JOURNAL OF ENVIRONMENTAL ENGINEERING AND LANDSCAPE MANAGEMENT
ISSN 1648–6897 / eISSN 1822-4199
2017 Volume 25(03): 241–250
https://doi.org/10.3846/16486897.2016.1223084

FEATURE SELECTION OF VARIOUS LAND COVER INDICES FOR MONITORING SURFACE HEAT ISLAND IN TEHRAN CITY USING LANDSAT 8 IMAGERY

Nikrouz MOSTOFI[a], Mahdi HASANLOU[b]

[a]*Department of Surveying Engineering, Islamic Azad University, South Tehran Branch, 1777613511 Tehran, Iran*

[b]*School of Surveying and Geospatial Engineering, Norh Karegar, College of Engineering, University of Tehran, 1439957131 Tehran, Iran*

Submitted 10 Nov. 2015; accepted 08 Aug. 2016

Abstract. Recently, scientists have been taking a great interest in Global warming issue, since the global surface temperature has been significantly increased all through last century. The surface heat island (SHI) refers to an urban area that has higher surface temperatures than its surrounding rural areas due to urbanization. In this paper, Tehran city is used as case study area. This paper tries to employ a quantitative approach to explore the relationship between land surface temperature and the most widespread land cover indices, and select proper (urban and vegetation) indices by incorporating supervised feature selection procedures using Landsat 8 imageries. In this regards, genetic algorithm is incorporated to choose best indices by employing kernel base one, support vector regression and linear regression methods. The proposed method revealed that there is a high degree of consistency between affected information and SHI dataset (*RMSE* = 0.9324, *NRMSE* = 0.2695 and R^2 = 0.9315).

Keywords: Surface heat island, Land use/cover, Support vector regression, linear regression model, Genetic algorithm.

Introduction

The surface temperature is a substantial factor in the study of urban climatology. It changes the air temperature of the lowest layers of the urban area. The surface temperature is also effective in determining the internal climates of buildings and disturbs the energy exchanges that impact the comfort of city life (Voogt, Oke 2003). The urban heat island (UHI) refers to an urban area that has higher surface temperatures than its surrounding rural areas due to urbanization (Xian, Crane 2006). The annual average air temperature of an urban area, with almost one million population, can be one to three degree warmer than its surrounding areas. This phenomenon can affect societies by increasing summertime, air conditioning costs, air pollution, heat related illness, greenhouse gas emissions and water quality. Today, more than fifty percent of the world's population are living in cities (UN DESA 2015), in this regard, urbanization has become a key factor in global warming issue. Tehran, the capital of Iran, one of megacities in the world, is the case study of this research. A megacity is mainly defined as a residential area with a total population in excess of ten million people (Dihkan *et al.* 2015). We have been encountering significant surface heat island (SHI) effect in this area due to rapid urbanization progress and the fact that twenty percent of population in Iran are currently living in Tehran houses.

SHI has been usually monitored and measured by in situ observations acquired from thermometer networks. Recently, remotely sensed observing and monitoring of SHIs has become accessible by incorporating thermal remote sensing technology and satellite data. Satellite thermal imageries, mainly high resolution imageries, have the advantage of providing a repeatable dense grid of temperature data, over a whole urban area, and even distinctive temperatures for individual buildings.

Previous studies of land surface temperatures (LST) and thermal remote sensing of urban and rural areas have been primarily conducted by using AVHRR or MODIS imageries (Streutker 2002; Imhoff *et al.* 2010). Now a days, most of researchers are using high resolution satellite imagery to monitor thermal anomalies in urban areas (Fabrizi *et al.* 2010; Ogashawara, Bastos 2012; Liu *et al.* 2015).

Corresponding author: Mahdi Hasanlou
E-mail: hasanlou@ut.ac.ir

Taylor & Francis
Taylor & Francis Group

In this study, newly launched Landsat series (Landsat 8) is used to monitor SHI, and retrieve the brightness temperatures and land use/cover types. The Landsat 8 carries two kind of sensors (Landsat 8 2016): The Operational Land Imager (OLI) sensor has former Landsat bands, with three new bands: a deep blue band for aerosol/coastal investigations (band 1), a shortwave infrared band for cirrus detection (band 9), and a Quality Assessment (AQ) band. The Thermal Infrared Sensor (TIRS) provides two high spatial resolution thirty meters thermal bands (band 10 and 11). These sensors both use corrected signal-to-noise ratio (SNR) radiometric performance quantized over a 12-bit dynamic range. Improved SNR performance would cause better determination of land cover type. Details of Landsat 8 band specification is illustrated in Table 1. Furthermore, Landsat 8 imageries incorporate two valuable thermal imagery bands with 10.9 μm and 12.0 μm wavelength. These two thermal bands would improve estimation of SHI by incorporating split-window algorithms, they will also increase the probability of the SHI and urban-modified climates to be monitored. Therefore, it is necessary to design and use new procedures that are able to simultaneously (a) handle the two new high resolution thermal bands of Landsat 8 imagery and (b) incorporate satellite in situ measurement into precise estimation of SHI.

Table 1. Landsat 8 OLI and TIRS bands

Bands	Wavelength (μm)	Res (m)
Band 1 – Coastal aerosol	0.43–0.45	30
Band 2 – BLUE	0.45–0.51	30
Band 3 – GREEN	0.53–0.59	30
Band 4 – RED	0.64–0.67	30
Band 5 - Near Infrared (NIR)	0.85–0.88	30
Band 6 – SWIR 1	1.57–1.65	30
Band 7 – SWIR 2	2.11–2.29	30
Band 8 – Panchromatic	0.50–0.68	15
Band 9 – Cirrus	1.36–1.38	30
Band 10 – Thermal Infrared (TIRS) 1	10.60–11.19	30
Band 11 – Thermal Infrared (TIRS) 2	11.50– 2.51	30

The UHIs can be affected by three main factors (Ogashawara, Bastos 2012): a) reduced evapotranspiration; b) the effects of energy transformation in urban area; and (c) anthropogenic energy production. Also, according to (Actionbioscience 2015), there are three types of UHIs: a) Boundary Layer Heat Island (BLHI); b) Canopy Layer Heat Island (CLHI); and c) Surface Heat Island (SHI). The main difference between BLHI and SHI is that BLHI refers to the warmth of the urban atmosphere while SHI refers to the warmth of the urban surface. Also, the major difference between CLHI and SHI is the place where temperature is appeared and detected. Usually, CLHI is detected by specified air temperature measurement (i.e. in situ data) in the canopy layer, while remotely sensed thermal data observe the spatial patterns and models of upwelling thermal radiance to estimate the LST (Voogt, Oke 2003) of the SHI.

