Yoo *J Big Data (2018) 5:2*
https://doi.org/10.1186/s40537-018-0113-z

Journal of Big Data

RESEARCH

Open Access

# Investigating important urban characteristics in the formation of urban heat islands: a machine learning approach

Sanglim Yoo*

*Correspondence:
syoo@bsu.edu
Department of Urban Planning, College of Architecture and Planning, Ball State University, Muncie, IN 47306, USA

**Abstract**

Despite the urban heat islands phenomenon has long been recognized as a major urban environmental problem, it was not until recently that this urban phenomenon gained attention from the discipline of urban planning. To integrate the findings of the urban heat islands research into the planning practice, the relationship between land surface temperatures and urban physical and socioeconomic characteristics should be addressed at the planning relevant spatial scale, a land parcel. Using a parcel as a unit of analysis, this study proposed to use a machine learning approach to identify important variables in the formation of urban heat islands in Indianapolis, Indiana. Applying random forest method to planning zones, this study identified planning zone specific urban physical and socioeconomic characteristics that are important for the interpretation of urban heat islands phenomenon of Indianapolis, Indiana. The main contribution of this study is twofold: to integrate urban physical and socioeconomic characteristics into a land parcel for the better interpretation of the result of urban heat islands study into planning practice and to apply machine learning approach to identify highly determinant variables in the formation of urban heat islands.

**Keywords:** Urban heat island effect, Biophysical vulnerability, Socioeconomic vulnerability, Machine learning, Random forest, Variable selection

## Introduction

The warming trend of US cities becomes significant since the late 1970s and its rate and magnitude of this trend severed during the late 1990s [1–3]. Land use changes due to urbanization can modify the energy balance in cities, and in turn, this affects the urban thermal environment, resulting in the urban heat islands (UHIs) phenomenon, meaning urban areas have higher air and surface temperature than their rural surroundings [4–6]. Long recognized in many disciplines of physical science, such as geography as an important climatological phenomenon, however, the UHIs has only recently gained serious attention in the field of urban planning [7]. As UHIs occur as a result of land cover transformation, mainly replacement of natural vegetation and natural land cover by impervious surfaced associated with urban land uses, key questions for the discipline of urban planning include what are the most contributing urban physical characteristics to the formation of UHIs, and their consequences to urban resident population, and consequently how to mitigate it.

Despite its commonly acknowledged importance as a major urban environmental issue, above stated questions were primarily investigated and addressed by the discipline of geography in the context of remote sensing and the assessment of social vulnerability. Moreover, the UHIs phenomenon has not been properly addressed by urban planners even as they have tried to become more environmentally conscious in both their thinking and practice [8]. The Indianapolis Metropolitan Statistical Area (MSA), which is the main subject of this study, is one of those US cities where UHIs have grown in largest numbers in recent years, but both state and local government have not yet to provide any concrete and effective solutions [2].

This lack of attention and response to UHIs and its environmental and social causes and consequences can be at least partly attributed to a scarcity of good empirical research which provides urban planners the kind of data and ideas that they need and able to utilize in the planning process. This does not necessarily mean that no meaningful UHIs research exists, however this research found three major shortcomings in the existing UHIs research. First, this study found that the most of the studies that addressed the relationship between UHIs and urban physical characteristics and social vulnerability to the heat-related weather event did not conducted on planning relevant urban scale. Many of them heavily on remotely sensed land surface characteristics of urban areas, such as land cover type and vegetation index which mainly measured at the microscale [9–11] and others that investigated social vulnerability to extreme heat-related disasters conducted on macro scale such as county, metropolitan statistical area, census tract level [2, 12–14]. Second, there is very few research that focused on the integration of physical factors that affect the formation of UHIs and variables that determine the socioeconomic vulnerability of the urban area. Third, there are very few empirical research that quantified the magnitude of urban physical and socioeconomic characteristics to the formation of UHIs. This research tried to fill this gap.

To address above stated issues, this study integrated urban physical and socioeconomic attributes into a land parcel which is the unit of planning practice, then investigated the relative importance of each attribute in the formation of UHIs using a machine learning approach. This study proposes to (1) integrate remotely sensed land surface characteristics, built area characteristics, and socioeconomic demographic characteristics of the urban area into planning relevant urban scale (2) apply the random forest (RF) method, a machine learning approach, as well as principal component analysis (PCA), a non-parametric data reduction technique widely used by social vulnerability studies, to identify and select important variables to the formation of UHIs, and (3) explain what makes the urban area hotter, how these factors are differentiated depends on the planning zones, and who are the most vulnerable to the heat-related weather event in each planning zones.

## Background

The relationships between remotely sensed land surface temperature and physical land surface characteristics have been intensively investigated in many disciplines of physical sciences, such as geography, as an important climatological phenomenon. Together with this trend, many social science studies have focused on addressing socioeconomically vulnerable groups of population and regions to the heat-related weather event. But in the field of urban planning, the UHI effect has only recently gained serious attention from planners [7]. This study strongly believed that it is largely due to the lack of a methodology that enables planners to integrate the findings from a grid level or from the much larger administrative boundaries level into a planning relevant scale. To guide the discussion of this paper to the topic of parcel based investigation of highly determinant physical land surface characteristics and socioeconomic characteristics contributing to the formation of UHIs, this section of the paper will discuss below: first, what biophysical and socioeconomic attributes have been addressed in various studies, second, in what scale these attributes were measured and integrated, and finally, how to address the importance of each attribute in the formation of UHIs.

### Variables for urban heat islands studies

As UHIs occur as a result of the land cover transformation, primarily replacement of natural vegetation and natural land cover by impervious surfaced associated with urban land uses, previous empirical UHIs research has mainly focused on the causal relationship between remotely sensed land surface temperature and remotely sensed physical land surface characteristics of urban areas such as elevation and slope [15, 16], land cover types [5, 10, 17, 18], percent impervious surface [5, 15, 19], percent tree canopy cover [5, 7, 19], vegetation abundance indices [16, 18–20].

There is another stream of UHIs research that investigated social vulnerability to extreme heat-related weather events. Social vulnerability to environmental hazard means the relative potential for physical harm and social disruption to subpopulations of societies and their larger subsystems based on socioeconomic status, age, gender, race and ethnicity, family structure, residential location and other demographic variables [21]. In the studies investigated the relationship between extreme heat relate weather events and socioeconomic status of the area, socioeconomic variables including total population [12], population density [15, 16, 19], total housing units [12], gender [12], race and ethnicity [12, 14, 16, 22, 23], number of older and younger population [12, 14, 23, 24], education [14, 23], median house value [12, 16, 25], median household income [13, 22], poverty [14, 22–24], housing tenure [22], number of mobile homes [12] were most widely used and found significant.

### Linking urban heat islands phenomenon with urban physical and socioeconomic attributes

Starting from the above-stated discussion of what biophysical and socioeconomic variables are to be included in the analysis, the unit of analysis, which means in what geographical unit all of the variables to be integrated is another important issue to be discussed. As Stone and Norman [7] addressed earlier in their study, in order to associate aforementioned remotely sensed urban surface attributes and socioeconomic characteristics with the planning and planning decision-making, it is essential that these attributes should integrate into a planning-relevant scale.

However, there is no commonly agreed upon guidelines for aggregating data into the certain unit of analysis or theory that provides any arguments in favor of certain data integration method. So many empirical studies have selected unit of analysis using cumulative results from prior empirical studies. There are studies that have adopted a unit of analysis consistent with the resolution of a satellite image, a grid cell [5, 18, 19, 25], and others used a unit that of socioeconomic variables were measured such as County [13], Census tract [16, 17], and Census block group [14, 26]. To the best of my knowledge, Stone and Norman [7] is the only study that used a parcel as a unit of analysis.

In order to associate aforementioned research results with the practice of urban planning and policy, it is essential to address the relationship between land surface temperatures in planning relevant urban scale. If urban planners may be able to mitigate the formation of UHIs through the modification of specific zoning and various regulation, urban physical characteristics highly correlated with higher land surface temperature in planning relevant urban scale should be identified first. As socioeconomically vulnerable population should be identified together with physical characteristics of the urban area, socioeconomic characteristics highly associated with higher land surface temperature also need to be addressed in that relevant urban scale.

However, it is methodologically not so simple because remotely sensed surface temperatures are measured in micro scale but resulting socioeconomic vulnerabilities are presented in much larger scale. Also, to effectively integrate the findings of the UHIs research, the result should be addressed in planning relevant scale. However, neither micro scales such as a pixel nor macro scales such as a Census block group and county cannot be directly used by urban planners to reduce the magnitude of UHI effects. This study tried to overcome this discrepancy in scale issue in following two ways: first, divide dataset by planning zones, as different planning zones represent different human activities so will need different UHIs mitigation strategies, second, integrate urban biophysical characteristics which measured in microscale and socioeconomic characteristics which measured in the Census block group level into land parcels for parcels are the spatial unit of urban planning and design.

