Application of Association Rule Mining for Exploring the Relationship between Urban Land Surface Temperature and Biophysical/Social Parameters

Umamaheshwaran Rajasekar and Qihao Weng

Abstract

This paper explores the relationship between remote sensing measurements of land surface temperature and biophysical/socioeconomic data by utilizing the association rule mining technique. The surfaces associated with urban uses typically radiate more heat as compared to its rural counterparts. There is a need to quantitatively analyze this contrast in temperature and the biophysical and social characteristics which influence it. Furthermore, in order to consider the urban heat island (UHI) effect, a parameterization is required to account for the urban surface characteristics impacts on the magnitude of land surface temperature (LST). The association rule mining model has demonstrated to bring in additional quantitative information concerning the relationships among urban parameters. The ASTER data from 2000 was used for the selection of appropriate variables to be used in the model. This information was then used for generating association rules between land-use land-cover (LULC) and LST information from 2000, 2001, and 2004. The results thus obtained quantitatively described the relationships between various urban parameters. It was found that there was little change in the percentage area of the LULC types from 2000 to 2004. This made the comparison of the results possible. In the case of the 2000 data, it was found that forest and impervious surfaces had strong association with temperature and scaled normalized difference vegetation index (SNDVI). Specific zones such as hospitals and universities had negative association with water. The comparison of data from 2000, 2001, and 2004 suggests that impervious surface and the zoning category of airport had a strong association. Nevertheless, the information extracted needs to be analyzed in greater detail in order to arrive at robust decision rules. Overall, the model so developed has demonstrated to be effective in predicting associations between urban LST and pertinent factors. This model could be useful for urban planners and environmental managers in quantifying rules that characterize a particular urban landscape.

Introduction

In the United States, the current urban growth rate, based on 1990 and 2000 census figures, is approximately 12.5 percent, with 80 percent of population residing within urban areas. As cities continue to grow, urban sprawl creates unique problems related to land-use, transportation, agriculture, housing, pollution, and development for policy makers (Shepherd and Burian, 2003). Demonstrated links exist between land-use land-cover (LULC) and temperature (Hawkins *et al.*, 2004, Weng *et al.*, 2004, Kalnay and Cai, 2003). There has been a significant amount of research conducted using thermal infrared measurements from ground and remotely sensed data with respect to urban and suburban characteristics leading to the well-known urban heat island (UHI) effect (Voogt and Oke, 2003; Chudnovsky *et al.*, 2004; Hung *et al.*, 2006; Khaikine *et al.*, 2006). Furthermore, changes in LULC and their impact on the UHI effect has been demonstrated by Golden (2004), Ghiaus *et al.* (2006) and Weng *et al.* (2004).

Factors such as urbanization, the conversion of other types of LULC associated with growth of population, and the economy impact the overall land surface temperature (LST) change in a specific area. These characteristics when amalgamated, contribute to the increase in the spread of UHI effect. This relationship between LULC and LST change has been measured qualitatively and has been characterized as overall change at a micro level using ground sensors during the day and night time. Fast *et al.* (2005) elegantly detailed the quantifiable parameters by installing data loggers at specific locations within a landscape. In this study, they attempted to determine the interaction of the physical variables with the environmental variables. Stathopoulou *et al.* (2004) studied the spatial temperature distribution and the intensity of the UHI of selected cities in Greece. Over the last three decades, great progress has been made with the advent of the space program associated with Earth observations (Kafatos *et al.*, 1998). Advancements in the field of UHI have been well reviewed by Arnfield (2003). Terra bytes of data are being collected from satellites, aerial sensors, telescopes, and other sensor platforms (Tan *et al.*, 2002). Some of the data collected suggests that skin temperatures for large areas could be mapped and studied more effectively by using satellite remote sensing data in the infrared region as compared to ground based sensor (Stathopoulou *et al.*, 2004).

Center for Urban and Environmental Change, Department of Geography, Indiana State University, 177 Science Building, Terre Haute, IN 47809 (qweng@indstate.edu).