Lately, quantitative algorithms for urban thermal environment and dependent factors have been studied, for example, the relationship of UHI with land cover types and its corresponding regression model (Xian, Crane 2006; Hasanlou, Mostofi 2015; Liu *et al.* 2015; Odindi *et al.* 2015). Similar works have been done and models of the relation between various vegetation indices and the surface temperature have been established (Chen *et al.* 2006; Xiong *et al.* 2012). This paper tried to employ a quantitative approach to track the relationship between SHI and common land cover indices and select proper indices, including the Normalized Difference Vegetation Index (NDVI) (Kriegler *et al.* 1969), Enhanced Vegetation Index (EVI) (Kriegler *et al.* 1969), Soil Adjusted Vegetation Index (SAVI) (Huete 1988), Normalized Difference Water Index (NDWI) (Gao 1996), Normalized Difference Bareness Index (NDBaI) (Zhao, Chen 2005; Chen *et al.* 2006), Normalized Difference Build-up Index (NDBI) (Zha *et al.* 2003), Modified Normalized Difference Water Index (MNDWI) (Xu 2006), Bare soil Index (BI) (Zha *et al.* 2003; Zhao, Chen 2005). Urban Index (UI) (Kawamura *et al.* 1996), Index based Built up Index (IBI) (Xu 2008) and Enhanced Built up and Bareness Index (EBBI) (Assyakur *et al.* 2012). Behind these indices, the tasselled cap transformation (TCT), which is calculated for Landsat 8 imagery, is used to compact spectral data into a few bands associated with physical scene characteristics with minimal information loss (Baig *et al.* 2014). The three TCT components, Brightness, Greenness and Wetness, are computed and incorporated to predict SHI effect. Therefore, the main objectives of this research are to develop a non-linear and kernel base analysis model for urban thermal environment area by incorporating support vector regression (SVR) method (Drucker *et al.* 1997), and also to compare proposed method with linear regression model (LRM) in which linear combination of incorporated land cover indices (features) is used. The primary aim of this paper is to establish a framework producing an optimum SHI by utilizing proper land cover indices form Landsat 8 imagery. In this regard, three scenarios have been implemented: a) incorporating LRM with full feature set without any feature selection; b) incorporating SVR with full feature set without any feature selection; and c) incorporating genetically selected suitable features in SVR method (GA-SVR). The results of this study can be used to increase the output performance of the SHI estimation in urban area using Landsat 8 imagery by adopting the

genetically SVR method with (a) an optimal land cover indices/feature space and (b) customized SVR parameters.

1. Material and methods

In this paper, Landsat 8 imagery is used as input imagery data to estimate SHI map, and also various urban and vegetation indices are calculated. Figure 1 shows the flowchart of proposed methods.

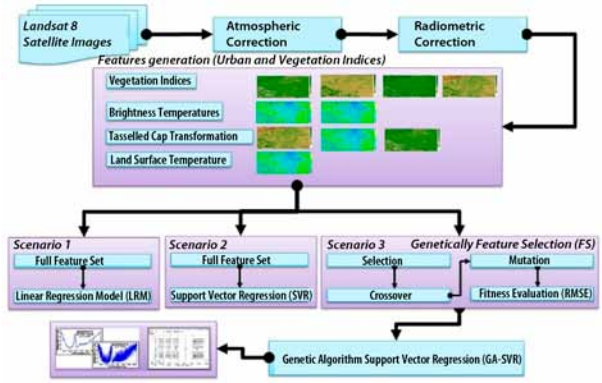

Fig. 1. Flowchart of proposed methods

As shown in Figure 1. atmospheric and radiometric correction is done after importing Landsat 8 images. In utilized procedure, the fast line-of-sight atmospheric analysis of spectral hypercubes (FLAASH) algorithm is used to reduce atmospheric and radiometric effect on incorporated images. The next step starts by producing urban and vegetation indices and also brightness temperature (BT) of thermal bands of Landsat 8 imagery in 10.9 μm and 12.0 μm (band 10, 11) (USGS 2015). Further, in this step, we calculate LST data for incorporated dataset using split window (SW) algorithm (Eq. (1)) introduced in (Jimenez-Munoz *et al.* 2014):

$$LST = c_0 + T_{b10} + c_1\left(T_{b10} - T_{b11}\right) + c_1\left(T_{b10} - T_{b11}\right)^2 + \left(c_3 + c_4\omega\right)\left(1-\varepsilon\right) + \left(c_5 + c_6\omega\right)\Delta\varepsilon, \quad (1)$$

where T_{b10} and T_{b11} are the at-sensor BTs (in kelvins), ε is the mean emissivity, $\varepsilon = 0.5\left(\varepsilon_{b10} + \varepsilon_{b11}\right)$, $\Delta\varepsilon$ is the emissivity difference, $\Delta\varepsilon = \left(\varepsilon_{b10} - \varepsilon_{b11}\right)$, ω is the total atmospheric water vapor content (in g×cm^{-2}) that we set to $\omega = 3$ as mentioned in (Jimenez-Munoz *et al.* 2014), and c_0 to c_6 are the SW coefficients that computed in (Jimenez-Munoz *et al.* 2014). In this study, to estimate ε_{b10} and ε_{b11}, we have simultaneously incorporated MODIS product (MOD11A2 2015) for the same area. As you know, the MOD11A2 is the level-3 MODIS global LST and emissivity, and 8-day data are composed from the daily 1-kilometer LST product. Also this product comprises day time and night time LSTs, quality assessment, observation times, view angles, bits of clear-sky days and nights, and emissivities estimated in bands 31 and 32 from land cover types that are similar to Landsat 8 thermal bands (ε_{b10} and ε_{b11}). Then, the last step is divided in three different scenarios, including LRM implementation (scenario 1), SVR incorporation (scenario 2) and GA-SVR implementation that adopted genetic algorithm (GA) as supervised feature selection (FS) procedure (scenario 3). In the next section, procedure of computing urban and vegetation indices behind utilized algorithm will be discussed.