### Testing variable importance

To investigate what are the most important determinant factors in the formation of the UHIs, who are the most severely affected by the consequence of the UHIs phenomenon, and where the phenomenon is the most intense, this study applied a machine learning approach. As there is no theory that explains the aforementioned relationship, thus the machine learning technique, which allows the computer to learn without being programmed and to identify patterns in data, was used to find highly determinant variables to the formation of UHIs in Marion County, Indiana. To quantify the causal relationship between the land surface temperature and explanatory variables, previous empirical studies have been used regression approaches such as simple regression and spatial regression approaches [19, 20, 26] and correlation analysis [18]. The PCA method, a mathematical procedure that transforms a number of correlated variables into a smaller number of uncorrelated variables called principal component, has been widely used in many of empirical UHIs and social vulnerability studies [12, 13, 19] to reduce the data dimension.

To the best of my knowledge, works of Hart and Sailor [5] and Rhee, Park, and Lu [20] are one of the few research that applied machine learning technique for the identification of highly determinant variable to the formation of UHIs. Hart and Sailor [5] applied Cubist method, a rule-based regression tree approach, to determine the most important variables affecting the UHIs intensity in Portland, Oregon. Rhee and others [20] applied the RF method to investigate the relationship between land cover pattern and surface temperature in Denver, Colorado to identify the relative importance of variables. Because there is no commonly accepted theory to guide modeling the relationship between urban attributes and UHIs, this study strongly believed the machine learning approach that allows the computers to learn from data without programming can be used as a powerful tool to identify the importance of urban attributes to the formation of UHIs. This study applies RF method to identify the relative importance of urban biophysical and socioeconomic variables to the formation of UHIs at the parcel level.

## Methodology

### Study area

Marion County, Indiana is the largest county in the state of Indiana and the 55th largest county in the US by its population [27]. The county seat is Indianapolis, the state capital and the largest city in Indiana. Indianapolis is 15th largest US city by its population, and it is one of those US cities where UHIs have grown in largest numbers in recent years [2, 27] (Fig. 1).

### Landsat 7 ETM+ processing for land surface temperature computation

A Landsat 7 Enhanced Thematic Mapper Plus (ETM+) image of Marion County, Indiana, path 21 row 32, which was acquired on 12 July 2013 was used for this study. During summer months between the year 2011 and year 2016, an image dated 12 July 2013 was chosen because of the higher atmospheric temperature before and after the image

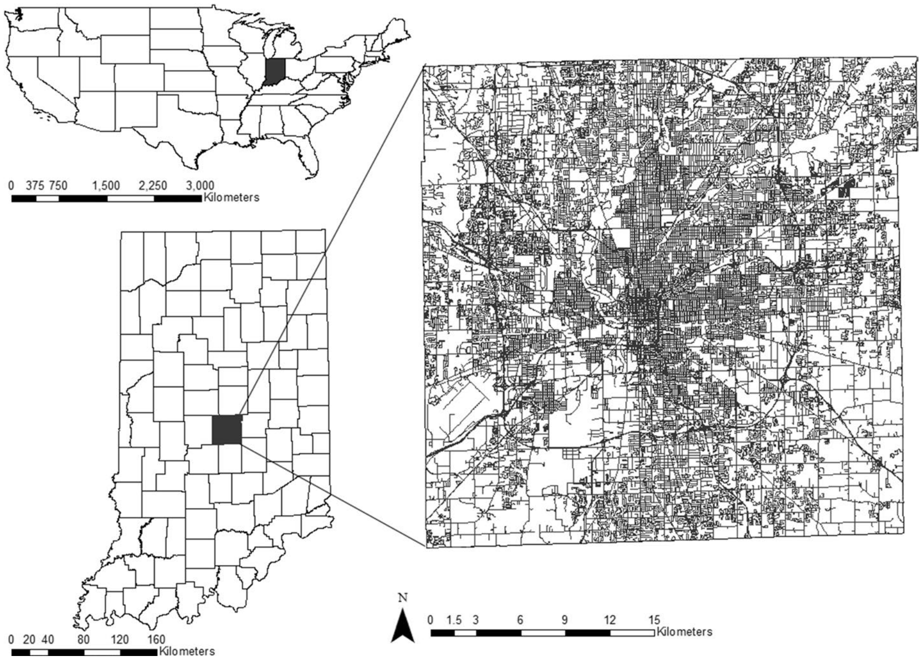

**Fig. 1** Study area map showing the location of Marion County, Indiana

acquisition date, and also due to the lower cloud cover. Image used in this study was taken at EST 4:18 p.m. According to US Environmental Protection Agency (US EPA) and US National Oceanic and Atmospheric Administration (US NOAA), year 2006 through 2015 was the warmest decade on record since thermometer-based observations began from 1850, and year 2013 was one of the hottest years on record [3, 28]. The image was rectified to a common Universal Transverse Mercator coordinate system based on 1:24,000 scale topographic maps and was resampled to a pixel size of 30 m by 30 m using a nearest neighbor resampling algorithm. The root mean square error (RMSE) achieved during the rectification was less than 0.5 pixel.

The thermal infrared (TIR) band of an image was converted to surface temperature following the process suggested by Weng and others [18]. The digital number (DN) of Landsat 7 ETM+ TIR band was first converted into spectral radiance using Eq. (1), and then converted to blackbody temperature ($T_B$) under the assumption of uniform emissivity using Eq. (2) below:

$$L_\lambda = 0.0370588 \times DN + 3.2 \tag{1}$$

$$T_B = \frac{K_2}{\ln\left(\frac{K_1}{L_\lambda} + 1\right)} \tag{2}$$

where $L_\lambda$ is spectral radiance, $T_B$ is effective at-satellite temperature, or blackbody temperature in Kelvin, $K_1$ and $K_2$ are pre-launched calibration constant [28].

For Landsat 7 ETM+, $K_1 = 666.09$ mW cm$^{-2}$ sr$^{-1}$ μm$^{-1}$, and $K_2 = 1282.71$ K were used [28]. The emissivity corrected land surface temperature ($T_s$) were finally computed as follows:

$$T_s = \frac{T_B}{1 + \left(\lambda \times T_B/\left(1.438 \times 10^{-2}\right)\right) \ln \varepsilon} \tag{3}$$

where λ is the wavelength of emitted radiance, and here $\lambda = 11.5$ μm was used, and ε is spectral emissivity [29]. ENVI 5.3 software was used for the land surface temperature calculation. Figure 2 shows the spatial distribution of summertime land surface temperature of the study area.

### Acquisition and processing of urban physical and socioeconomic variables

The formation of UHIs is the result of a complicated combination of land cover changes, structural characteristics of the built area, meteorological conditions, landscape structure, and socioeconomic conditions of the urban area. The consequences of the UHIs phenomenon differentiate among various socioeconomic groups. Based on the in-depth literature review and the availability of data for the study area, this study selected following urban physical characteristics and socioeconomic characteristics to investigate the causal relationship with land surface temperatures (Table 1).

All of the remotely sensed data including land surface temperature and NDVI which derived from Landsat 7 ETM + and NLCD 2011 data including percent developed imperviousness and percent tree cover were measured in 30-m by 30-m resolution. Data explaining urban physical structure including average and maximum building height,

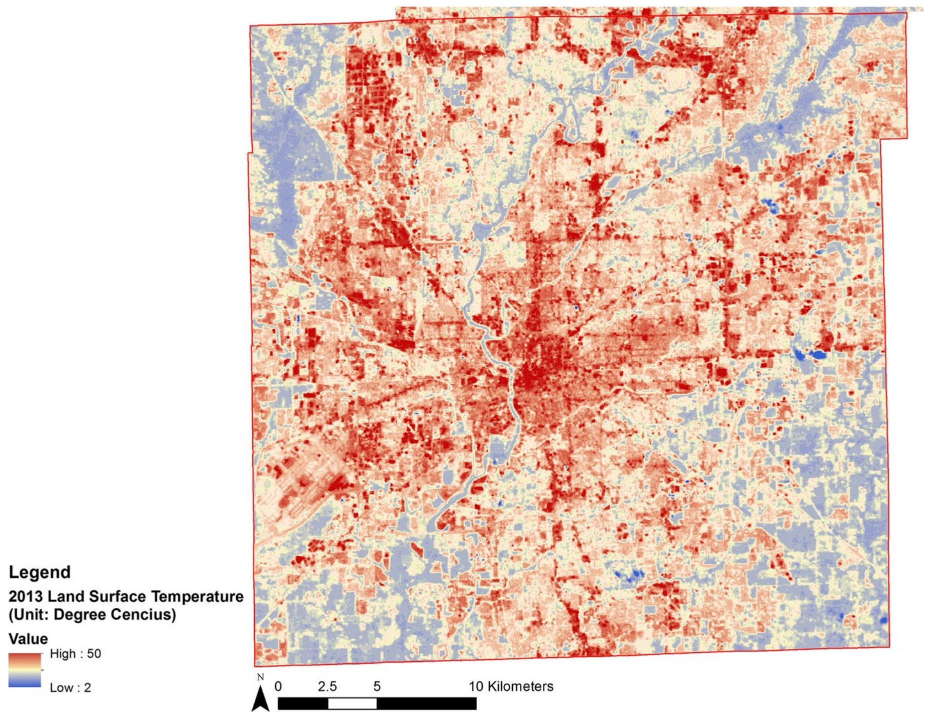

**Fig. 2** Spatial distribution of summertime land surface temperature of study area, July 2013

and building footprint were calculated for each building using LiDAR image in 1-m by 1-m resolution using ENVI 5.3 software. All of the socioeconomic variables were collected at the level of Census block group from the year 2013 American Community Survey (ACS) to be consistent with the Landsat 7 ETM+ image which also acquired at year 2013.