Photogrammetric Engineering & Remote Sensing
Vol. 75, No. 3, April 2009, pp. 385–396.

0099-1112/09/7503–0385/$3.00/0

The literature already contains a considerable amount of studies, which have contributed to the use of remote sensing imagery for understanding the UHI effect. Lo *et al.*, (1997) analyzed UHI effect using Advanced Thermal and Land Application Sensor (ATLAS). Weng (2001 and 2003), Weng *et al.*, (2004 and 2006) have conducted studies to analyze the relationship of UHI effect with respect to the urban factors such as LULC, vegetation, and population. Streutker (2003) measured the growth of UHI in Houston, Texas using the split-window infrared channels of the Advanced Very High Resolution Radiometer (AVHRR). Jung *et al.*, (2005) modeled the effect of UHI on the vegetation using the hyper-spectral remote sensing imageries. Kato and Yamaguchi (2005) analyzed the effect of UHI using ASTER and ETM+ imagery. Hung *et al.*, (2006) assessed the UHI in the Asian mega-cities using the images from Aqua and Terra missions.

Statistical analysis may play an important role in linking urban temperatures to related factors. The interactions between the LULC (Dousset and Gourmelon, 2003; Weng, 2001) and NDVI (Gillies *et al.*, 1997; Eliasson, 1996; Weng *et al.* 2004; Lo *et al.*, 1997, Gallo and Owen 1999) with UHIs (including both surface and air temperature models) have been studied and quantified previously using linear statistical models and multivariate analysis. These techniques are well established but are effective in analyzing the quantitative relationship between limited qualitative variables and especially estimating large scale parameters. In recent years with the advent of computing, there is a parallel development in the field of methods and models within statistics to facilitate the analysis of huge repositories of datasets (Cabena, 1998). In this study, we attempt to address the issue of analyzing the relationship between several remote sensing derived parameters and GIS parameters with respect to LST using the technique of association rule mining (Agarwal and Srikant, 1994). The rationale being that the relationships derived using this model would be very beneficial for urban planners and environmental managers in simulating the type of impact certain LULC would effect and also aid in understanding the effect of new developments which might arise in the near future.

However, little has been done in the UHI research to quantitatively estimate the associations between the environmental, physical, and demographic variables with specific LULC types using data mining techniques. This study attempts to address this issue using the technique of association rule mining (Agrawal and Srikant, 1994) by conducting a case study in Indianapolis, United States, with ASTER image and ancillary geospatial data. The rationale being that the relationships derived using this model would be very beneficial for urban planners and environmental managers in analyzing and simulating the type of impact certain LULC would effect and also aid in understanding the effect of new developments which might arise in the near future. More specifically, the objectives of this study are: (a) to isolate and determine variables which strongly associate with LULC, (b) to test the variables selected to determine and access associations between socio-economic and LST data for years 2000, 2001, and 2004, and (c) to analyze the effect of urban and/or environmental variables on various LULC types based on the results from the previous two steps.

Study Ar ea and Data

Indianapolis/Marion County, Indiana, USA, was chosen as the study area. It possesses several characteristics that make this area an appropriate choice for such a study. Indianapolis has a single, central city. Large urban areas in the vicinity have not influenced its growth. The city is located on a flat plain and is relatively symmetrical, having possibilities of expansion in all directions. Like many other American cities, Indianapolis is rapidly increasing in population and in area. The areal expansion occurs through encroachment into the adjacent agricultural and non-urban land. Certain decision-making forces, such as density of population, distance to work, property value, and income structure, encourage some sectors of metropolitan Indianapolis to expand faster than others. Extracting information of LULC from satellite images allows for monitoring urban changes over time (Weng *et al.*, 2004). Figure 1 depicts the study area and its surrounding.