1.1. Calculating TOA radiance/reflectance and at-sensor BT

The OLI and TIRS bands data can be converted to top of atmosphere (TOA) spectral radiance and planetary reflectance using the reflectance/radiance rescaling parameters provided in the product metadata file (MTL file) (USGS 2015). The following equation (Eq. (2) and Eq. (3)) is used to convert digital number (DN) values to TOA radiance/reflectance for OLI data as follows:

$$L_\lambda = M_L DN + A_L; \quad (2)$$

$$\rho_{\lambda'} = M_P DN + A_P, \quad (3)$$

where L_λ is TOA spectral radiance (watts/(m^2×srad×μm)), M_L and M_P are band-specific multiplicative rescaling factor extracted from MTL file, A_L and A_P are band-specific additive rescaling factor extracted from MTL file, DN is quantized and calibrated standard digital number values and $\rho_{\lambda'}$ is TOA planetary reflectance, without correction for solar angle. Also, we can correct TOA reflectance by considering the sun angle (Eq. (4)):

$$\rho_\lambda = \frac{\rho_{\lambda'}}{\sin\left(\theta_{SE}\right)}, \quad (4)$$

where ρ_λ is TOA planetary reflectance and θ_{SE} is local sun elevation angle. The scene center sun elevation angle in degrees is extracted from MTL file. For TIRS bands data, we use conversion from spectral radiance to BT using the thermal constants provided in MTL file which derived from Planck's law (Eq. (5)):

$$BT = K_2 / \ln\left(1 + K_1 / L_\lambda\right), \quad (5)$$

where BT is at-sensor BT in kelvin and K_1 / K_2 are band-specific thermal conversion constant extracted from MTL file for each TIRS bands (USGS 2015). By utilizing Eq. (1) to Eq. (5), input features and land cover indices will be produced.

1.2. Urban and vegetation indices

In this study, the most common urban and vegetation indices are used. These indices can be divided to two main type: a) urban indices; and b) vegetation indices. The widespread and common urban indices are shown in the Table 2. The most of these indices would extract urbanization parameters related to spectral difference of near infrared, visible and short wave infrared bands of Landsat 8 Imagery

Table 2. Extracted urban indices form Landsat 8 imagery

Name of urban index	Formulation
Normalized Difference Bareness Index (NDBaI)	$NDBaI = \frac{SWIR1 - TIRS1}{SWIR1 + TIRS1}$
Normalized Difference Build-up Index (NDBI)	$NDBI = \frac{SWIR1 - NIR}{SWIR1 + NIR}$
Bare Soil Index (BI)	$BI = \frac{(SWIR1 + RED) - (NIR + BLUE)}{(SWIR1 + RED) + (NIR + BLUE)}$
Urban Index (UI)	$UI = \frac{SWIR2 - NIR}{SWIR2 + NIR}$
Index-based Built-Up Index (IBI)	$IBI = \frac{\frac{2 \times SWIR1}{SWIR1 + NIR} - \left(\frac{NIR}{NIR + RED} - \frac{GREEN}{GREEN + SWIR1}\right)}{\frac{2 \times SWIR1}{SWIR1 + NIR} + \left(\frac{NIR}{NIR + RED} - \frac{GREEN}{GREEN + SWIR1}\right)}$
Enhanced Built-Up and Bareness Index (EBBI)	$EBBI = \frac{SWIR1 - NIR}{10\sqrt{SWIR1 + TIRS1}}$

Table 3. Extracted vegetation indices form Landsat 8 imagery

Name of vegetation index	Formulation
Normalized Difference Vegetation Index (NDVI)	$NDVI = \frac{NIR - RED}{NIR + RED}$
Enhanced Vegetation Index (EVI)	$EVI = G\frac{NIR - RED}{NIR + C_1 RED - C_2 BLUE + L}$ $L = 1; C_1 = 6; C_2 = 7.5; G = 2.5$
Soil Adjusted Vegetation Index (SAVI)	$SAVI = \frac{NIR - RED}{NIR + RED + L}(L + 1)$ $0 < L < 1 \Rightarrow L = 0.5$
Normalized Difference Water Index (NDWI)	$NDWI = \frac{NIR - SWIR1}{NIR - SWIR1}$
Modified Normalized Difference Water Index (MNDWI)	$MNDWI = \frac{GREEN - NIR}{GREEN + NIR}$
Tasselled Cap Transformation (TCT)	Brightness
TCT	Greenness
TCT	Wetness

(Table 1). All indices from Table 2 are calculated based on incorporating digital number (DN) of Landsat 8 bands.

As before, in Table 3 computable vegetation indices extracted from spectral bands like, near infrared, visible and short wave infrared bands of incorporated dataset are illustrated. All these indices are calculated based on incorporating reflectance/radiance of related Landsat 8 bands by using procedure introduced in (USGS 2015).

1.3. Linear regression model

Given a data set $\{y_i, x_{i1}, \ldots, x_{iq}\}_{i=1}^{n}$ of n statistical units, a linear regression model (LRM) assumes that the relationship between the dependent parameter y_i and the q vector of regressors x_i is linear. This relationship is modeled through a disturbance term or error parameter ε_i – an unobserved random parameter that adds noise to the linear relationship between the dependent parameter (in this study SHI data) and regressors (in this study land cover indices) (Kutner *et al.* 2004). Thus the model takes the form (Eq. (6)):

$$y_i = \beta_i x_{i1} + \ldots + \beta_q x_{iq} + \varepsilon_i = X_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n. \tag{6}$$

Where T defines the transpose, therefore $X_i^T \beta$ is the inner product between vectors x_i and β. Often these n equations are stacked together and written in vector form as (Eq. (7)):

$$y = X\beta + \varepsilon, \tag{7}$$

$x_{i1}, \ldots, x_{iq}$, are called regressors or independent variables (in this study land cover indices). The matrix X is sometimes called the design matrix. y_i, is called the measured variable or dependent variable (in this study SHI data). The criteria as to which variable in a dataset is modeled as the dependent parameter and which is modeled as the independent parameter may be based on an assumption that the value of one of the parameters is caused by, or directly influenced by the other variables. As an alternative, there may be a reason to model one of the parameters in terms of the others, in which case there would be no need for presumption of any causes. β, is a q dimensional vector. Its elements are also called regression coefficients. Statistical prediction and conclusion in linear regression focuses on β. The elements of this parameter vector are perceived as the partial derivatives of the dependent parameter with respect to the various independent parameters.