According to the zoning ordinance of Marion County, Indiana, there are72 types of zoning districts, which were reclassified by permitted land uses and human activities. This study reclassified original zoning districts into eight categories by grouping them by the similarity in land uses and resulting human activities, including Central Business District (CBD), commercial, industrial urban, industrial suburban, urban dwelling, suburban dwelling, agricultural dwelling, and mixed-use district. Zoning districts such as airport, speedways, cemetery, sanitary landfill, power substation were excluded from the analysis, even though the imperviousness and surface temperatures of these zones were relatively high, these land uses have little significance in human settlement. Description of eight planning zones considered in this study is summarized in Table 2.

The Normalized Difference Vegetation Index (NDVI) is a function of the visible and near-infrared reflectance from plant canopy, the reflectance of the same spectra from the atmospheric reflectance [18]. As this index provides an estimate of the abundance of actively photosynthesizing vegetation, thus it can be used to infer general vegetation condition [23, 26, 30]. NDVI is computed as follows:

$$\mathrm{NDVI} = (\mathrm{TM4} - \mathrm{TM3})/(\mathrm{TM4} + \mathrm{TM3}) \quad (4)$$

**Table 1 Summary of physical and socioeconomic variables**

| Variable | Definition | Unit | Measurement | Source |
|---|---|---|---|---|
| *Dependent variable: land surface temperature* | | | | |
| TEMP | Land surface temperature of 12 July 2013 | °C | Pixel (30 m × 30 m) | Landsat 7 ETM+ |
| *Independent variable: physical variables* | | | | |
| Zone | Planning zone, reclassified as below<br>Central Business District (CBD)<br>Mixed use<br>Commercial<br>Industrial urban<br>Industrial suburban<br>Urban dwelling<br>Suburban dwelling<br>Agricultural dwelling | NA | Planning zone | Indianapolis and Marion County, Department of Metropolitan Development |
| Area | Total area of a parcel | $m^2$ | Parcel | Indy GIS |
| NDVI | Normalized Difference Vegetation Index | NA | Pixel (30 m × 30 m) | Landsat 7 ETM+ |
| TREE | Percent tree canopy | % | Pixel (30 m × 30 m) | NLCD 2011, USGS |
| Imperv | Percent developed imperviousness | % | Pixel (30 m × 30 m) | NLCD 2011, USGS |
| Height_Avr | Average building height | m | Building | LiDAR, 2011 |
| Height_Max | Maximum building height | m | Building | LiDAR, 2011 |
| Bldg_Ftpr | Total building footprint | $m^2$ | Building | Indianapolis and Marion County, Department of Metropolitan Development |
| *Independent variable: socioeconomic variables* | | | | |
| POP_DEN | Population density of a Census block group | $\#/m^2$ | CBG | ACS 2013 |
| POP_UR_18 | Total number of population under age 18 | # | CBG | ACS 2013 |
| POP_OV_65 | Total number of population over age 65 | # | CBG | ACS 2013 |
| Female | Total number of female population | # | CBG | ACS 2013 |
| NON_WT | Total number of non-white population | # | CBG | ACS 2013 |
| TOT_HSUNT | Total number of household unit | # | CBG | ACS 2013 |
| RENT | Total number of rental unit | # | CBG | ACS 2013 |
| MedHsValue | Median house value | $ | CBG | ACS 2013 |
| MEDHHINC | Median household income | $ | CBG | ACS 2013 |
| Edu_Less | Total number of population have education less than 9th grade | # | CBG | ACS 2013 |
| Mobile | Total number of mobile home | # | CBG | ACS 2013 |

where TM3 and TM4 refer to the reflectivity of Landsat TM band 3 and band 4, with the wavelength of 0.63–0.69 and 0.76–0.90 μm respectively [19, 29]. The value varies from − 1 to + 1, and the greater the value, the better the vegetation condition. NDVI was calculated using ENVI 5.3 software. Area of parcels, building footprints, average building heights, and maximum building heights were calculated using ArcGIS 10.4.1 and ENVI 5.3 software. Percent tree canopy and percent developed imperviousness were derived from National Land Cover Database (NLCD) 2011 by US Geological Survey (USGS).

Subject of socioeconomic vulnerability to extreme heat-related weather event and heat-related death was intensively documented by previous vulnerability studies [12–14,

**Table 2 Summary of planning zones in Marion County, Indiana**

| Planning zone | Zoning district | Description |
|---|---|---|
| CBD | Central Business District one<br>Central Business District two<br>Central Business District three | The area containing the general downtown area<br>This planning zone represents the typical urban core of Indianapolis to be developed at very high density<br>Maximum lot coverage: 100% |
| Commercial | Office buffer district<br>Neighborhood commercial district<br>Community regional district<br>General commercial district<br>High-intensity commercial district<br>Special commercial district | This planning zone provides a freestanding area that office use, compatible office type uses such as medical and dental facilities, education service, public and semipublic uses, and an extensive range of retail sales and personal, professional, business services, regional size shopping center, and retail services and services functions whose operations are typically characterized by automobiles, outdoor display, sales of merchandise, major outdoor storage or display of sizable merchandise<br>The maximum height of buildings and structure: ranges from 38 to 65 ft. |
| Urban dwelling | Dwelling district 3–11<br>Planning unit development district (PUD) | This planning zone provides for low, medium, or high-density residential development including single family detached dwelling, single-family attached dwelling, two families attached dwelling, multifamily dwelling, triplex, fourplex, high-rise apartment, and mobile dwelling<br>The maximum height of the primary building: ranges from 35 ft. to unlimited (24+ stories)<br>Minimum open spaces: ranges from 70 to 8% |
| Suburban dwelling | Dwelling suburban district<br>Dwelling district one<br>Dwelling district two | This planning zone is intended for suburban areas to permit low to minimum density residential development<br>Minimum lot area, single family: 15,000 ft.$^2$, two family: 20,000 ft.$^2$<br>Minimum open space: 75%<br>The maximum height of the primary building: 35 ft.<br>Minimum main floor area: 1200 ft.$^2$ |
| Agricultural dwelling | Dwelling agricultural district | This planning zone holds the agricultural lands of Marion County and provides for a variety of agricultural uses<br>A single-family dwelling is intended to be permitted as a part of such agricultural uses<br>Minimum lot area: 3 acres<br>Minimum open space: 85%<br>The maximum height of the primary building: 35 ft.<br>Minimum main floor area: 1200 ft.$^2$ |
| Industrial urban | Light industrial urban<br>Medium industrial urban<br>Heavy industrial urban<br>Restricted industrial urban | This planning zone permits an industrial park land use in urban area to prevent objectionable characteristics and possibly harmful activities from the lot lines<br>Light industrial area is used as a buffer zone between the more intense industrial district and protected district such as a dwelling, park, hospital, or school |
| Industrial suburban | Light industrial suburban<br>Medium industrial suburban<br>Heavy industrial suburban<br>Restricted industrial suburban | This planning zone permits an industrial park land use in suburban area to prevent objectionable characteristics and possibly harmful activities from the lot lines<br>Light industrial area is used as a buffer zone between the more intense industrial district and protected district such as a dwelling, park, hospital, or school |
| Mixed use | Mixed use one district<br>Mixed use two district<br>Mixed use three district<br>Mixed use four district | This planning zone includes the development of high-rise office uses and apartment intermix, a mix of residential uses, office, personal services, retail, and eating and drinking businesses, a compact, mixed-use village development comprised of moderate and high-density housing<br>The height of the building: ranges from 18 ft. to unlimited |

16, 31–34], and many of the previous studies commonly addressed characteristics such as population density, proportion of youth and elder population, female population, non-white population, total housing units, total rental units, total mobile homes, education level, and indices that can explain household economic statuses such as median household income and median house value. Based on the findings of previous empirical studies, this study used aforementioned socioeconomic variables available from American Community Survey 2013. These selected variables were collected at the level of Census block group, which represents the smallest geographic unit for which the desired information was available (Table 1).

### Dataset generation: linking land surface temperature with physical and socioeconomic characteristics into the land parcel

To generate the dataset, all of the metrics described in Table 1 were integrated into the land parcel. As a parcel is commonly acknowledged as an important spatial unit of contemporary urban planning and design [35], this study used a parcel as a unit of data integration and as a unit of analysis. Since the average parcel size of Marion County, Indiana was 1526.4 $m^2$, each parcel will have approximately 1.7 number of 30-m by 30-m pixels in average. In the case of large parcels, which may contain many pixels, it is possible to have spatially varying surface characteristics such as surface temperature, percent developed imperviousness and percent tree canopy within a parcel. It will be problematic because if a single value, such as percent impervious cover of a pixel, is allocated to a parcel, then spatially varying characteristics that computed in micro-scale will not be counted. In that chase, lots of valuable information measured in a grid level will be lost. To prevent this information loss, this study allowed duplication. Simply put, a parcel may be recognized as multiple observations depends on its size and also depends on the number of pixels observed within a parcel. Spatially varying urban surface characteristics were input into a parcel, resulting in multiple observation which shares macro-level socioeconomic characteristics but has different micro level physical characteristics. In Fig. 3, the example of a Census block group that contains residential, industrial, and commercial zones, multiple land parcels, building footprints, and raster grids is illustrated. In Marion County, there are total 335,489 parcels, and when allowing duplication, then the total number of observations were increased to 942,193, reflecting the number of pixels in the study area. This study divided total 942,193 observations into eight planning zones which share similar urban land surface characteristics. By dividing the original dataset into eight subsets, this study could increase its processing speed and efficiency.

