates the need to consider them in studies aimed at estimating the LST. For model 3B, higher MAE and RMSE values may be associated with the amount of dates available for training (2) and testing (1), which proved to be insufficient for this case.

Figures 3, 4 and 5 show the boxplots with the evaluation metrics (r, MAE and RMSE respectively). Model 3B has higher r and lower MAE and RMSE than the others, which would characterize it as the best performing model. However, as shown in the test data in table 3, the same model had higher MAE and RMSE values, as well as median values for r. This situation is called overfitting and occurs when the model fits the training data. Thus, even though training presents good results, the model is not able to adjust to new data, in other words, the testing data. One of the causes of overfitting is the limited size of training set (Ying, 2019), which justifies what happened in the 3B model.

Still based on the boxplots, it is possible to notice the similarity of results of models 1 and 2 when compared to the testing data in table 3. The graphics show the best performance of model 1 in relation to 2, with higher correlation coefficient values and lower MAE and RMSE values.

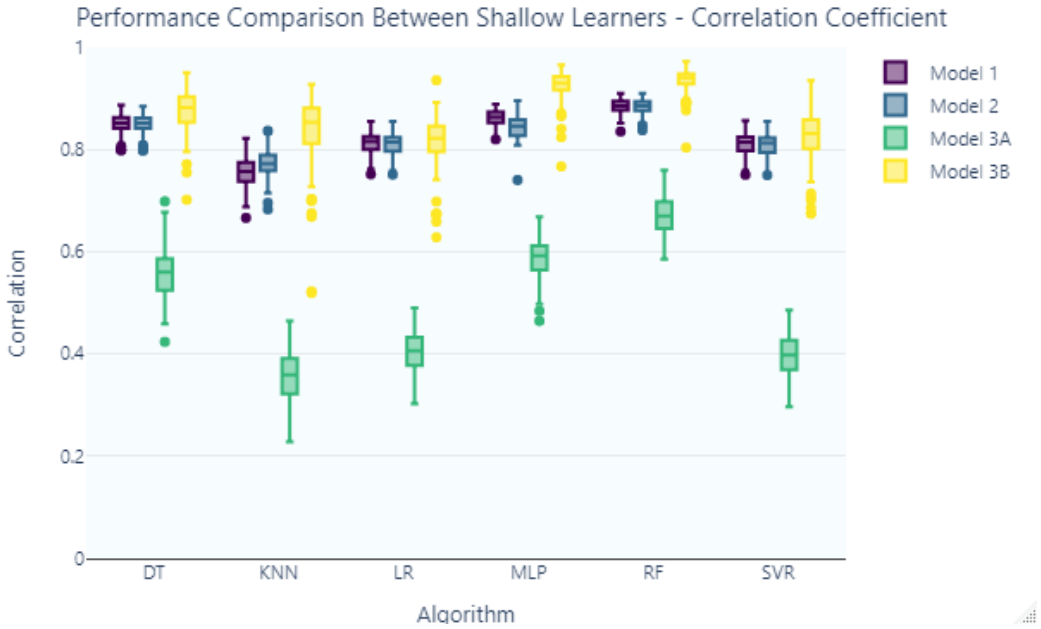

Figure 3. Boxplot for correlation coefficient (r) of algorithms applied to different datasets

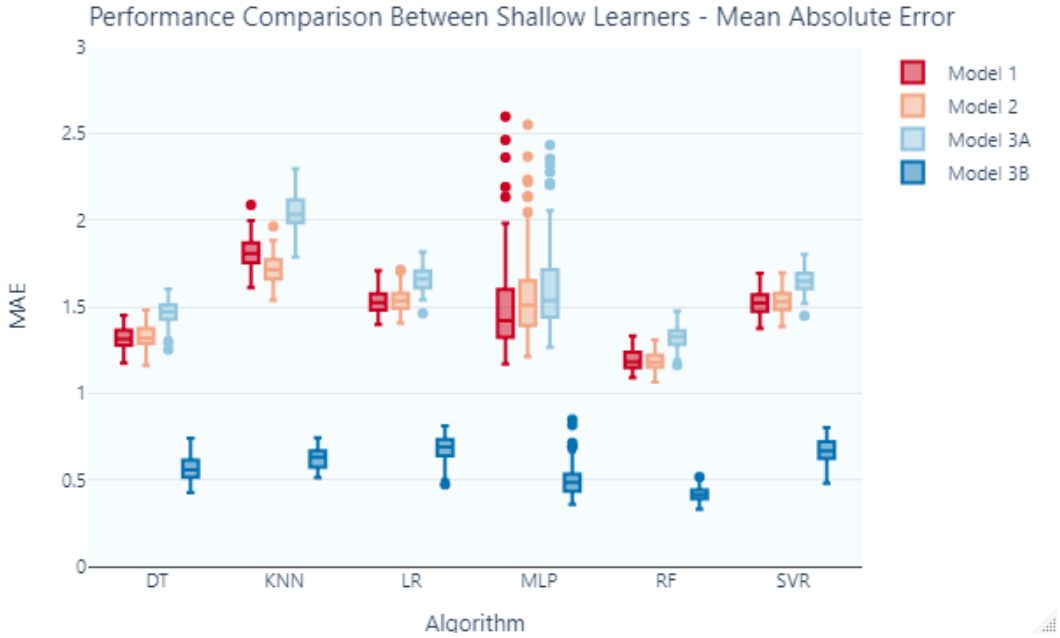

Figure 4. Boxplot for mean absolute error (MAE) of algorithms applied to different datasets