1.4. Support vector regression

The SVR is a supervised learning method which emerged in late 1970s (Drucker *et al.* 1997). SVR allows computing a strong nonparametric model of the relationship between urban/vegetation indices and SHI change. This method is also widely used in most remote sensing

applications like SST and LST estimation (Moser, Serpico 2009), biophysical parameter estimation and other vegetation index monitoring from multispectral satellite images (Durbha *et al.* 2007). Consider a set of training points, $\{(x_1,z_1),\ldots,(x_l,z_l)\}$, where $x_i \in \mathbb{R}^n$ is a feature vector and $z_i \in \mathbb{R}^1$ is the target output. Under given parameters $C>0$ and $\in>0$ the standard form of support vector regression (Eq. (8)) (Drucker *et al.* 1997; Smola, Schölkopf 2004) is:

$$\min_{\omega,b\xi,\xi^*} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right)$$

subject to $\omega^T\phi(x_i)+b-z_i \le \in +\xi_i$

$$z_i - \omega^T\phi(x_i) - b \le \in +\xi_i^*;$$
$$\xi_i,\xi_i^* \ge 0,\ i=1,\ldots,l. \tag{8}$$

By introducing Lagrange multipliers and exploiting the optimality constraints, the decision function has the following explicit form (Eq. (9)):

$$\sum_{i=1}^{l}\left(\alpha_i+\alpha_i^*\right)K(x_i,x)+b,$$
$$0\le\alpha_i\le C,\ 0\le\alpha_i^*\le C. \tag{9}$$

Where l is the number of support vectors (SVs) and the kernel function (Eq. (10)):

$$K(x_i,x)=\sum_{j=1}^{m}\phi_j(x)\phi_j(x_i) \tag{10}$$

and α_i^* are Lagrange multipliers. In order to run the SVR method with high efficiency, some constraint must be considered, (a) tuning the SVR parameters including $C, \in$ and kernel parameters, (b) optimizing input space (selecting suitable features). In this paper, we are focusing on both constraints by utilizing search procedure to tune the parameters and feature selection method, and to optimize input space (Genetic Algorithm).

1.5. Genetic algorithm

Genetic algorithm (GA) is the most widespread technique among evolutionary algorithms. This method allows us to search potential solutions to optimize problems in reasonable time, particularly when the search space area is very extensive (Goldberg, Holland 1988). This method is heuristic, based on population of individuals (e.g., chromosomes), that each individual performs a candidate solution (Goldberg, Holland 1988) and can be illustrated as a bit in string mode. Each individual is evaluated by fitness function. This function measures the quality of an individual. In GA method, the population commences randomly, or by different strategy based on the problem in question. The population answers some number of evolutions. During GA's process individuals are evolved and reproduced using GA's operations such as: mutation, crossover, and selection. Its main goal is to select and find the individual with the best fitness value (Goldberg, Holland 1988).

We modeled the problem of feature/index selection as follows: each individual has a size of N features/genes, as shown in Figure 2, and each gene represents a binary random vector number including 0s and 1s, associated with a features/indices. Bit strings of 0s and 1s are chosen for coding. Since N features form one combination, chromosome is arranged as comprised of N individual feature sequence numbers, which are arranged in a serial mode. The length of each parameter is automatically detected according to the number of features/indices in the data set (Fig. 2). As previously stated, the quality of each candidate solution is evaluated according to a fitness function. Our fitness function here is the root mean square error (RMSE) of a regression performed by a SVR or LRM method. Stochastic two-point was used as the selection operator. Additionally, Gaussian mutation, as well as uniform crossover were used as GA operators. Finally, migration direction was set to forward mode in addition to the elitism mechanism.

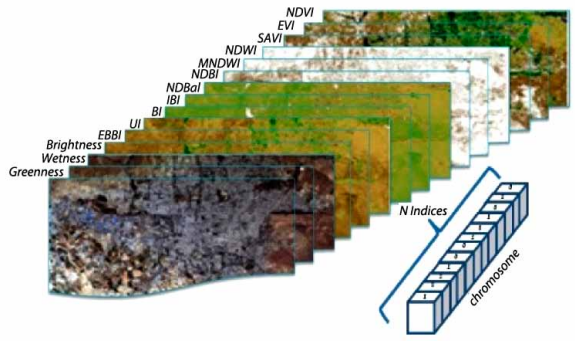

Fig. 2. Chromosome format in genetically feature selection

2. Results

In this study, Landsat 8 imagery acquired from Tehran city is used as dataset. This dataset belongs to summer time (June 15, 2014), which air temperature was nearly 40 °C (Fig. 3). As mentioned in previous section, by incorporating Landsat 8 LST retrieval algorithm (Eq. 1) and contemporary MODIS product estimating emissivity of two thermal bands (i.e. ε_{b10} and ε_{b11}), SHI is estimated. Urban, vegetation and TCT (Brightness, Greenness and Wetness) indices from DN/Ref Landsat 8 images are calculated as well. Calculated indices and information using this dataset is shown in Figure 3.

To establish models in all three scenarios, 2400 points of in situ data were extracted from Tehran urban region. Then, 720 (30% of data) random points were selected as training data and 1680 (70% of data) random points as testing data. Also, some common criteria like, mean square error (MSE), root mean square error (RMSE), Normalized root mean square error (NRMSE) and R-squared

(R^2) were used to examine the output result of each method and scenario.