### Statistical analysis: random forest

With the advancement of analytical techniques, big data empowers planners and planning decision-makers by helping them to better understand current situations and to predict future more precisely and accurately [36]. Machine learning is one of the proven analytic tools to harnessing the power of big data. It is one of the analytical approaches that uses computer algorithms to repeatedly learn from the data. As the analytical tool continues to learn from the data, the prediction accuracy of the model improves over time [36]. Because the main purpose of this study is in identifying highly determinant

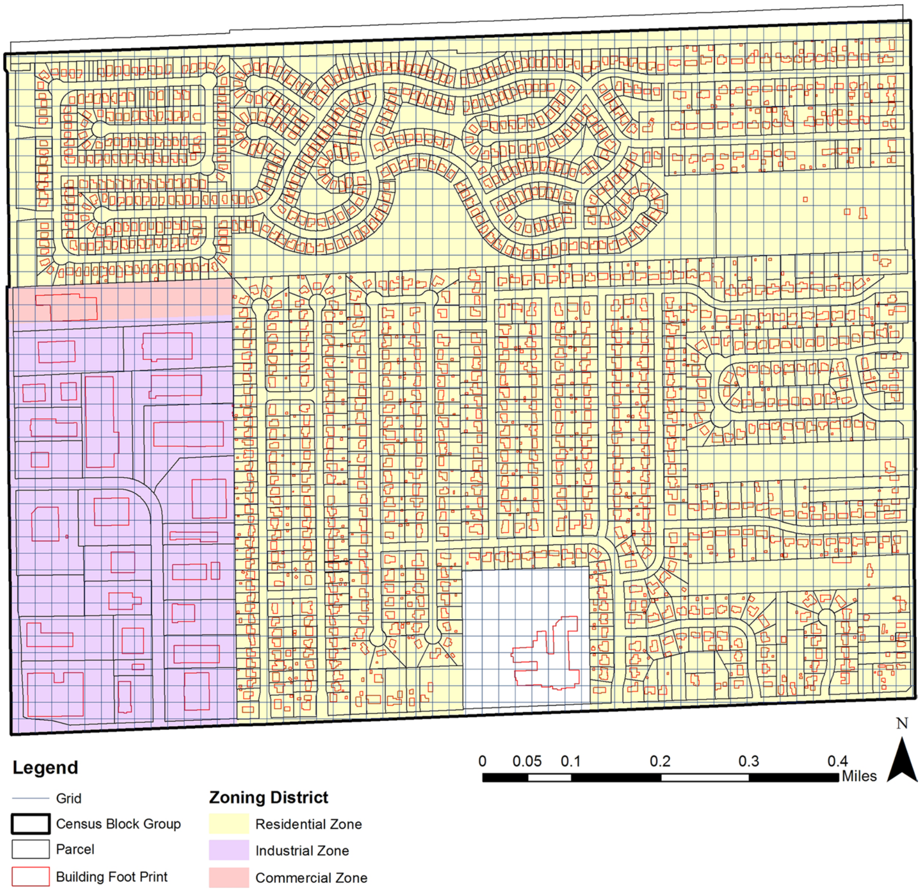

**Fig. 3** Example of a Census block group containing multiple planning zones, land parcels, building footprints, and raster grids

characteristics of urban area in the formation of UHIs, this study used the RF method as the main analysis method.

The RF is a classification and regression technique introduced by Breiman [37], and this method is extremely useful in the study highly determinant variable selection because it provides variable importance measure as a part of the analysis results [37, 38]. The RF has been demonstrated to have improved accuracy in comparison to other machine learning methods and also, it is unexcelled in accuracy among current algorithm and it also runs efficiently on large dataset [37]. Moreover, it gives estimates of what variables are important in the classification and it is the main reason that this study used the RF methods over other algorithms to identify the relative importance of urban physical and socioeconomic characteristics to the formation of spatially varying land surface temperature across Marion County. The classification and regression trees are series of binary rule-based decisions that dictate how an input variable is related to its dependent variable. A forest is a collection of trees, and a RF consists of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of predictor values. A RF is random in two ways: (1) each tree is based on a random subset of the observations, and (2) each split within each tree is created based on a random subset of candidate variables [39]. When RF constructs a tree, an average of 36.8% (approximately 1/3rd of cases) of the observations are not used for any individual tree, and these hold-out cases are referred to as out-of-bag (OOB) [37, 39, 40]. In an OOB method, the errors from data excluded for each regression tree generation are used to inform RF of the relative strength and correlation of that tree [37]. This OOB data is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance. The variable importance measure of a variable can be explained as the contribution variables make to the construction of the tree. For a fixed number of trees, a variable with a larger importance score relative to other variables indicates that the variable is important for classification. Therefore, rather than estimate a specific relationship between the independent variables and the response as in data modeling, the variable importance measures are robust statistics pertaining to a variable's importance in the RF's emulation of the natural mechanism behind the data [41].

This study used RF add-on package with R statistical software. The R software offers several options for the RF model and its outputs. Among these, the variable importance plot (VIP) and percent increase in mean squared error (MSE) provide relative importance of the independent variables in the prediction of the dependent variable. To calculate percent increase in MSE and produce a VIP plot, the RF algorithm estimate the importance of each predictor by computing how much the error increases for a given tree when OOB data for each predictor are randomly permuted while all other predictors are left unchanged [38, 40]. Only two user-specified parameters are required to run RF: the number of trees in the forest, *ntree*, and the number of variables randomly sampled at each split, *mtry* [42]. Following Liaw and Wiener's [38] suggestion, this study operated RF with a default value of *ntree* 500 and that of *mtry* that is the square root of the number of variables in the dataset.

There is no universally accepted variable selection strategy for RF and suggested selection criteria from previous studies are limited [43]. Díaz-Uriarte and Alvarez de Andrés [44] suggested eliminating 20% of the variables having the smallest variable importance and building a new forest with the random variables. They stated the proportion of variables to eliminate is an arbitrary parameter of their method and does not depend on the data. This study followed the strategy proposed by Díaz-Uriarte and Alvarez de Andrés [44], eliminating 20% of variables having the smallest variable importance.

### Dimension reduction: principal component analysis

There are two objectives for variable selection. The first is to identify all the important variables, even with some redundancy, highly related to the dependent variable for explanatory and interpretation purpose, and the second is to find a sufficient parsimonious set of important variables for good prediction of the dependent variable [39, 42]. To achieve these two objectives, this study selected variable with RF method then validate its variable selection results with that of the PCA method.

The PCA is a non-parametric mathematical procedure that transforms a number of correlated variables into a smaller number of uncorrelated variables called principal component. The purpose of the PCA is to reduce a large set of variables to a smaller set that still contains most of the information in the large set without redundancy [45]. Thus this study conducted the PCA to identify principal components of the dataset and to use the PCA selection result as a verification tool for the variable selection results by rule-based machine learning technique whether the principal components are successfully identified as important variables. This study used SAS 9.4 software to identify principal components and the correlation between the principal components and the original variables.

## Results

As aforementioned in the previous section, if a variable is important, then rearranging the values of that variable will decrease in prediction accuracy. The amount of decrease in prediction accuracy is quantified as percent increase in MSE. Thus, if a variable is important, then the variable importance plot generated by the RF method will show relatively higher value than other variables.

In RF method, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error [37]. It is estimated internally, during the run as the RF method contains an inherent cross-validation procedure, which is used to identify the relative importance of each independent variable. Due to the randomness of the RF's clustering algorithm, the RF may return slightly different result in every trial [46], so this study run 10 consecutive rounds of the RF for each planning zone and averaged the percent increase in mean square error (MSE) for each variable. The variable importance measures showing percent increase in MSE for each independent variable by each planning zone are summarized from Figs. 4, 5, 6, 7, 8, 9, 10, 11. Together with RF method, the PCA was performed to identify the principal components for each zone and to identify the variables used to construct each principal component. As described in the previous section, this study applied the PCA to verify the variable selection results by machine learning technique whether the principal components are successfully identified as important variables. Results of the PCA for eight planning zones are summarized in Table 3.