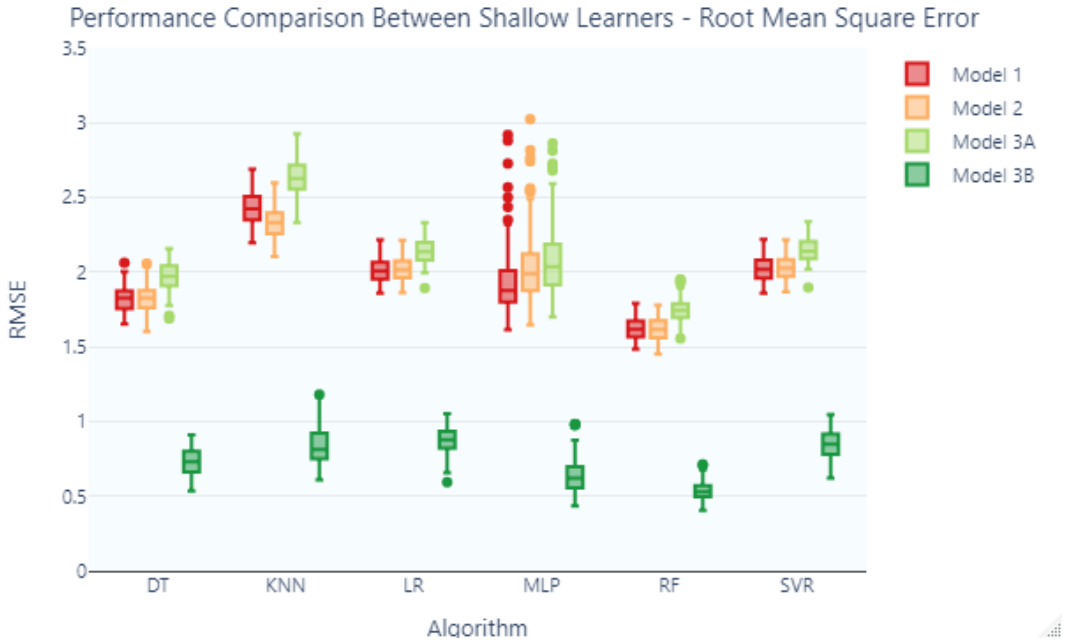

Figure 5. Boxplot for root mean square error (RMSE) of algorithms applied to different datasets

After finding the best algorithm applied to the best model, we created thematic maps showing the performance of DT in predicting the LST. The maps show the predicted LST and the error, calculated by the difference between the predicted and actual temperature. Model 1 has 3 testing dates (Table 4), so we create maps for each test date of the model.

Table 4. Information regarding the testing data from model 1

Testing Dates	Day	Season in the South Hemisphere
1	25 October 2019	Spring
2	04 May 2020	Autumn
3	26 July 2021	Winter

The maps were prepared following the division of census sectors and are shown in figure 6. As a complement, table 5 shows the minimum and maximum values for predicted LST and LST error for each date represented on the map.

An analysis of the error values by census sector shows that on the spring date only two sectors had a predicted LST higher than the real LST, since all other error values are negative. For that date, 92.5% of the census sectors present error values within the range of -5.1 to -1 and only 7.1% are within the range of -1 to 1. In autumn, about 78.6% of the sectors present an error value within the range of -1 to 1, which indicates that the errors were smaller. Finally, in winter, the predicted LST was higher

than the real one in most census sectors due to the positive error values. Census sectors with error values within the range of -1 to 1 correspond to 44.6% in winter.

Table 5. Minimum and maximum values for Predicted LST and LST Error

Day	Predicted LST (ºC)		LST Error (ºC)	
	Min	Max	Min	Max
25 October 2019	34.65	38.12	- 5.11	1.60
04 May 2020	27.29	33.02	- 2.43	4.22
26 July 2021	26.19	34.43	- 3.48	5.61