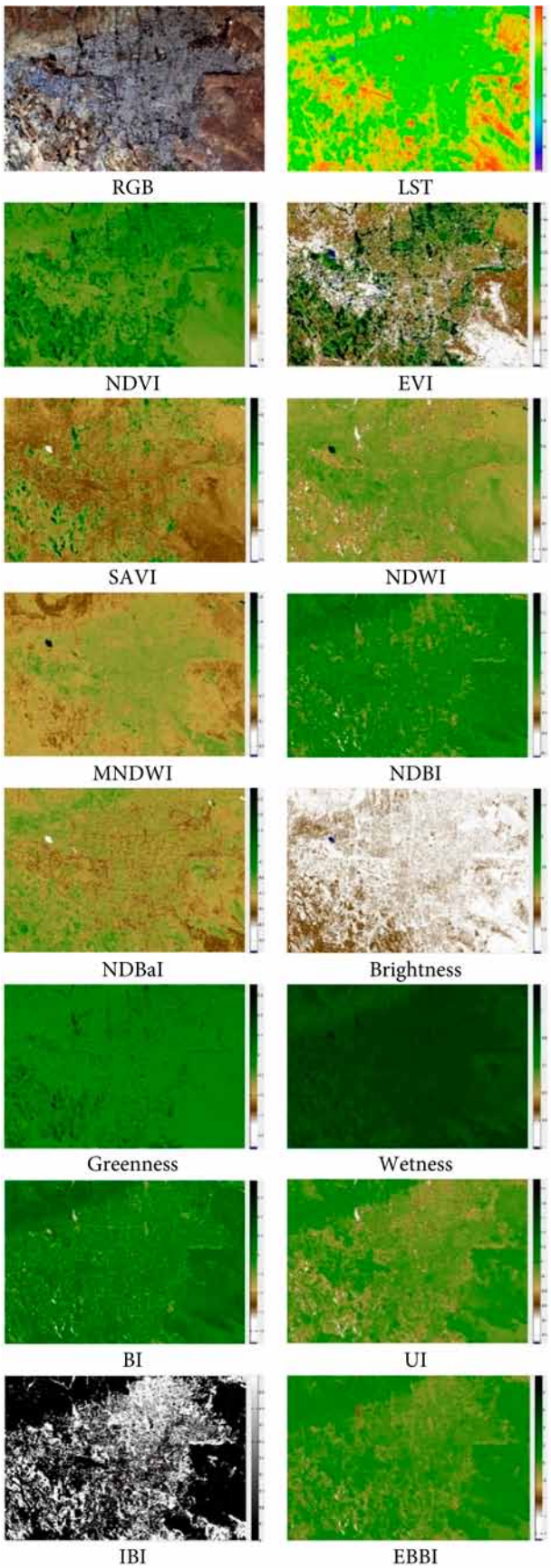

Fig. 3. Urban, vegetation and TCT indices for Tehran dataset

2.1. Scenario 1

In this scenario, we used and implemented LRM method by using linear combination of fourteen features including EVI, NDVI, NDWI, NDBI, NDBaI, MNDWI, BI, UI, IBI, EBBI, SAVI, Wetness, Greenness and Brightness as formulized in Eq. (11). Then, by incorporating training data, the estimation of unknown variable is possible. The output result of LRM method is presented in Table 4.

$$SHI = \beta_1 + \beta_2 EVI + \beta_3 NDVI + \beta_4 NDWI + \beta_5 NDBI + \beta_6 NDBaI + \beta_7 MNDWI + \beta_8 BI + \beta_9 UI + \beta_{10} IBI + \beta_{11} EBBI + \beta_{12} SAVI + \beta_{13} Wetness + \beta_{14} Greenness + \beta_{15} Brightness. \quad (11)$$

Table 4. The output result of LRM method using full fourteen input dataset (scenario 1)

	Estimate	SE	tStat	pValue
Intercept	−54.4486	13.8228	−3.939	0.00010
EVI	−8.2797	1.5004	−5.5183	0.00000
NDVI	−29.6874	3.9611	−7.4947	0.00000
NDWI	−15.9532	5.7056	−2.7961	0.00530
NDBI	−23.3553	11.5275	−2.0261	0.04310
NDBaI	−55.8113	0.9728	−57.3746	0.00000
MNDWI	−9.0744	6.219	−1.4591	0.14500
BI	0.7867	2.0305	0.3874	0.69860
UI	22.0356	4.4001	5.008	0.00000
IBI	−0.0522	0.0991	−0.5267	0.59860
EBBI	79.4409	13.1277	6.0514	0.00000
SAVI	54.911	5.6451	9.7272	0.00000
Wetness	56.0247	8.491	6.5981	0.00000
Greenness	0.6995	6.0152	0.1163	0.90750
Brightness	46.4579	3.0543	15.2106	0.00000

As shown in Table 4, estimated coefficient value for each feature (β_i) is illustrated in first column, the second column contains standard error (SE) of the estimation, the third column shows t statistic (tStat) for a test in which the coefficient is equal to zero and the last column contains p-value for the t statistic. We can also examine our model by incorporating estimated coefficient and using testing data. The result of LRM method is presented in Table 5.

Table 5. The performance of LRM method (scenario 1)

	MSE	RMSE	NRMSE	R^2
Training	0.5672	0.7531	0.3483	0.8762
Testing	0.5486	0.7407	0.3419	0.8785

From Table 5, it is obvious that, there is a good degree of consistency between SHI data and estimated parameters with testing RMSE around 0.74 °C and high compatibility with R^2 around 0.88.

On the other hand, a normal probability illustration of the residuals of a fitted linear model is shown in Figure 4a and added variable plot for whole model is shown in Figure 4b. Figure 4b illustrates the progressive effect on the reflex of specified terms by omitting the effects of all other terms. The slope of the estimated line is the coefficient of the linear compound of the determinate terms projected onto the best fitting direction. Also, from Figure 4b a horizontal line does not fit between the confidence boundaries, but referring to result extracted from Table 5, it is an acceptable result.

2.2. Scenario 2

As before, in this scenario, we used SVR method by using all feature set (fourteen features), as mentioned in flowchart (Fig. 1), next step is performed by SVR technique to relate extracted urban, vegetation and TCT indices to SHI data (Eq. 12).

$$SHI = f(NDVI, EVI, SAVI, NDWI, MNDWI, Brightness, Greenness, Wetness, NDBaI, NDBI, BI, UI, IBI, EBBI). \quad (12)$$

In this regards, as mentioned in previous section we adopted SVR technique. In SVR, the parameter C computes the trade-off between the flatness and the degree to which deviations larger than ε are tolerated in the optimization formulation. In this manner, if C is too large, then the objective is to minimize the empirical risk without regard to flatness part in the optimization formulation. The bigger the ε is, the fewer support vectors will be included. Therefore, more "flat" estimation is a consequence of bigger ε values. In fact, both C and ε values affect the flatness (model complexity). In this paper, C value is computed by (Eq. (13)) base on (Cherkassky, Ma 2004).

$$C = \max(Training\ data) - \min(Training\ data). \quad (13)$$

Also, a Gaussian radial basis function (RBF) kernel (Eq. (10)) is used; this function is widely used in remote sensing algorithms (Hasanlou *et al.* 2013). Before the regression estimating stage, simple normalizing must be applied to the training dataset. The main advantage of normalizing is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to reduce numerical complexity during the computation. The next step is the training procedure, during which some critical SVR parameters, ε, and in the RBF kernel, γ, must be specified.