In this research, we used satellite-derived thermal infrared data instead of ground base air surface temperature data. In spite of the advancement in ground sensors for measuring the air temperature at a high temporal scale, it is satellite remote sensing which helps in attaining high spatial resolution data of the skin surface temperature. The association of the temperature and the LULC type as derived from the ground based sensors cannot be directly incorporated to satellite remote sensing images due to high uncertainty in the data; the uncertainty being albedo, cloud

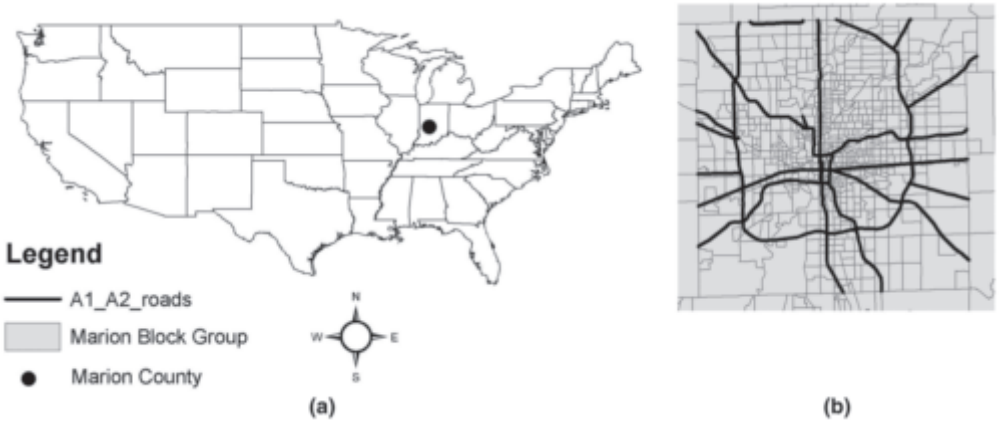

Figure 1. (a) the Indiana state and the region of study Marion County which encompasses the City of Indianapolis, and (b) the block groups and the transportation networks of Indianapolis City which was under study.

cover, wind, shadows due to buildings, and canopy structure. Furthermore, the number and spread of existing air surface temperature data loggers situated in and around the study area were inadequate to cover the entire study area with the precision necessary for a local study. Therefore, ASTER-derived temperature data was used for this study. ASTER thermal images were acquired on dates of 03 October 2000 (12:00:51 local time), 16 June 2001 (11:55:29 local time) and 05 April 2004 (11:46:39 local time). Geospatial data sets such as transportation networks, 2000 population census, and zoning data were also used for this research.

Data Preparation

Land Surface Temperature

In this study, ASTER band 13 (10.25 μm to10.95 μm) was used to calculate the LST due to the spectral width of this band bring close to the peak radiation of the black body spectrum given off by the urban surface in the study area (Lu and Weng, 2006). There are two steps which are involved in the computation of LST. First was the conversion of sensor derived spectral radiance to the at-sensor brightness temperature (i.e., considering that the emitting source is a perfect black body). Since the emitting source is never a perfect black body, the second step was the correction for spectral emissivity taking into consideration the LULC types. The conversion formula for the first step as derived by Dash *et al.*, (2002) was used for the computation. The formula is as follows:

$$T_c = C_2 / \lambda_c \, ln \, ((C_1 / \lambda_c^5 \, \pi \, L_\lambda) + 1) \quad (1)$$

where T_c is the brightness temperature in degrees Kelvin (K) from the central wavelength, L_λ is the spectral radiance in $Wm^{-3}sr^{-1}$, λ_c is the sensor's central wavelength in our case the central wavelength of the band 13 of the ASTER imagery, C_1 is the first radiation constant ($3.74151 * 10^{-16} Wm^{-2}sr^{-1}\mu m^{-1}$), and C_2 is the second radiation constant (0.0143879mK)

The black body temperature values were then corrected for the spectral emissivity (ε). This was done using the LULC classification derived from ASTER data (using optical bands). The LULC categories were then assigned an emissivity value according to the scheme provide by Snyder *et al.*, (1998). The LULC categories and their corresponding emissivity values were used for the calculation provided in Table 1.