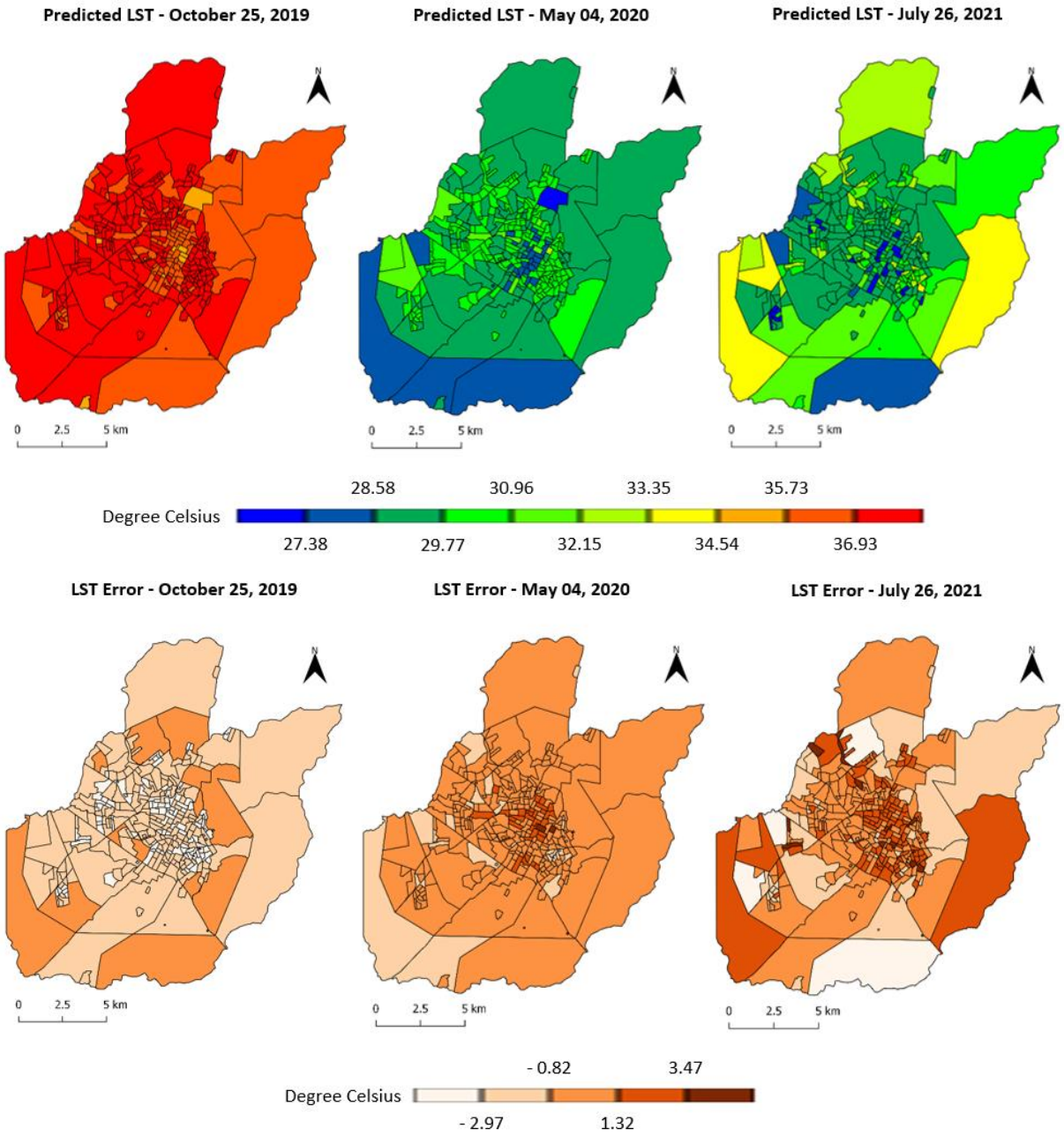

Figure 6. LST thematic maps showing the DT performance in the best model (model 1).

After the elaboration of the LST maps, the spatial distribution of the SUHI was mapped according to the real LST and the predicted LST as shown in Figure 7. Table 6 shows the variation of LST in SUHI and NON-SUHI areas.

During spring, SUHI areas are present in urban sectors where impervious surfaces are predominant. During autumn, SUHI are still concentrated in the urban perimeter, but they are also present in urban sectors with a high concentration of pervious surfaces. Conversely, during winter, an inversion occurs, as SUHI areas are identified in rural and urban sectors with a predominance of pervious surfaces, while non-SUHI areas correspond to urban sectors characterized by impervious surfaces. The presence of SUHI in rural areas and pervious surfaces during winter can be explained by the lack of water vapor, characteristic of the season, added to the fact that rural areas are characterized by exposed soil during this season.

The areas were identified as SUHI at 40.41 ºC for spring, 30.20 ºC for autumn and 29.67 ºC for winter. The mean LST of SUHI areas are greater than the mean LST of non-SUHI areas by 1.77 ºC for spring, 1.58 ºC for autumn and 2.35 ºC for winter.

Regarding the SUHI mapping based on the predicted LST, it is possible to notice that the algorithm overestimated the SUHI areas during the spring and underestimated it during the winter. The areas were identified as SUHI at 38.01 ºC for spring, 30.35 ºC for autumn and 30.45 ºC for winter. As the same way as the real LST, the mean of the predicted LST of SUHI areas are also greater than the mean LST of non-SUHI areas by 1.50 ºC for spring, 1.19 ºC for autumn and 2.53 ºC for winter.