A simple tool to check a grid of parameters is provided by cross-validation (CV) error (i.e. RMSE as fitness function) with 5-fold. Range of grid search method for estimating ε parameter is [0,5] and for γ RBF parameter is $[2^{-7},2^{7}]$. In this manner, Table 6 shows the optimum SVR parameters estimation for Tehran Landsat 8 imagery. It is obvious from Table 6 that the optimum SVR parameters for Tehran scene are $\in=0$, $\gamma=2$ and C = 22.40, which fulfils minimum RMSE.

By incorporating the optimum estimated parameters ($\in,\gamma$ and C) with minimum validation error (RMSE), and training dataset, the SVR model would be generated. Then, the performance of the selected final SVR model is computed for Tehran scene using testing data (Table 7).

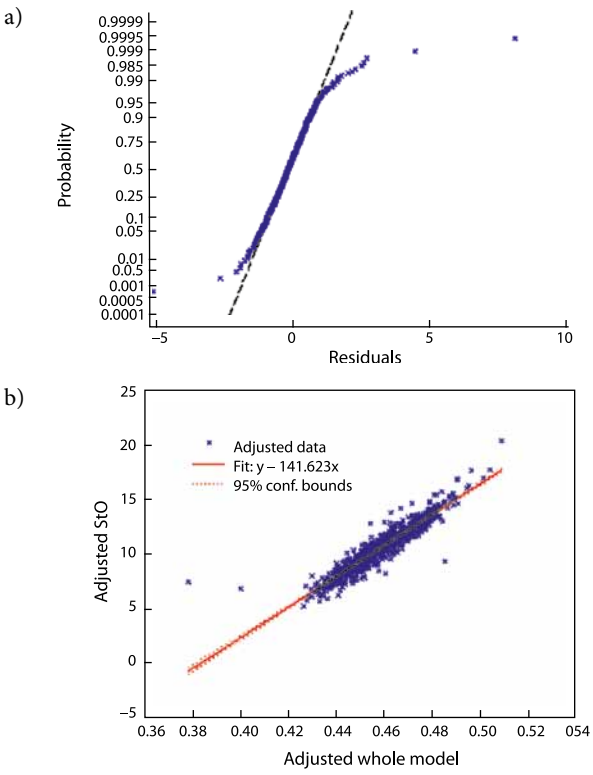

Fig. 4. a) Normal probability plot of residuals and b) Added variable plot for whole model

Table 6. Optimum SVR parameters estimation for Landsat 8 Tehran scene with C = 22.4013

	∈=0	1	2	3	4	5
$\gamma=2^{-7}$	8.9248	8.964	9.044	9.261	9.491	9.882
2^{-6}	8.1931	8.302	8.436	8.825	9.095	9.660
2^{-5}	7.2267	7.327	7.672	8.162	8.583	9.270
2^{-4}	5.8522	5.994	6.619	7.266	7.977	8.760
2^{-3}	3.9949	4.453	5.289	6.239	7.285	8.105
2^{-2}	2.372	2.939	3.974	5.242	6.397	7.409
2^{-1}	1.5473	2.110	3.174	4.467	5.729	6.894
2^{0}	1.4013	1.801	2.843	4.050	5.457	6.598
2^{1}	**1.3833**	1.683	2.552	3.770	5.179	6.578
2^{2}	1.5092	1.620	2.468	3.711	5.203	6.759
2^{3}	1.7264	1.825	2.685	3.973	5.360	7.122
2^{4}	1.9631	2.233	3.288	4.468	5.929	7.605
2^{5}	2.4885	2.957	4.228	5.561	6.974	8.220
2^{6}	3.554	4.100	5.487	6.811	8.160	9.216
2^{7}	5.1897	5.94	7.131	8.257	9.371	10.385

Table 7. The performance of final SVR model (scenario2)

	MSE	RMSE	NRMSE	R^2
Training	0.7507	0.8664	0.2424	0.9442
Testing	1.1155	1. 0562	0.3053	0.9100

As it is clear from Table 7, there is a high degree of consistency between incorporated information for each feature in kernel method and SHI data. For example, correlation coefficient in training data is $R^2 = 0.94$ and in testing data is $R^2 = 0.91$ and improvement of NRMSE comparing to scenario 2 (NRMSE = 0.3023).

2.3. Scenario 3

In this scenario, we applied GA-SVR method to optimize input space of SVR method. This means that by utilizing all feature set (fourteen features) and the GA procedure as feature selection optimizing tool, the suitable and appropriate features (including urban, vegetation and TCT indices) are selected as input space of SVR technique. Then reduced features would estimate SHI data. An individual's chromosome, i.e., the features present in an individual, was initialized in a random way and the parameters were set according to results of preliminary experiments. Table 8 presents all parameters set in GA. The samples were randomly chosen; however, the total number of samples has an important impact on the performance. The higher the number of samples are; the more time would be consumed calculating the fitness for each individual. In order to ensure high reliability of results, 10 runs of GA for each dataset were performed. Then those features that appeared more frequently were selected.

Table 8. GA's Parameters

Parameters	Value
Population size	20
Crossover rate	0.7
Elitism Ration	1
Mutation Ratio	0.05
Crossover Method	Two point
Max Iterations	50
Elite count	1

Since we have decided to let GA find the optimal features, it is also important to note that result of this scenario is optimal features (including urban, vegetation and TCT indices) with minimum RMSE validation error and maximum R^2. Figure 5 shows the best and the mean of fitness values (RMSE) in each generation for a single run of GA.

Five indices are selected as the optimal and best features among fourteen features, including, NDBI, NDBaI,

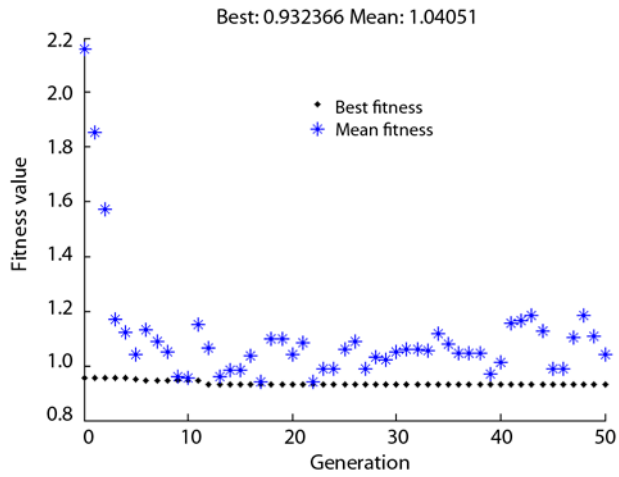

Fig. 5. Fitness values achieved by the proposed GA-SVR (scenario 3)

BI, Greenness and Brightness indices and the rest of the features are omitted from SVR input dataset (Eq. (14)).

$$SHI = f\left(NDBI, NDBaI, BI, Greenness, Brightness\right) \quad (14)$$

By incorporating these selected features, procedure of training and testing in SVR technique is commenced. Result and performance of this scenario (GA-SVR) is computed for Tehran Landsat 8 scene (Table 9). As it is clear from Table 9, the minimum NRMSE (0.26) has been achieved by this scenario. NRMSE facilitates the comparison between datasets or models with different scales and it is often expressed as a percentage, where lower values indicate less residual variance. Also, in this scenario, R-squared coefficient for testing data is $R^2 = 0.93$, calculated by using GA and SHI data, revealing high compatibility between selected features.