### Central Business District

The CBD zone was one of the hottest planning zone of the Marion County, Indiana with the highest land surface temperature of 41 °C and the lowest of 25 °C. Following the variable selection strategy suggested by Díaz-Uriarte and Alvarez de Andrés [44], this study eliminated 4 of variables having the smallest of variance and selected other 14 variables as important variables. For CBD zone, urban biophysical variables including the percent developed imperviousness, building footprint, NDVI, maximum and average building height, percent tree canopy showed relatively high importance (Fig. 4). All of these variables were measured at the grid level. Together with these biophysical variables, socioeconomic variables including total number of female population, total number of population have education less than 9th grade, total population age under 18, median house value, total population age over 65, total number of non-white population, and median house value were selected with relatively higher variable importance (Fig. 4). With total 14 variables, the RF explains 73.25% of total variance in the dataset and the root mean squared error (RMSE) was 0.797 °C with the relative root mean squared error (rRMSE) was 4.98% (Fig. 4). The PCA method selected three principal components with 66.1% of total variance explaining demographic characteristics, household level

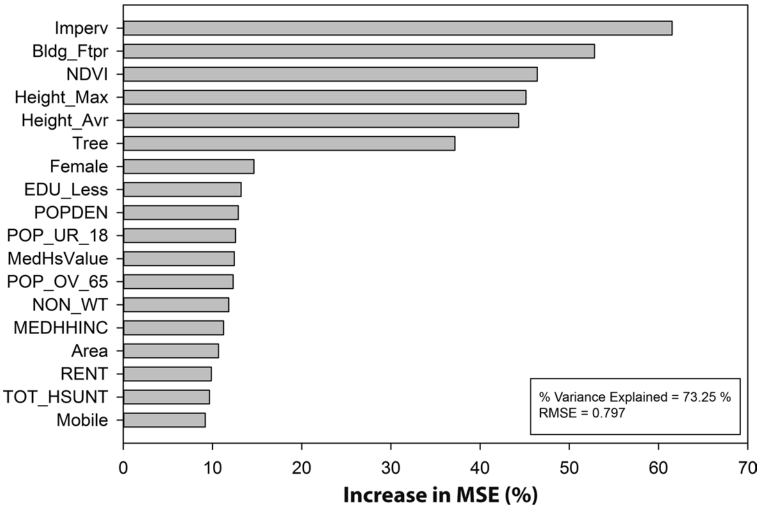

**Fig. 4** Variable importance plot showing increase in MSE (%) when each variable is held in out-of-bag data in the random forest model for the CBD zone

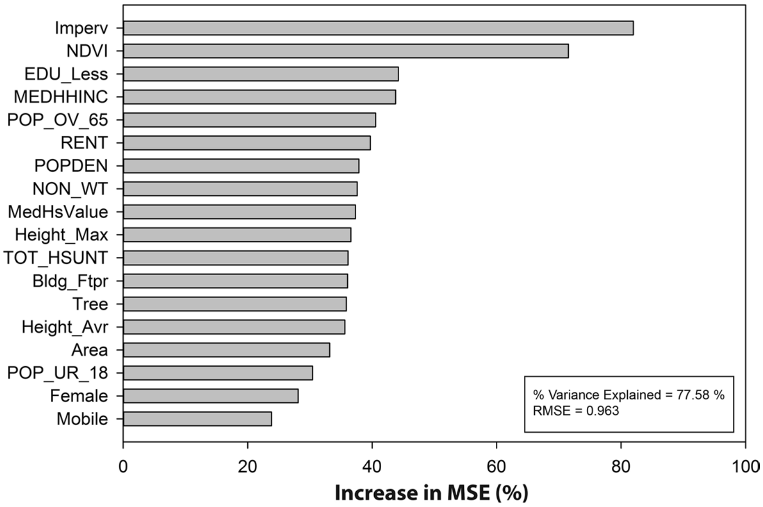

**Fig. 5** Variable importance plot showing increase in MSE (%) when each variable is held in out-of-bag data in the random forest model for the commercial zone

economic status, and urban vertical characteristics of the dataset. Variables included in the three principal components were also included in the 14 variables selected by the RF method, excluding the total number of mobile home variable which identified as the least important variable in explaining the formation of UHIs by the RF method (Table 3).

### Commercial district

Similar to the CBD zone, the commercial district zone is one of the hottest zones in Marion County, Indiana. Its maximum land surface temperature was 40 °C and the minimum was 17 °C. Total 14 variables including the percent developed imperviousness, NDVI, total number of population have education less than 9th grade, median household

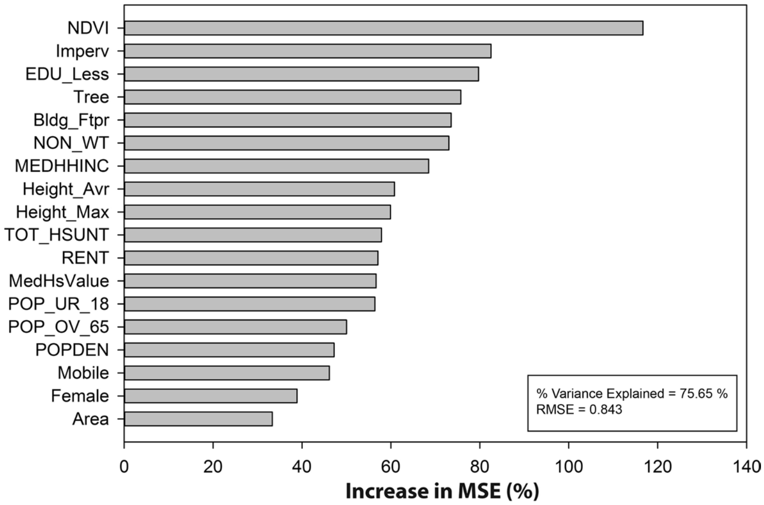

**Fig. 6** Variable importance plot showing increase in MSE (%) when each variable is held in out-of-bag data in the random forest model for the urban dwelling zone

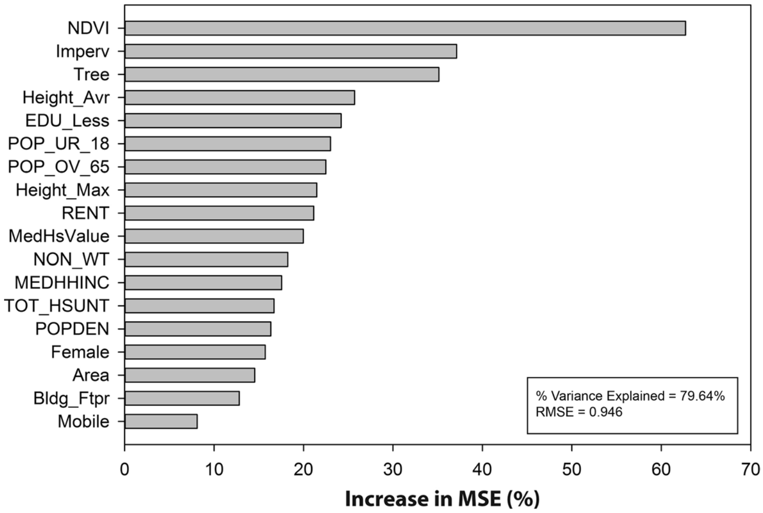

**Fig. 7** Variable importance plot showing increase in MSE (%) when each variable is held in out-of-bag data in the random forest model for the suburban dwelling zone

income, total number of population over age 65, total number of rental unit, population density of a Census block group, total number of non-white population, median house value, maximum building height, percent tree canopy, and average building heights were selected by the RF method (Fig. 5). Different from the CBD zone which showed the relatively higher importance of urban biophysical variables over the socioeconomic variables, the commercial district zone showed mixed signal. Biophysical variables such as the percent developed imperviousness and NDVI were selected with high importance, but other biophysical variables explaining building heights showed relatively lower variable importance than other two variables. Instead, the socioeconomic variables explaining education level, age, household level economic status were selected with relatively

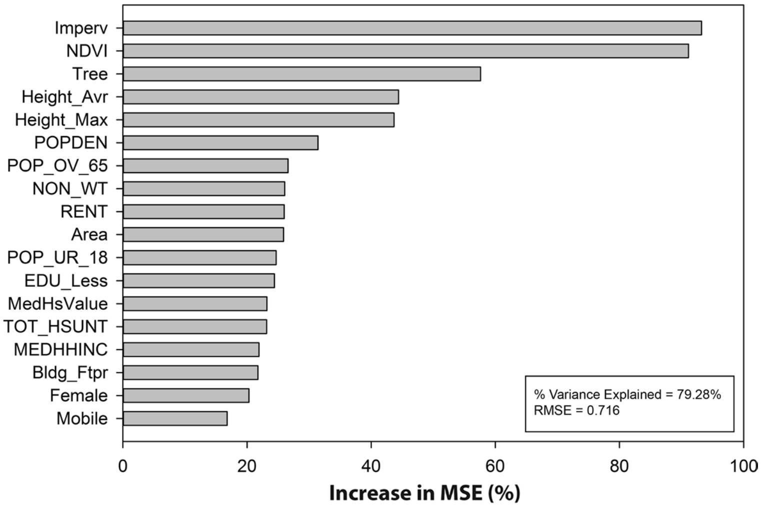

**Fig. 8** Variable importance plot showing increase in MSE (%) when each variable is held in out-of-bag data in the random forest model for the agricultural dwelling zone