A comparison between the real LST and the predicted LST in relation to the number of census sectors classified in the same class (SUHI or Non-SUHI) shows that the predicted LST was able to classify 65% of the sectors correctly during spring, 67% during autumn and 64% during winter. These values are directly related to the error values shown in the previous maps.

Table 6. LST values in SUHI and NON-SUHI areas

Day	Real LST – Min (ºC)		Real LST – Max (ºC)		Real LST – Mean (ºC)	
	SUHI	NON-SUHI	SUHI	NON-SUHI	SUHI	NON-SUHI
25 October 2019	40.41	35.12	43.12	40.40	41.05	39.28
04 May 2020	30.20	26.13	32.75	30.19	30.71	29.13
26 July 2021	29.67	26.11	33.33	29.66	30.70	28.35

Day	Predicted LST Min (ºC)		Predicted LST Max (ºC)		Predicted LST Mean (ºC)	
	SUHI	NON-SUHI	SUHI	NON-SUHI	SUHI	NON-SUHI
25 October 2019	38.01	34.83	38.11	37.17	38.02	36.52
04 May 2020	30.35	27.29	33.02	30.22	30.48	29.29
26 July 2021	30.45	26.19	34.43	30.37	31.38	28.85

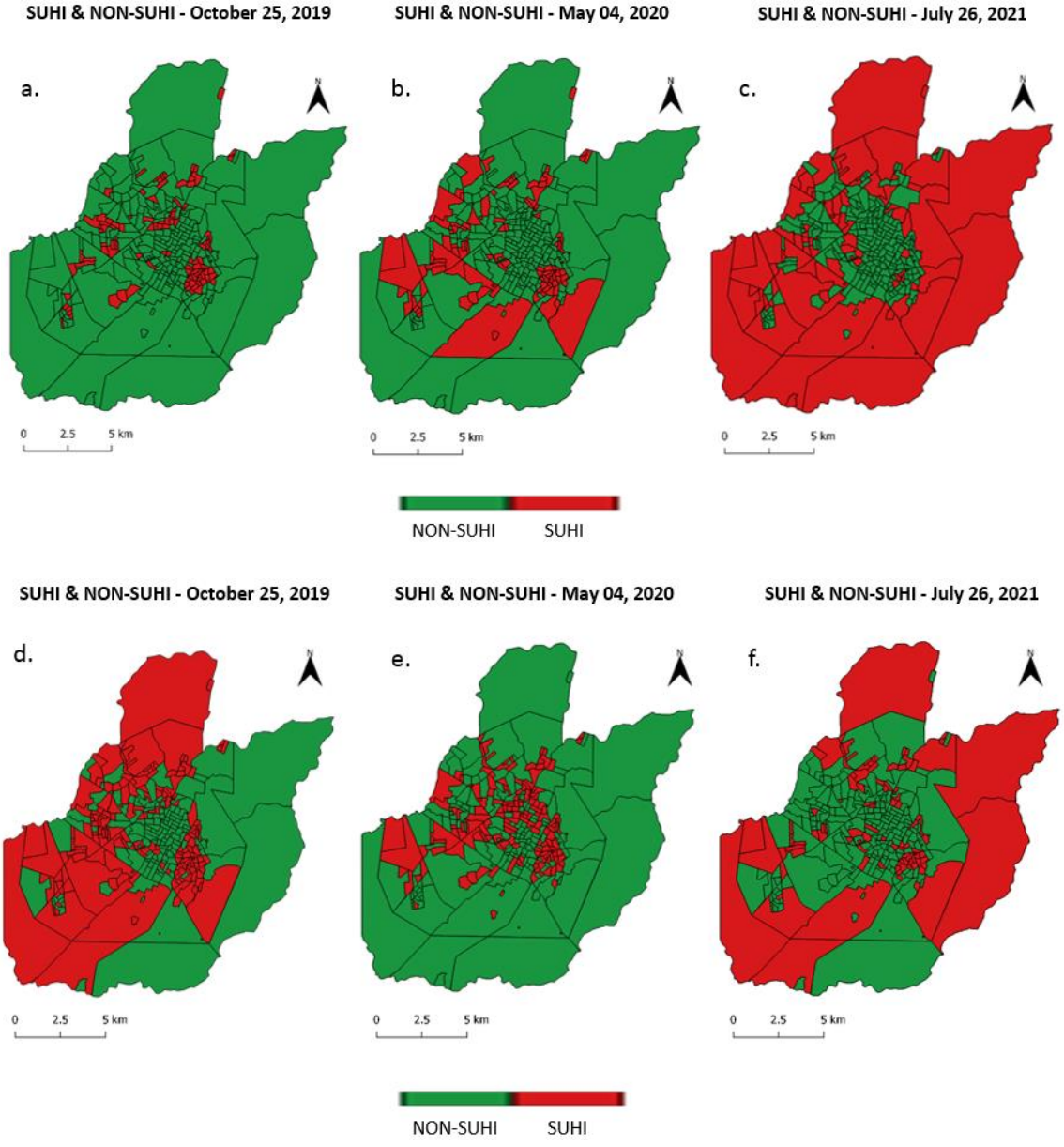

Figure 7. SUHI thematic maps according to the real LST (a, b, c) and the predicted LST (d, e, f).