The emissivity corrected LST was computed as follows (Lu and Weng, 2006):

$$LST = T_c / (1 + (\lambda * T_c / \rho) \, ln \, \varepsilon \quad (2)$$

where λ is the wavelength of the emitted radiance (for which the peak response and the average of the limiting wavelengths was used), $\rho = h * c/\sigma$ where σ = Boltzmann constant ($1.38 * 10^{-23} JK^{-1}$), h = Planck's constant ($6.626 * 10^{-34}$J.s, and c = velocity of light ($2.998 * 108 ms^{-1}$), which equals to $1.438 * 10^{-2}$mK.

TABLE 1. 2004 LULC CLASSIFICATION AND THEIR CORRESPONDING SPECTRAL EMISSIVITY

Class	LULC Type	Spectral Emissivity Assigned
1	Impervious Surfaces	0.966
2	Barren Lands	0.977
3	Grass Lands	0.972
4	Agriculture	0.973
5	Forest Land	0.987
6	Water	0.991

The so developed LST image was numeric data. Since the association rule mining is only possible with the nominal or ordinal data, the image was classified into categorical information. After experimenting with several classification algorithms, natural break algorithm was selected for the final model. The natural breaks method is based on the assumption that the data fall naturally into meaningful groups (Smith, 1986). Our assumption is that this would make the inference from the classified system easy to understand and also the values which fall within a class could be made consistent over time (i.e., while analyzing temporal datasets). The same algorithm was also found to be best for association rule mining within other geographical applications such as analysis of urban socioeconomic and land-cover change (Mennis and Liu, 2005; Mennis, 2006). Table 2 details the classes and their minimum and maximum values for years 2000, 2001 and 2004.

LULC

The LULC data was developed from the ASTER 2000 image using semi-automatic technique. An unsupervised classification method (Iterative Self-Organizing Data Analysis) was chosen to classify ASTER data with the maximum iterations of 30. One hundred twenty clusters were created and labeled in reference to 2003 and 2005 aerial photos. Reclassification was then executed for the confusing regions. Post classification smoothing and image refinement were also conducted to improve the accuracy of image classification. Classification accuracy for each image was assessed against the 2003 county aerial photo. A stratified random sampling method was applied to choose 50 samples in every LULC category. The overall accuracy was above 85 percent. For a detailed description of the method implemented and the results acquired refer to (Liu and Weng (2008). Figure 2 shows the LULC map of seven classes (excluding the background). In the final classification of Marion county (the study area), the class "wetlands" was not present, therefore it was removed and is therefore not listed in Table 3.

Scaled Normalized Difference Vegetation Index (SNDVI)

The NDVI in this study was calculated from the three ASTER images using Equation 3.

$$NDVI = (NIR - R)/(NIR + R) \quad (3)$$

where, *NIR* and *R* refer to very near infrared bands VNIR-Band3N (B3) and VNIR-Band2 (B2), respectively, within the ASTER image data set.

The resulting image was scale transformed from scale −1-1 to 0-2 (basically the values were converted from negative domain to positive domain to facilitate processing) forming SNDVI. These values were then converted into categorical data of ten classes using the natural breaks algorithm. The categorical data and their corresponding maximum and minimum values are given in Table 2

Population Density Dataset

One of the main reasons for inclusion of population data was to study the relation between the population density and the LST since population is one of the major factors which indicate the urban sprawl and spread. It is also the reason why the UHI effect is proved to be more prominent in the developed cities. The population density dataset was developed from the 2000 population census. The spatial map of the population along with the block identifiers was obtained directly from the Indiana Geological Survey. There were three levels of population data representing block, block group, and tract level data. Block group level population

TABLE 2. LAND SURFACE TEMPERATURE AND SNDVI DERIVED AND CLASSIFIED FROM ASTER DATA SET ALONG WITH THEIR RESULTS FOR THE YEARS 2000, 2001, AND 2004