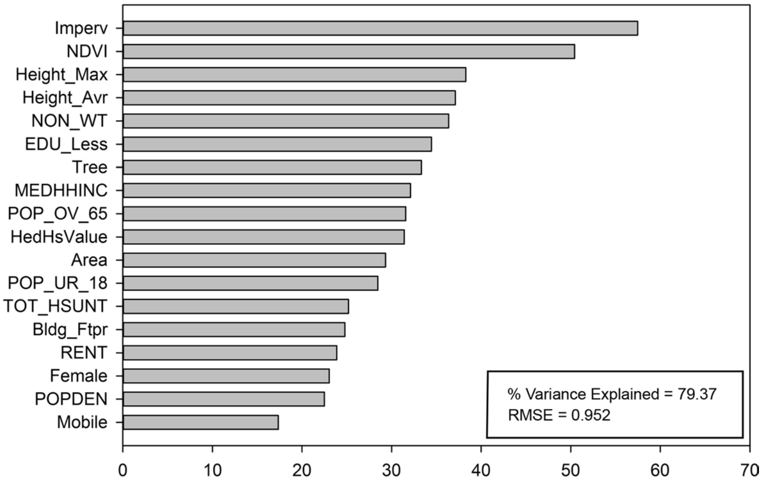

**Fig. 9** Variable importance plot showing increase in MSE (%) when each variable is held in out-of-bag data in the random forest model for the industrial urban zone

higher importance. With total 14 variables, the RF explains 77.58% of total variance in the dataset and the root mean squared error (RMSE) was 0.963 °C with the relative root mean squared error (rRMSE) was 4.19% (Fig. 5). The PCA method selected three principal components with 49.2% of total variance explaining demographic characteristics, household level economic status, and urban vertical characteristics (Table 3). Most of the variables used to construct three PCA excluding the total number of female population and the total number of the population under age 18 were selected in the total 14 RF variables (Table 3).

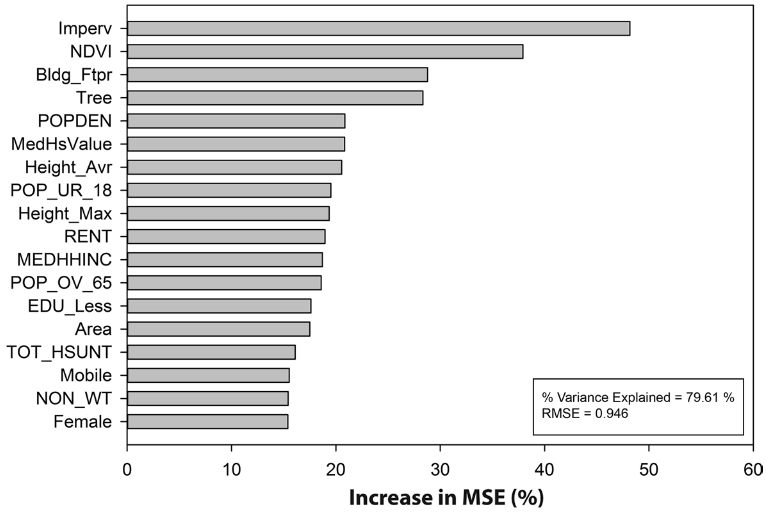

**Fig. 10** Variable importance plot showing increase in MSE (%) when each variable is held in out-of-bag data in the random forest model for the industrial suburban zone

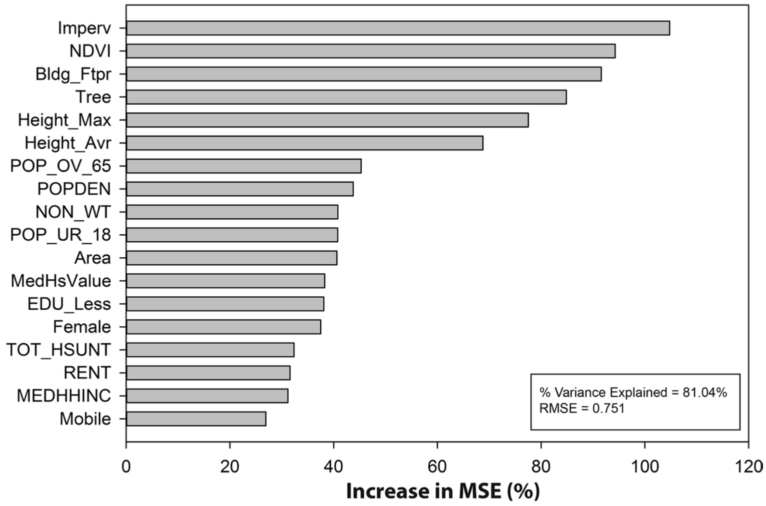

**Fig. 11** Variable importance plot showing increase in MSE (%) when each variable is held in out-of-bag data in the random forest model for the mixed-use zone

### Urban dwelling district

Similar to the result of the commercial district zone, the urban dwelling district zone selected both of the urban biophysical characteristics and socioeconomic characteristics with higher importance. Selected 14 variables are NDVI, the percent developed imperviousness, total number of population have education less than 9th grade, percent tree canopy, total building footprint, total number of non-white population, median household income, average building height, maximum building height, total number of household unit, total number of rental unit, median house value, total population under age 18, and total population over age 65 (Fig. 6). Demographic characteristics such as age and gender showed relatively lower importance, household level socioeconomic measures

**Table 3 Summary of the principal components analysis results by planning zones**

| Planning zone | Principal components | | Variables | % Variance explained |
|---|---|---|---|---|
| CBD | PC 1 | Demographic characteristics | POP_OV_65, female, NON_WT, mobile | 36.7 |
| | PC 2 | Household level economic status | MedHsValue, MEDHHINC | 18.0 |
| | PC 3 | Urban vertical characteristics | Height_Avr, Height_Max | 11.4 |
| | | | | Total % variance explained = 66.1 |
| Commercial | PC 1 | Demographic characteristics | POP_UR_18, female, TOT_HSUNT | 24.8 |
| | PC 2 | Household level economic status | MedHsValue, MEDHHINC | 12.6 |
| | PC 3 | Urban vertical characteristics | Height_Avr, Height_Max | 11.9 |
| | | | | Total % variance explained = 49.2 |
| Urban dwelling | PC 1 | Demographic characteristics | POP_UR_18, female, TOT_HSUNT | 22.7 |
| | PC 2 | Household level economic status | MedHsValue, MEDHHINC | 15.8 |
| | PC 3 | Urban vertical characteristics | Height_Avr, Height_Max | 11.6 |
| | | | | Total % variance explained = 50.1 |
| Suburban dwelling | PC 1 | Demographic characteristics | POP_UR_18, female, NON_WT, RENT | 24.4 |
| | PC 2 | Urban physical characteristics | Imperv, Height_Avr, Height_Max, Bldg_Ftpr | 14.5 |
| | PC 3 | Household level economic status | MedHsValue, MEDHHINC | 12.5 |
| | | | | Total % variance explained = 51.3 |
| Agricultural dwelling | PC 1 | Demographic characteristics | POP_UR_18, female, TOT_HSUNT | 28.0 |
| | PC 2 | Household level economic status | MedHsValue, MEDHHINC | 14.1 |
| | PC 3 | Urban vertical characteristics | Height_Avr, Height_Max | 12.9 |
| | | | | Total % variance explained = 55.0 |
| Industrial urban | PC 1 | Demographic characteristics | Female, TOT_HSUNT, Rent | 25.4 |
| | PC 2 | Urban biophysical characteristics | NDVI, Imperv | 16.1 |
| | PC 3 | Urban vertical characteristics | Height_Avr, Height_Max | 12.9 |
| | | | | Total % variance explained = 54.4 |
| Industrial suburban | PC 1 | Demographic characteristics | POP_UR_18, female, TOT_HSUNT | 30.0 |
| | PC 2 | Urban vertical characteristics | Height_Avr, Height_Max | 13.6 |
| | PC 3 | Household level economic status | MedHsValue, MEDHHINC | 11.2 |
| | | | | Total % variance explained = 54.8 |
| SE | PC 1 | Demographic characteristics | POP_UR_18, female, TOT_HSUNT | 31.6 |
| | PC 2 | Urban vertical characteristics | Height_Avr, Height_Max | 13.1 |
| | | | Total % variance explained = 44.7 | |

such as education level, median household income, and median household value shoed relatively higher importance. Similar to the variable selection result of the commercial district zone, urban dwelling district zone showed the balanced result in selecting biophysical and socioeconomic variables. With total 14 variables, the RF explains 75.65% of total variance in the dataset and the RMSE was 0.843 °C with the rRMSE was 4.01% (Fig. 6). The PCA selected three principal components explaining 50.1% of variance including principal components of demographic characteristics, household level economic status, and urban vertical characteristics. Same as the commercial district zone, total female population variable selected to explain the demographic characteristics principal component was not included in the RF selected 14 variables (Table 3).

### Suburban dwelling district

The zoning ordinance for the suburban dwelling district set a lower limit for the open space of a parcel as 70%, limiting excessive coverage by building structures and impervious covers. Due to that regulation, the highest land surface temperature of this planning zone was recorded as 29 °C and the lowest was 19 °C. The RF selected 14 variables including NDVI, percent developed imperviousness, percent tree cover, average building heights, total number of population have education less than 9th grade, total number of population age under 18, total number of population age over 65, maximum building height, total number of rental unit, median house value, total number of non-white population, median household income, total number of household unit, and population density. With total 14 variables, the RF explains 79.64% of total variance in the dataset and the RMSE was 0.946 °C with the rRMSE was 9.46% (Fig. 7). The PCA selected three principal components explaining the demographic characteristics, urban physical characteristics, and household level economic characteristics with 51.3% of total variance explained by the method. The total building footprint variable and the total number of female population variable were selected for PCA method mot not identified by the RF method as important variables (Table 3).