Finally, the level of association between the attributes and the LST is shown in figure 8 and table 7. The pink color shows the attributes that are positively correlated to the LST and the blue color shows the attributes that are negatively correlated.

Based on the graphic, it is possible to notice a strong correlation of LST with the seasons, which highlights the importance of considering seasonality in temperature studies. Regarding the vegetation index, the graphic shows that high vegetation indexes indicate lower LST.

The Impervious Surfaces (ISA) and Built-up Area (NDBI) commonly present a positive correlation with LST, since the higher the ISA and NDBI indexes, the higher the LST. However, in this case, the negative correlation may be justified by the influence of the seasons, since as seen in the SUHI maps, impervious surfaces do not always have higher temperatures, as in winter, for example. Therefore, the data show the influence of seasons in urban climate studies one more time.

Regarding the socioeconomic variables, it is shown that the mean income has a negative correlation with LST, which means that high income has possibly more vegetation and, consequently, lower temperature. Likewise, population density presents a positive correlation with LST. The inclusion of these variables shows how the urban environment varies according to the economic characteristics of the place.

Based on the values presented in table 7, it is possible to notice that the seasons were the attribute with the highest correlation with temperature, with an approximate value of 0.51. The justification may be based on the high variability of the attribute, since there are dry and rainy periods throughout the year and the presence of water vapor directly influences the temperature. In this sense, the other attributes, such as socioeconomic and impermevious surfaces, tend to present constant characteristics or less variability, as is the case of vegetation. For this reason, they present correlation values within the range of -0.1 and 0.1, which indicates a lower influence when compared to seasonality.

Table 7. Values of Association between LST and the Attributes.

Attribute	Value	Attribute	Value	Attribute	Value
Season	0.51072	Population Density	0.00455	Mean NDBI	-0.02812
Min NDVI	0.06382	Number of Households	-0.00398	Max NDBI	-0.04543
SD NDBI	0.06162	Max NDVI	-0.00592	Min NDBI	-0.06475
Mean NDVI	0.03686	Median NDBI	-0.01248	Mean Income	-0.06852
Median NDVI	0.02902	SD NDVI	-0.01671	ISA	-0.07052

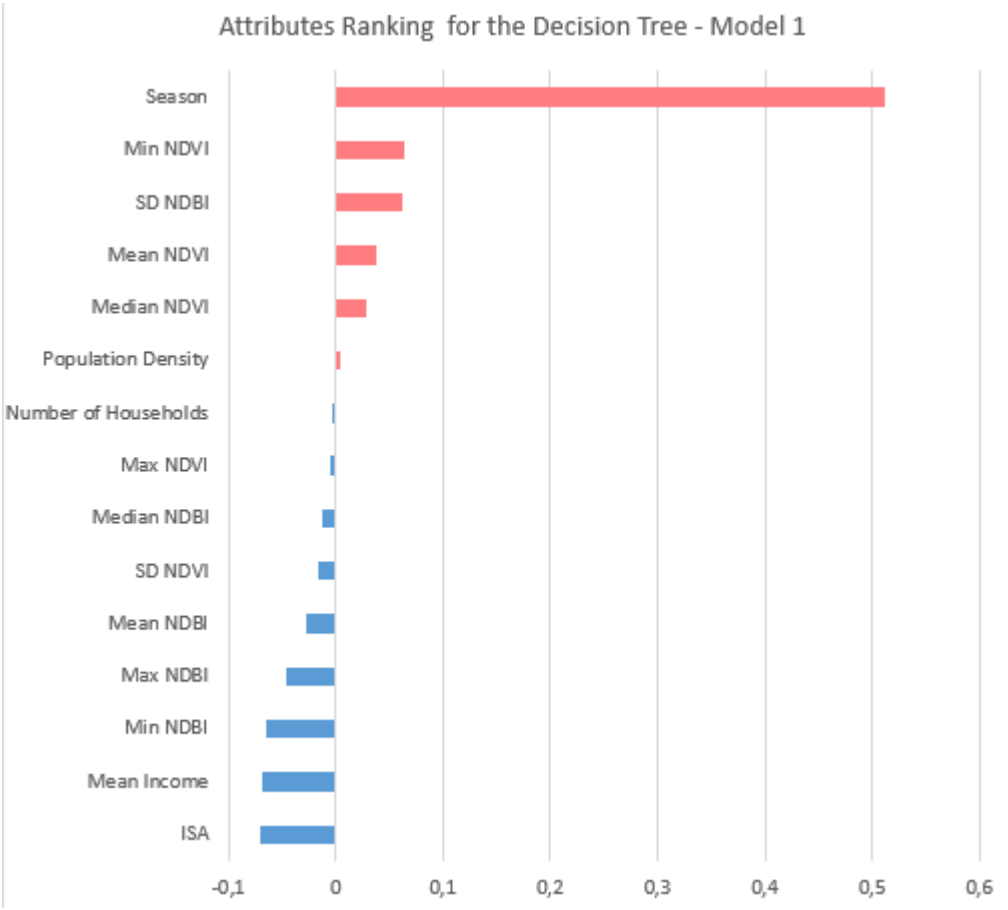

Figure 8. Level of Association between LST and the Attributes.