Table 9. The performance of GA-SVR model (scenario 3)

	MSE	RMSE	NRMSE	R^2
Training	0.8852	0.9409	0.2633	0.9338
Testing	0.8693	0.9324	0.2695	0.9315

2.4. Discussion

To sum up, utilizing three different scenarios with various regression methods would enable us to evaluate our estimation procedures better. Further, adopting NRMSE criteria for comparing purpose, enables the evaluation of training and testing result extracted from three scenarios. In this regards, scenario 3 has represented high performance in relating the input space features (input indices) to SHI data. Also, incorporating genetic FS method has some advantages comparing to previous scenarios, a) reduced dimensionality of input space increasing reliability and reducing computation complexity; and b) revealed the best and proper indices that used to connect

input space and SHI data. Using genetically FS enables us to monitor more affective indices that influence estimating SHI data (i.e. NDBI, NDBaI, BI, Greenness and Brightness indices).

Conclusions

All range of Landsat 8 spectral bands, particularly thermal bands, have been used to estimate SHI of Tehran city. In this study, urban indices including NDBaI, NDBI, BI, UI, IBI and EBBI have been calculated using recent urban parameters and factors. Further, to investigate vegetation factors better, more common vegetation and water indices including NDVI, EVI, SAVI, NDWI, MNDWI and TCT information including Brightness, Greenness and Wetness have been used. By utilizing these information and indices, the modeling and monitoring process of SHI is more practical. Also as a part of this study, three scenarios were implemented to compare the performance of each scenario. In scenario one, all calculated features/indices (full features) were used as input space of linear regression model. Result of scenario one is illustrated in Table 5. Then, in scenario two, same as previous scenario, full features were used as input space but with kernel base method (i.e. support vector regression). This scenario is more complicated than scenario one but it can handle high dimensional data and has better performance result (Table 7). Finally, in scenario three, we used supervised feature selection procedures (genetic algorithm) to select proper and affective features (indices). Estimated result (Table 9) revealed that scenario three has more reliable performance using NRMSE criteria comparing to two others (NRMSE = 0.2695 for scenario one, NRMSE = 0.3053 for scenario two and NRMSE = 0.3419 for scenario three). Also, incorporating genetic FS in these three scenarios indicated that, to estimate SHI data using Landsat 8 images, it is better to use more affective and optimum indices like NDBI, NDBaI, BI, Greenness and Brightness indices. This study would be completed by incorporating supervised feature extraction (FE) method to select suitable transform features from urban and vegetation information.

Acknowledgements

The authors would like to thank the Islamic Azad University's (South Tehran Branch) research deputy office for supporting this work. Also, we would like to thank the Earth Resources Observation and Science (EROS) – U.S. Geological Survey and the Land Remote Sensing Program of the U.S. Geological Surveying collaboration with NASA for generously availing the MODISLST and Landsat 8 data.

References

Actionbioscience. 2015. *Urban heat islands: hotter cities* [online], [cited 20 May 2015]. Available from Internet: http://www.actionbioscience.org/environment/voogt.html

As-syakur, A. R.; Adnyana, I. W. S.; Arthana, I. W.; Nuarsa, I. W. 2012. Enhanced built-up and bareness index (EBBI) for mapping built-up and bare land in an urban area, *Remote Sensing* 4(10): 2957–2970. https://doi.org/10.3390/rs4102957

Baig, M. H. A.; Zhang, L.; Shuai, T.; Tong, Q. 2014. Derivation of a tasselled cap transformation based on Landsat 8 at-satellite reflectance, *Remote Sensing Letters* 5(5): 423–431. https://doi.org/10.1080/2150704X.2014.915434

Chen, X.-L.; Zhao, H.-M.; Li, P.-X.; Yin, Z.-Y. 2006. Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes, *Remote Sensing of Environment* 104(2): 133–146. https://doi.org/10.1016/j.rse.2005.11.016

Cherkassky, V.; Ma, Y. 2004. Practical selection of SVM parameters and noise estimation for SVM regression, *Neural Networks* 17(1): 113–126. https://doi.org/10.1016/S0893-6080(03)00169-2

Dihkan, M.; Karsli, F.; Guneroglu, A.; Guneroglu, N. 2015. Evaluation of surface urban heat island (SUHI) effect on coastal zone: the case of Istanbul megacity, *Ocean & Coastal Management* 118(Part B): 309–316.

Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A.; Vapnik, V. 1997. Support vector regression machines, *Advances in Neural Information Processing Systems* 9: 155–161.

Durbha, S. S.; King, R. L.; Younan, N. H. 2007. Support vector machines regression for retrieval of leaf area index from multi angle imaging spectro radiometer, *Remote Sensing of Environment* 107(1–2): 348–361. https://doi.org/10.1016/j.rse.2006.09.031

Fabrizi, R.; Bonafoni, S.; Biondi, R. 2010. Satellite and ground-based sensors for the urban heat island analysis in the city of Rome, *Remote Sensing* 2(5): 1400–1415. https://doi.org/10.3390/rs2051400

Gao, B. 1996. NDWI–A normalized difference water index for remote sensing of vegetation liquid water from space, *Remote Sensing of Environment* 58(3): 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3

Goldberg, D. E.; Holland, J. H. 1988. Genetic algorithms and machine learning, *Machine learning* 3(2): 95–99. https://doi.org/10.1023/A:1022602019183

Hasanlou, M.; Mostofi, N. 2015. *Investigating urban heat island effects and relation between various land cover indices in Tehran City using Landsat 8 Imagery*. MDPI, f004.