### Agricultural dwelling district

The zoning ordinance for the suburban dwelling district set a lower limit for the open space of a parcel as 75%, limiting excessive coverage by building structures and impervious covers. Thus, in the agricultural dwelling district zone, most of the urban biophysical variables excluding the total building footprints were identified as highly determinant variables in the formation of UHIs. The RF selected 14 variables including 5 biophysical variables and 9 socioeconomic variables such as population density of a Census block group, total number of population over 65, total number of non-white population, total number of rental unit, total area of a parcel, total number of population under age 18, total number of population have education less than 9th grade, median house value, and total number of household unit. The RF explains 79.28% of total variance in the dataset and the RMSE was 0.716 °C with the rRMSE was 4.21%. The PCA found three principal components explaining the demographic characteristics, household level economic status, and urban vertical characteristics with 55% of total variance explained by the method. Variables including the total number of female population and the total building footprint were identified as that of explaining principal components but not included in the RF selection result.

### Industrial urban district

In the industrial urban district zone, variables explaining biophysical characteristics of the area were selected with a higher level of importance. The RF selected variables including percent developed imperviousness, NDVI, maximum building height, average building height, total number of non-white population, total number of population have education less than 9th grade, percent tree canopy, population density of a Census block group, median house value, total number of population over age 65, total area of a parcel, total number of population under age 18, and total household unit. With these 14 variables, the RF explains 79.37% of total variance in the dataset and the RMSE was 0.952 °C with the rRMSE was 5.34% (Fig. 9). Industrial district zone was the only planning zone in Marion County that the variables explaining the urban biophysical characteristics were identified as one of the principal components. Variables including NDVI and percent developed imperviousness were identified as a principal component together with demographic characteristics and urban vertical characteristics. The PCA did not select any variables that explaining the household level socioeconomic characteristics as the principal component. The PCA result explains total 54.4% of the total variance of the dataset (Table 3).

### Industrial suburban district

Similar to the industrial urban district zone, variables explaining biophysical characteristics of the area were selected with a higher level of importance in the industrial suburban district zone. The RF selected other socioeconomic variables including population density of a Census block group, median house value, the total number of the population under age 18, the total number of the rental unit, median household income, the total number of the population over age 65, and total population have education less than 9th grade. With these 14 variables, the RF explains 79.61% of total variance in the dataset and the RMSE was 0.946 °C with the rRMSE was 4.11% (Fig. 10). The PCA found three principal components explaining the demographic characteristics, urban vertical characteristics, and household level economic status with 54.8% of total variance explained by the method. Among the variables identified to contribute to the construction of the principal components, a total number of female population and total household unit were not included in the RF variable selection results (Table 3).

### Mixed-use district

The uniqueness of the mixed-use district's land use pattern can be summarized by its vertical characteristics as its zoning ordinance does not regulate the maximum height of the building. In accordance with this uniqueness, the maximum building height and the average building height were identified as highly determinant variable in the formation of UHIs (Fig. 11). Together with building height variables, built area biophysical characteristics variables including percent developed imperviousness, NDVI, total building footprint, percent tree canopy, and the total area of a parcel were identified as important variables. Also, socioeconomic variables including total number of population over age 65, population density of a Census block group, total number of non-white population, total number of population under age 18, median house value, total number of population have education less than 9th grade, and total number of female population were found relatively important. With these 14 variables, the RF explains 81.04% of total variance in the dataset and the RMSE was 0.751 °C with the rRMSE was 3.41% (Fig. 11). Unlike previous planning zones, the PCA result for the mixed-use district zone selected two principal components including demographic characteristics and urban vertical characteristics. None of the socioeconomic variables were identified as one of the principal components. With these two principal components, total 44.7% of the variance of the dataset was explained (Table 3).

## Discussion and conclusion

The RF results successfully identified most of the variables used to construct the principal components for eight planning zones (Table 3) and also explained more than 73% of the variance in the dataset from all of the planning zones in Marion County, Indiana (Figs. 4, 5, 6, 7, 8, 9, 10, 11). The most important urban characteristics in the formation of UHIs of the Marion County, Indiana area was the percent developed imperviousness and NDVI, regardless of the planning district zone. This illustrates the importance of green spaces on the spatial variability of summertime land surface temperature. It also suggests that for the successful mitigation of UHIs in Marion County, Indiana, controlling biophysical characteristics of the urban area will bring more effective result than structural characteristics of the area.

As this study hypothesized that different sets of variables from each planning zone to be selected with different relative importance reflecting each planning zone's surface characteristics and socioeconomically different context, the RF method appeared to confirm this expectation. For example, planning zones with relatively higher pavedness such as CBD zone and the mixed-use district zone selected most of the biophysical variables with relatively higher importance than other independent variables used for the variable selection (Figs. 4, 5, 6, 7, 8, 9, 10, 11). Also, in the urban dwelling district zone, socioeconomic variables explaining educational attainment, ethnicity, household level economic status selected with relatively higher importance (Fig. 6), however in the suburban dwelling district zone, the relative importance of age-related variables were found higher (Fig. 7). These findings support the argument of this study expressing the need of planning zone specific UHIs mitigation strategies and action plans. The result of PCA, which is summarized in Table 3, also supported the main argument of this study that for different planning zones, different types of variables should be considered with different importance.

This study also confirmed that the RF selection results emphasized the importance of urban biophysical characteristics than the PCA results. Through previous empirical UHIs research, urban biophysical characteristics were proven as the most significant driving factors in the formation UHIs but these variables were seldom considered in the discussion of social vulnerability to the same environmental phenomenon. In this context, attention needs to be paid to the result that the relatively higher importance of biophysical variables identified by the RF method. Without exception, the percent developed imperviousness and NDVI were selected as the highly determinant variable in for formation of UHIs for all planning zones (Figs. 4, 5, 6, 7, 8, 9, 10, 11). Also, more developed planning zones, such as CBD district, where the lot coverage by impervious materials and buildings is higher than other planning zones, the relative importance of variables that explain the urban biophysical characteristics was found more significant than other socioeconomic variables (Fig. 4). Many previous UHIs research used number of rental units, parcel size, total number of house units, and total number of mobile home in the consideration of social vulnerability, however this study found with the application of RF method, these four variables are the least important variables in the formation of UHIs in the CBD district of Marion County, Indiana. Above mentioned findings should be considered when making UHIs mitigation plan by local government. As the main purpose of modern urban planning is resolving urban issues with the minimum economic costs and conflicts, the findings of this study will greatly contribute to the decision-making process determining where to focus and what should be done with higher priority.

Many of the previous empirical UHIs studies that tested the relative magnitude of independent variables to the formation of UHIs selected variables simply relying on the data availability and the result of previous studies. This is partly due to the lack of theoretical argument in favor of modeling method and variable selection method. This study believed more attention should be paid to the rule-based machine learning approach because the machine learning approach does not require to build a predictive model, thus are more flexible than conventional regressions. Consistently, this study found the rule-based RF variable selection results were better representations of urban biophysical and socioeconomic characteristics of Marion County, Indiana.

Another importance of this study lies in the proposition of more through variable selection method in the planning relevant scale. This study introduced a method for integrating socioeconomic data that measured in Census block group level, or even larger scale, and biophysical data that measured in a grid scale into a planning relevant scale. Also, the RF was plied to eight different planning zones, and consequently identified important variables for each planning zone, this zone specific variable importance can be used for developing the planning zone specific UHIs mitigation strategies.

### Abbreviations

CBD: Central Business District; DN: digital number; ETM+: enhanced thematic mapper plus; MSE: mean squared error; NDVI: The Normalized Difference Vegetation Index; NLCD: National Land Cover Database; OOB: out-of-bag; PCA: principal component analysis; RF: random forest; RMSE: root mean squared error; rRMSE: relative root mean squared error; TIR: thermal infrared; UHIs: urban heat islands; US EPA: US Environmental Protection Agency; USGS: US Geological Survey; US NOAA: US National Oceanic and Atmospheric Administration; VIP: variable importance plot.

### Acknowledgements

The author would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of this manuscript.

### Competing interests

The author declares no competing interests.

### Availability of data and materials

Dataset used for this manuscript that the author-generated from publicly available data including satellite images and Census Data is a part of the main arguments of this study. Thus the final dataset will not be available. In Table 1 of the main manuscript, the author summarized the publicly available original data sources in detail.

### Ethics approval and consent to participate

Not applicable.