Discussion

The approach proposed in this study evaluated the performance of machine learning algorithms in estimating LST in order to optimize the SUHI characterization. The potential of the algorithms was measured based on environmental and socioeconomic variables given as input data for the models. Our interest in including

socioeconomic variables aimed to investigate the contribution of these variables in surface temperature, as the UHI is commonly characterized only by biophysical parameters, such as vegetation and built-up area index.

The application of machine learning algorithms is present in surface temperature (Osborne and Sanches, 2019), air temperature (Dos Santos, 2020; Vulova et al., 2020) and UHI studies (Yao et al., 2020). In all the studies mentioned, the decision tree-based algorithms outperformed the other algorithms. In our study, DT achieved better results due to lower MAE (1.49 °C) and RMSE (1.88 °C) values, in addition to high r (0.96). Furthermore, the correlation coefficient, when analyzed in isolation, obtained a higher value for the RF (table 3). To quantify the effects of urban form on LST, Sun et al. (2019) used seven variables, including NDVI. The study showed that the urban metrics used explain more than 90% of the temperature variations in the case of RF.

Regarding the amount of satellite images used in this study, the adoption of different dates to better characterize the LST in different seasons proved to be efficient. This is because considering isolated dates, in addition to not representing the real LST variation, in the case of the machine learning approach, it can present overfitting, as in the case of the 3B model. As for the seasons, models 3A and 3B, which considered only two seasons of the year, presented much lower correlation coefficients compared to the models containing all seasons. Choe and Yom (2019) also used multiple dates in order to obtain a more accurate model to estimate LST. In total, the work used 35 Landsat images between 2013 and 2018 considering all seasons. The data was applied in a deep learning model (deep multi-layer perceptron) and showed better accuracy during winter. Another study (Amorim, 2020) considers dry and rainy seasons to study UHI in Presidente Prudente. The study concludes that the intensity of UHI is greater during the dry seasons because the absence of water vapor increases the temperature difference between urban and rural areas, causing the phenomenon of UHI.

Despite obtaining relevant results, the present work was limited by the amount of available dates, especially during spring and summer. In addition, the socioeconomic variables analyzed were extracted from the last census in Brazil, in 2010. Another limitation is the characterization of SUHI only in the daytime period due to the data available for the study area.

Furthermore, despite the DT algorithm presenting satisfactory results in the LST prediction, it did not achieve the same performance in the SUHI mapping, since the mapping is done based on the mean and standard deviation of the LST, which makes the error values have a strong influence on the final result.

Recommendations for future studies would be to evaluate the occurrence of SUHI with more recent socioeconomic data and estimating LST with other techniques, which can be either changing the parameters of machine learning algorithms or applying deep learning techniques to increase the accuracy of the SUHI mapping.

The strategy proposed in this work proved to be a promising alternative in the detection and characterization of SUHI. One of the interests was to include socioeconomic variables in the analysis of surface temperature changes, as the landscape configuration is also influenced by these parameters (Yoo, 2018). Despite having achieved high accuracy, the inclusion of new data can help to improve the performance of models that did not show satisfactory values (Table 3). This improvement can happen through the addition of new dates or the inclusion of other algorithms (Dos Santos, 2020). Furthermore, by using data from satellite images and using machine learning, the procedure presented in this study is of low cost and faster than traditional methods, which means it can be adopted in the urban planning of any city.

Conclusions

This study evaluated the performance of six machine learning regression algorithms in estimating LST for characterization of SUHI in four different situations. The best performance was achieved by DT algorithm in the first situation (model 1), which considers 16 attributes, including environmental and socioeconomic variables, in addition to considering all seasons. Although LST achieves high accuracy, the characterization of SUHI can still be improved by changing parameters or applying other techniques. The contribution of this study is to state that decision tree-based algorithms, including RF, can be applied in other locations for similar situations. Whether for the SUHI or LST study, the high temporal and spatial resolution provided by satellite images contribute to a more robust approach, as it better represents the variation in temperature data. Future studies can benefit from the information

presented here as a diagnosis to improve urban planning and contribute to the formation of smart cities.

References

Ali, N.; Neagu, D.; Trundle, P. 2019. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. SN Appl. Sci. 1, 1–15, doi:10.1007/s42452-019-1356-9.

Amorim, M. C. de C. T. (2020). Daily evolution of urban heat islands in a Brazilian tropical continental climate during dry and rainy periods. Urban Climate, 34, 100715.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., & Wood, E. F. (2018). Present and future Köppen-Geiger climate classification maps at 1-km resolution. Scientific data, 5(1), 1-12.