Hasanlou, M.; Samadzadegan, F.; Homayouni, S. 2013. SVM-based hyperspectral image classification using intrinsic dimension, *Arabian Journal of Geosciences* 8(1): 477–487. https://doi.org/10.1007/s12517-013-1141-9

Huete, A. R. 1988. A soil-adjusted vegetation index (SAVI), *Remote Sensing of Environment* 25(3): 295–309. http://dx.doi.org/10.1016/0034-4257(88)90106-X

Imhoff, M. L.; Zhang, P.; Wolfe, R. E.; Bounoua, L. 2010. Remote sensing of the urban heat island effect across biomes in the continental USA, *Remote Sensing of Environment* 114(3): 504–513. https://doi.org/10.1016/j.rse.2009.10.008

Jimenez-Munoz, J. C.; Sobrino, J. A.; Skokovic, D.; Mattar, C.; Cristobal, J. 2014. land surface temperature retrieval methods from Landsat-8 thermal infrared sensor data, *IEEE Geoscience and Remote Sensing Letters* 11(10): 1840–1843. https://doi.org/10.1109/LGRS.2014.2312032

Kawamura, M.; Jayamana, S.; Tsujiko, Y. 1996. Relation between social and environmental conditions in Colombo Sri Lanka and the urban index estimated by satellite remote sensing data. The International Archives of the Photogrammetry, *Remote Sensing and Spatial Information Sciences* 31: 321–326.

Kriegler, F. J.; Malila, W. A.; Nalepka, R. F.; Richardson, W. 1969. Preprocessing transformations and their effects on multispectral recognition, in *Remote Sensing of Environment*, VI. 97.

Kutner, M.; Nachtsheim, C.; Neter, J.; Li, W. 2004. *Applied linear statistical models*. 5th ed. Boston: McGraw-Hill/Irwin.

Landsat 8. 2016. *Landsat 8* [online], [cited 21 May 2016]. Available from Internet: http://landsat.usgs.gov/landsat8.php

Liu, K.; Su, H.; Zhang, L.; Yang, H.; Zhang, R.; Li, X. 2015. Analysis of the urban heat island effect in Shijiazhuang, China using satellite and airborne data, *Remote Sensing* 7(4): 4804–4833. https://doi.org/10.3390/rs70404804

MOD11A2. 2015. *NASA land data products and services* [online], [cited 22 May 2015]. Available from Internet: https://lpdaac.usgs.gov/products/modis_products_table/mod11a2

Moser, G.; Serpico, S. B. 2009. Automatic parameter optimization for support vector regression for land and sea surface temperature estimation from remote sensing data, *IEEE Transactions on Geoscience and Remote Sensing* 47(3): 909–921. https://doi.org/10.1109/TGRS.2008.2005993

Odindi, J. O.; Bangamwabo, V.; Mutanga, O. 2015. Assessing the value of urban green spaces in mitigating multi-seasonal urban heat using MODIS land surface temperature (LST) and Landsat 8 data, *International Journal of Environmental Research* 9(1): 9–18.

Ogashawara, I.; Bastos, V. da S. B. 2012. A quantitative approach for analyzing the relationship between urban heat islands and land cover, *Remote Sensing* 4(11): 3596–3618. https://doi.org/10.3390/rs4113596

Smola, A. J.; Schölkopf, B. 2004. A tutorial on support vector regression, *Statistics and Computing* 14(3): 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Streutker, D. R. 2002. A remote sensing study of the urban heat island of Houston, Texas, *International Journal of Remote Sensing* 23(13): 2595–2608. https://doi.org/10.1080/01431160110115023

UN DESA. 2015. *World's population increasingly urban with more than half living in urban areas* [online], [cited 20 May 2015]. Available from Internet: http://www.un.org/en/development/desa/news/population/world-urbanization-prospects-2014.html

USGS. 2015. *Using the USGS Landsat 8 product* [online], [cited 20 May 2015]. Available from Internet: http://landsat.usgs.gov/Landsat8_Using_Product.php

Voogt, J. A.; Oke, T. R. 2003. Thermal remote sensing of urban climates, *Remote Sensing of Environment* 86(3): 370–384. https://doi.org/10.1016/S0034-4257(03)00079-8

Xian, G.; Crane, M. 2006. An analysis of urban thermal characteristics and associated land cover in Tampa Bay and Las Vegas using Landsat satellite data, *Remote Sensing of Environment* 104(2): 147–156. https://doi.org/10.1016/j.rse.2005.09.023

Xiong, Y.; Huang, S.; Chen, F.; Ye, H.; Wang, C.; Zhu, C. 2012. The impacts of rapid urbanization on the thermal environment: a remote sensing study of Guangzhou, South China, *Remote Sensing* 4(7): 2033–2056. https://doi.org/10.3390/rs4072033

Xu, H. 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery, *International Journal of Remote Sensing* 27(14): 3025–3033. https://doi.org/10.1080/01431160600589179

Xu, H. 2008. A new index for delineating built-up land features in satellite imagery, *International Journal of Remote Sensing* 29(14): 4269–4276. https://doi.org/10.1080/01431160802039957

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. https://doi.org/10.1080/01431160304987

Zhao, H.; Chen, X. 2005. Use of normalized difference bareness index in quickly mapping bare areas from TM/ETM+, in *Geoscience and Remote Sensing Symposium, 2005. IGARSS '05*, 29 July 2005, Seoul, South Corea. IEEE International, 1666–1668.

Nikrouz MOSTOFI. He received the BSc degree in Civil Eng.-Surveying & Geomatics Eng. from industrial K. N. Toosi University, Tehran, Iran in 2001, and the MSc degree in photogrammetry engineering from the Tehran University, Tehran, Iran in 2005. Since 2014 he has been PhD in Remote Sensing and Geospatial Information System in Islamic Azad University, Sciences and Researchs Branch. Also, he is a member of Islamic Azad University, South Tehran Branch, and Department of Surveying Engineering, Tehran, Iran. His current research interests are using urban heat island monitoring, close range photogrammetry and machine learning algorithms in photogrammetry and remote sensing.

Mahdi HASANLOU. He received the BSc degree in Surveying and Geomatics Eng. from the University of Tehran, Tehran, Iran, in 2003, the M.Sc. degree in Remote Sensing from University of Tehran, Tehran, Iran, in 2006, and the PhD degree in Remote Sensing from University of Tehran, Tehran, Iran, in 2013. Since 2013, he has been as an assistant professor in the School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Iran, His research activities are mainly focused on Thermal, optical and SAR remote sensing for urban and agro-environmental applications.