**Funding**
Not applicable.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 19 October 2017 Accepted: 13 January 2018
Published online: 24 January 2018

**References**

1. Habeeb D, Vargo J, Stone B Jr. Rising heat wave trends in large US cities. Nat Hazards. 2015;76(3):1651–65. https://doi.org/10.1007/s11069-014-1563-z.
2. Stone B, Vargo J, Habeeb D. Managing climate change in cities: will climate action plans work? Landsc Urban Plan. 2012;107(3):263–71. https://doi.org/10.1016/j.landurbplan.2012.05.014.
3. U.S. Environmental Protection Agency. Climate change indicators in the United States, 2016. 4th ed. 2016. EPA 430-R-16-004. https://www.epa.gov/climate-indicators. Accessed 11 Oct 2017.
4. Arnfield AJ. Two decades of urban climate research: a review of turbulence, exchanges of energy and water, and the urban heat island. Int J Climatol. 2003;23(1):1–26. https://doi.org/10.1002/joc.859.
5. Hart MA, Sailor DJ. Quantifying the influence of land-use and surface characteristics on spatial variability in the urban heat island. Theor Appl Climatol. 2009;95(3):397–406. https://doi.org/10.1007/s00704-008-0017-5.
6. Oke TR. The energetic basis of the urban heat island. Q J R Meteorol Soc. 1982;108(455):1–24. https://doi.org/10.1002/qj.49710845502.
7. Stone B, Norman JM. Land use planning and surface heat island information: a parcel-based radiation flux approach. Atmos Environ. 2006;40:3561–73. https://doi.org/10.1016/j.atmosenv.2006.01.015.
8. Eliasson I. The use of climate knowledge in urban planning. Landsc Urban Plan. 2000;48(1–2):31–44. https://doi.org/10.1016/S0169-2046(00)00034-7.
9. Weng Q, Yang S. Managing the adverse thermal effects of urban development in a densely populated Chinese city. J Environ Manag. 2004;70(2):145–56. https://doi.org/10.1016/j.jenvman.2003.11.006.
10. Yuan F, Bauer ME. Comparison of impervious surface area and normalized difference vegetation index as indicator of surface urban heat islands effects in Landsat imagery. Remote Sens Environ. 2007;106:375–86. https://doi.org/10.1016/j.jenvman.2003.11.006.
11. Weng Q. Thermal infrared remote sensing for urban climate and environmental studies: methods, applications, and trends. J Photogramm Remote Sens. 2009;64(4):335–44. https://doi.org/10.1016/j.isprsjprs.2009.03.007.
12. Cutter SL, Mitchell JT, Scott MS. Revealing the vulnerability of people and places: a case study of Georgetown County, South Carolina. Ann Assoc Am Geogr. 2000;90(4):713–37. https://doi.org/10.1111/0004-5608.00219.
13. Cutter SL, Boruff BJ, Shirley WL. Social vulnerability to environmental hazards. Soc Sci Q. 2003;84(2):242–61. https://doi.org/10.1111/1540-6237.8402002.
14. Huang G, Zhou W, Cadenasso ML. Is everyone hot in the city? Spatial pattern of land surface temperatures, land cover and neighborhood socioeconomic characteristics in Baltimore, MD. J Environ Manag. 2011;92(7):1753–9. https://doi.org/10.1016/j.jenvman.2011.02.006.
15. Chen Z, Gong C, Wu J, Yu S. The influence of socioeconomic and topographic factors on nocturnal urban heat islands: a case study in Shenzhen, China. Int J Remote Sens. 2012;33(12):3834–49. https://doi.org/10.1080/01431161.2011.635717.
16. Jenerette GD, Harlan SL, Brazel A, Jones N, Larsen L, Stefanov WL. Regional relationships between surface temperature, vegetation, and human settlement in a rapidly urbanizing ecosystem. Landsc Ecol. 2007;22(3):353–65. https://doi.org/10.1007/s10980-006-9032-z.
17. Vargo J, Habeeb D, Stone B. The importance of land cover change across urban-rural typologies for climate modeling. J Environ Manag. 2013;114:243–52. https://doi.org/10.1016/j.jenvman.2012.10.007.
18. Weng Q, Lu D, Schubring J. Estimation of land surface temperature–vegetation abundance relationship for urban heat island studies. Remote Sens Environ. 2004;89(4):467–83. https://doi.org/10.1016/j.rse.2003.11.005.
19. Meng D, Li Z, Zhao W, Gong H. Quantitative exploration of the mechanisms behind the urban thermal environment in Beijing. Prog Nat Sci. 2009;19(12):1757–63. https://doi.org/10.1016/j.pnsc.2009.07.005.
20. Rhee J, Park S, Lu Z. Relationship between land cover patterns and surface temperature in urban areas. GIScі Remote Sens. 2014;51(5):521–36. https://doi.org/10.1080/15481603.2014.964455.
21. Cutter SL. The science of vulnerability and the vulnerability of science. Ann Assoc Am Geogr. 2003;93(1):1–12. https://doi.org/10.1111/1467-8306.93101.
22. Jesdale BM, Morello-Frosch R, Cushing L. The racial/ethnic distribution of heat risk-related land cover in relation to residential segregation. Environ Health Perspect. 2013;121(7):811–7. https://doi.org/10.1289/ehp.1205919.
23. Johnson DP, Wilson JS. The socio-spatial dynamics of extreme urban heat events: the case of heat-related deaths in Philadelphia. Appl Geogr. 2009;29(3):419–34. https://doi.org/10.1016/j.apgeog.2008.11.004.
24. Hamilton B, Erickson CL. Urban heat islands and social work: opportunities for intervention. Adv Soc Work. 2012;13(2):420–30.
25. Stone B, Vargo J, Liu P, Habeeb D, DeLucia A, Trail M, Hu Y, Russell A. Avoided heat-related mortality through climate adaptation strategies in three US cities. PLoS ONE. 2014;9(6):e100852. https://doi.org/10.1371/journal.pone.0100852.
26. Buyantuyev A, Wu J. Urban heat islands and landscape heterogeneity: linking spatiotemporal variations in surface temperatures to land-cover and socioeconomic patterns. Landsc Ecol. 2010;25(1):17–33. https://doi.org/10.1007/s10980-009-9402-4.

27. U.S. Census Bureau. Annual estimate of the resident population 2010–2016. American Community Survey 5—year estimates. 2016. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk. Accessed 11 Oct 2017.
28. U.S. National Oceanic and Atmospheric Administration. Global climate report—annual 2016. 2016. https://www.ncdc.noaa.gov/sotc/global/201613. Accessed 11 Oct 2017.
29. Landsat Project Office. Landsat 7 science data users handbook. 2016. https://landsat.gsfc.nasa.gov/wp-content/uploads/2016/08/Landsat7_Handbook.pdf. Accessed 11 Oct 2017.
30. Tucker CJ. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens Environ. 1979;8(2):127–50. https://doi.org/10.1016/0034-4257(79)90013-0.
31. Harlan SL, Declet-Barreto JH, Stefanov WL, Petitti DB. Neighborhood effects on heat deaths: social and environmental predictors of vulnerability in Maricopa County, Arizona. Environ Health Perspect. 2013;121(2):197–204. https://doi.org/10.1289/ehp.1104625.
32. Lundgren L, Jonsson A. Assessment of social vulnerability: a literature review of vulnerability related to climate change and natural hazards. Norrköping: Center for Climate Science and Policy Research; 2012.
33. Reid CE, O'Neill MS, Gronlund CL, Brines SJ, Brown DG, Diez-Roux AV, Schwartz J. Mapping community determinants of heat vulnerability. Environ Health Perspect. 2009;117(11):1730–6. https://doi.org/10.1289/ehp.0900683.
34. Tate E. Social vulnerability indices: a comparative assessment using uncertainty and sensitivity analysis. Nat Hazards. 2012;63(2):325–47. https://doi.org/10.1007/s11069-012-0152-2.
35. Liu X, Long Y. Automated identification and characterization of parcels with OpenStreetMap and points of interest. Environ Plan B Urban Anal City Sci. 2015;43(2):341–60. https://doi.org/10.1177/0265813515604767.
36. Desouza KC, Smith KL. Big data and planning. Chicago: American Planning Association; 2016.
37. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
38. Liaw A, Wiener M. Classification and regression by random forest. R News. 2002;2(3):18–22.
39. Grömping U. Variable importance assessment in regression: linear regression versus random forest. Am Stat Assoc. 2009;63(4):308–9. https://doi.org/10.1198/tast.2009.08199.
40. Walker WS, Kellndorfer JM, LaPoint E, Hoppus M, Westfall J. An empirical in SAR optical fusion approach to mapping vegetation canopy height. Remote Sens Environ. 2007;109:482–99. https://doi.org/10.1016/j.rse.2007.02.001.
41. Archer KJ, Kimes RV. Empirical characterization of random forest variable importance measures. Comput Stat Data Anal. 2008;52(4):2249–60. https://doi.org/10.1016/j.csda.2007.08.015.
42. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. Pattern Recognit Lett. 2010;31(14):2225–36.
43. Yoo S, Im J, Wagner JE. Variable selection for hedonic model using machine learning approaches: a case study in Onondaga County, NY. Landsc Urban Plan. 2012;107(3):293–306. https://doi.org/10.1016/j.landurbplan.2012.06.009.
44. Díaz-Uriarte R, de Andrés SA. Gene selection and classification of microarray data using random forest. BMC Bioinform. 2006;7(1):3–16. https://doi.org/10.1186/1471-2105-7-3.
45. National Institute of Standards and Technology. NIST/SEMATECH e-handbook of statistical methods. http://www.itl.nist.gov/div898/handbook/. Accessed 11 Oct 2017.
46. Walton JT. Subpixel urban land cover estimation: comparing Cubist, random forests, and support vector regression. Photogramm Eng Remote Sens. 2008;74(10):1213–22.