Brazilian Institute of Geography and Statistics (IBGE), 2010. Census 2010. Available online: http://ibge.gov.br/ (Accessed 22 February 2021).

Brazilian Institute of Geography and Statistics (IBGE), 2021. Available online: https://www.ibge.gov.br/cidades-e-estados/sp/presidente-prudente.html (Accessed on 25 August 2021).

Buyantuyev, A., & Wu, J. (2010). Urban heat islands and landscape heterogeneity: linking spatiotemporal
variations in surface temperatures to land-cover and socioeconomic patterns. Landscape Ecology, 25(1),
17–33.

Carrasco, R. A., Pinheiro, M. M. F., Junior, J. M., Cicerelli, R. E., Silva, P. A., Osco, L. P., & Ramos, A. P. M. (2020). Land use/land cover change dynamics and their effects on land surface temperature in the western region of the state of São Paulo, Brazil. Regional Environmental Change, 20(3).

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247–1250.

Choe, Y.-J., & Yom, J.-H. (2019). Improving accuracy of land surface temperature prediction model based on deep-learning. Spatial Information Research.

Dos Santos, R. S. (2020). Estimating spatio-temporal air temperature in London (UK) using machine learning and earth observation satellite data. International Journal of Applied Earth Observation and Geoinformation, 88, 102066.

Ebrahimy, H., & Azadbakht, M. (2019). Downscaling MODIS land surface temperature over a heterogeneous area: An investigation of machine learning techniques, feature selection, and impacts of mixed pixels. Computers & Geosciences.

Gomez-Martinez, F., de Beurs, K. M., Koch, J., & Widener, J. (2021). Multi-Temporal Land Surface Temperature and Vegetation Greenness in Urban Green Spaces of Puebla, Mexico. Land, 10(2), 155.

Guha, S., Govil, H., & Mukherjee, S. (2017). Dynamic analysis and ecological evaluation of urban heat islands in raipur city, india. Journal of Applied Remote Sensing, 11(3), 36020.

Guha, S., Govil, H., Dey, A., & Gill, N. (2018). Analytical study of land surface temperature with NDVI and NDBI using Landsat 8 OLI and TIRS data in Florence and Naples city, Italy. European Journal of Remote Sensing, 51(1), 667–678.

Isaya Ndossi, M.; Avdan, U. Application of Open Source Coding Technologies in the Production of Land Surface Temperature (LST) Maps from Landsat: A PyQGIS Plugin. Remote Sens. 2016, 8, 413.

Jenerette, G. D., Harlan, S. L., Brazel, A., Jones, N., Larsen, L., & Stefanov, W. L. (2006). Regional relationships between surface temperature, vegetation, and human settlement in a rapidly urbanizing ecosystem. Landscape Ecology, 22(3), 353–365.

Kafy, A.-A., Abdullah-Al-Faisal, Rahman, M. S., Islam, M., Rakib, A. A., Islam, M. A., … Sattar, G. S. (2020). Prediction of seasonal urban thermal field variance index using machine learning algorithms in Cumilla, Bangladesh. Sustainable Cities and Society, 102542.

Li, H., Zhou, Y., Li, X., Meng, L., Wang, X., Wu, S., & Sodoudi, S. (2018). A new method to quantify surface urban heat island intensity. Science of The Total Environment, 624, 262–272.

Li, J., Song, C., Cao, L., Zhu, F., Meng, X., & Wu, J. (2011). Impacts of landscape structure on surface urban heat islands: A case study of Shanghai, China. Remote Sensing of Environment, 115(12), 3249–3263.

Linares-Rodriguez, A., Ruiz-Arias, J. A., Pozo-Vazquez, D., & Tovar-Pescador, J. (2013). An artificial neural network ensemble model for estimating global solar radiation from Meteosat satellite images. Energy, 61, 636–645.

Ma, Y., Kuang, Y., & Huang, N. (2010). Coupling urbanization analyses for studying urban thermal environment and its interplay with biophysical parameters based on TM/ETM+ imagery. International Journal of Applied Earth Observation and Geoinformation, 12(2), 110–118.

Marques Ramos, A. P., Prado Osco, L., Elis Garcia Furuya, D., Nunes Gonçalves, W., Cordeiro Santana, D., Pereira Ribeiro Teodoro, L., … Pistori, H. (2020). A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices. Computers and Electronics in Agriculture, 178, 105791.

Mathew, A., Chaudhary, R., & Gupta, N. (2015). Study of Urban heat Island Effect on Ahmedabad City and Its Relationship with Urbanization and Vegetation Parameters. International Journal of Computer & Mathematical Sciences.

Osborne, P. E., Alvares-Sanches, T. (2019). Quantifying how landscape composition and configuration affect urban land surface temperatures using machine learning and neutral landscapes. Computers, Environment and Urban Systems, 76, 80–90.

Sentinel; European Space Agency (ESA). Level-2A Algorithm Overview; ESA Standard Document; ESA: Paris, France, 2015. Available online: https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-2-msi/level-2a/algorithm (Accessed on 30 September 2021).

Singh, D., & Singh, B. (2019). Investigating the impact of data normalization on classification performance. Applied Soft Computing, 105524.

Sobrino, J.; Jiménez-Muñoz, J. C. & Paolini, L., 2004. Land surface temperature retrieval from LANDSAT TM 5. Remote Sensing of Environment, 90, 434-440.

Štepanovský, M.; Ibrová, A.; Buk, Z.; Velemínská, J. 2017. Novel age estimation model based on development of permanent teeth compared with classical approach and other modern data mining methods. Forensic Sci. Int., 279, 72–82

Sun, Y., Gao, C., Li, J., Wang, R., & Liu, J. (2019). Quantifying the Effects of Urban Form on Land Surface Temperature in Subtropical High-Density Urban Areas Using Machine Learning. Remote Sensing, 11(8), 959.

Tang, J., Di, L., Xiao, J., Lu, D., & Zhou, Y. (2017). Impacts of land use and socioeconomic patterns on urban heat Island. International Journal of Remote Sensing, 38(11), 3445–3465.

USGS (United States Geological Service) (2019) Landsat 8 (L8) data users handbook. Department of the Interior U.S. Geological Survey. EROS Sioux Falls, South Dakota.

Voogt, J., & Oke, T. (2003). Thermal remote sensing of urban climates. Remote Sensing of Environment, 86(3), 370–384.

Vulova, S. V., Meier, F., Fenner, D., Nouri, H., & Kleinschmit, B. (2020). Summer Nights in Berlin, Germany: Modeling Air Temperature Spatially With Remote Sensing, Crowdsourced Weather Data, and Machine Learning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 1–1.

Weng, Q., Lu, D., & Schubring, J. (2004). Estimation of land surface temperature–vegetation abundance relationship for urban heat island studies. Remote Sensing of Environment, 89(4), 467–483.

Weng, Q., Rajasekar, U., & Hu, X. (2011). Modeling Urban Heat Islands and Their Relationship with Impervious Surface and Vegetation Abundance by Using ASTER Images. IEEE Transactions on Geoscience and Remote Sensing, 49(10), 4080–4089.

Yang, C., He, X., Yan, F., Yu, L., Bu, K., Yang, J., Chang, L. & Zhang, S. (2017). Mapping the Influence of Land Use/Land Cover Changes on the Urban Heat Island Effect—A Case Study of Changchun, China. Sustainability, 9(2), 312.

Yao, Y., Chang, C., Ndayisaba, F., & Wang, S. (2020). A New Approach for Surface Urban Heat Island Monitoring based on Machine Learning Algorithm and Spatiotemporal Fusion Model. IEEE Access, 1–1.

Ying, Xue. An overview of overfitting and its solutions. In: Journal of Physics: Conference Series. IOP Publishing, 2019. p. 022022.

Yoo, S. (2018). Investigating important urban characteristics in the formation of urban heat islands: a machine learning approach. Journal of Big Data, 5(1).

Yuan, F., & Bauer, M. E. (2007). Comparison of impervious surface area and normalized difference vegetation index as indicators of surface urban heat island effects in Landsat imagery. Remote Sensing of Environment, 106(3), 375–386

Zha, Y., Gao, J., & Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. International Journal of Remote Sensing, 24(3), 583–594.

Zhang, X., Zhong, T., Wang, K., & Cheng, Z. (2009). Scaling of impervious surface area and vegetation as indicators to urban land surface temperature using satellite data. International Journal of Remote Sensing, 30(4), 841–859.

3 CONSIDERAÇÕES FINAIS

O presente trabalho verificou o desempenho de algoritmos de aprendizagem de máquina em prever a LST com base em variáveis ambientais e socioeconômicas. A partir dos resultados dos algoritmos foi possível a criação de mapas temáticos apresentando a distribuição espacial da LST estimada, bem como a distribuição espacial do erro do algoritmo, calculado pela diferença de valor entre a LST estimada e a real.

O cálculo da distribuição espacial da SUHI foi feito com base na média e no desvio padrão da LST real e estimada. O mapeamento da SUHI com base na LST estimada alcançou acertos de 64% a 67% quando comparados à classificação feita pela LST real. Portanto para a caracterização da SUHI é possível aplicar novos parâmetros ou novas técnicas, como deep learning, para melhorar a predição. Quanto às variáveis abordadas no estudo verificou-se que as estações do ano apresentam forte correlação com a LST.

O estudo é relevante para o planejamento das cidades, pois oferece informações a respeito de diferentes parâmetros que alteram a configuração da paisagem urbana. Com base nas informações aqui apresentadas é possível identificar o fator de maior impacto nas mudanças de temperatura que acabam desencadeando fenômenos climáticos como as ilhas de calor urbano (UHI). Por se tratar de uma questão climática e urbana, o estudo pode beneficiar diferentes cidades, uma vez que pode ser replicado em outras localidades, além dos dados poderem auxiliar prefeituras na gestão urbana.