From LST 2000	To	Class	% Pixels	From SNDVI 2000	To	Class	% Pixels
0	290.99	1	1.22	0.3	0.63	1	1.78
290.99	294.77	2	5.54	0.63	0.74	2	5.36
294.77	297.29	3	18.53	0.74	0.81	3	6.2
297.29	299.81	4	32.6	0.81	0.88	4	8.8
299.81	302.33	5	26.88	0.88	0.94	5	13.45
302.33	304.85	6	11.12	0.94	1.01	6	12.76
304.85	307.37	7	3.04	1.01	1.07	7	14.66
307.37	311.15	8	0.82	1.07	1.13	8	15.52
311.15	316.19	9	0.2	1.13	1.19	9	13.68
316.19	322.49	10	0.04	1.19	1.45	10	7.77
LST 2001				**SNDVI 2001**			
0	296.37	1	7.06	0.3	0.69	1	1.94
296.37	300.29	2	22	0.69	0.77	2	5.46
300.29	302.9	3	21.92	0.77	0.85	3	6.2
302.9	305.51	4	20.73	0.85	0.93	4	7.48
305.51	308.12	5	15.12	0.93	1.01	5	10.79
308.12	312.04	6	10	1.01	1.08	6	12.82
312.04	315.95	7	2.52	1.08	1.16	7	14.04
315.95	321.18	8	0.5	1.16	1.23	8	14.4
321.18	327.7	9	0.11	1.23	1.31	9	14.06
327.7	334.23	10	0.03	1.31	1.52	10	12.8
LST 2004				**SNDVI 2004**			
267.62	279.41	1	0.02	0.46	0.81	1	1.13
279.41	285.79	2	2.79	0.81	0.93	2	1.21
285.79	288.69	3	9.63	0.93	1.03	3	5.75
288.69	290.82	4	16.98	1.03	1.1	4	8.13
290.82	292.56	5	24.78	1.1	1.16	5	11.72
292.56	294.3	6	21.84	1.16	1.22	6	19.01
294.3	296.23	7	14.79	1.22	1.28	7	20.37
296.23	298.94	8	7.07	1.28	1.35	8	15.01
298.94	303.96	9	1.8	1.35	1.44	9	11.07
303.96	317.11	10	0.31	1.44	1.67	10	6.57

statistics (658 blocks within the Marion County/the city proper of Indianapolis) were selected for our study, since it was a more detailed representation as compared to the other two. The assumption was that more variation at a pixel level would improve the strength of relationships between the variables of interest. The population map was then converted into the population density raster dataset by dividing the populations by the map area, i.e., block group population data was divided by the area of the specific block group. The resultant density raster was further classified into ten classes using the natural break algorithm. Figure 2a represents the population density dataset developed for the study area.

Zoning Data

The municipal zoning data depicting different land-use zones of Marion County was used for this study. From the initial classification, the data was reduced to 13 zones. The reduction was based on basic zone types. For example, different sub-zoning types of residential classes were classified as one single type, i.e., residential. The final 13 zones are shown in Figure 2b.

Transportation Buffer Zones

One of the main rationales for using transportation data base was that earlier studies have proved that the impervious surfaces tend to generate more heat. Furthermore, within Indianapolis City, similar to other cities, many of the major commercial/industrial structures are situated along the highways. Therefore, we hypothesized that inclusion of the highways and the buffer zones from them would give us more understanding of the relation between these infrastructures and the UHI effect. The transportation buffer zones were created from the road network maps. Since the base LST image which was used for mining relationships was of 15 m (all ASTER bands were resampled to 15 m) resolution, only the major roads of class A1 and A2 (as designated by the Department of Transportation) were included in the study based on expert knowledge considering the local settings. The reason for not including other road types was mainly due to their relatively narrow road widths. Since 15 m ASTER LST was used for the analysis, the influence of these roads over the 15 m grid cells was relatively small. Whereas, the classes A1 and A2 roads were mainly highways and freeways which are present and/or pass through Marion County. In order to find the influence and association (if any) of these highways on the LST, buffer zones of 400 m on each side of the road were created, for a total distance of 4 km. This zonal map was then converted into raster and resampled to the same dimension as that of LST (15 m). Figure 2c and 2d show the transportation networks.

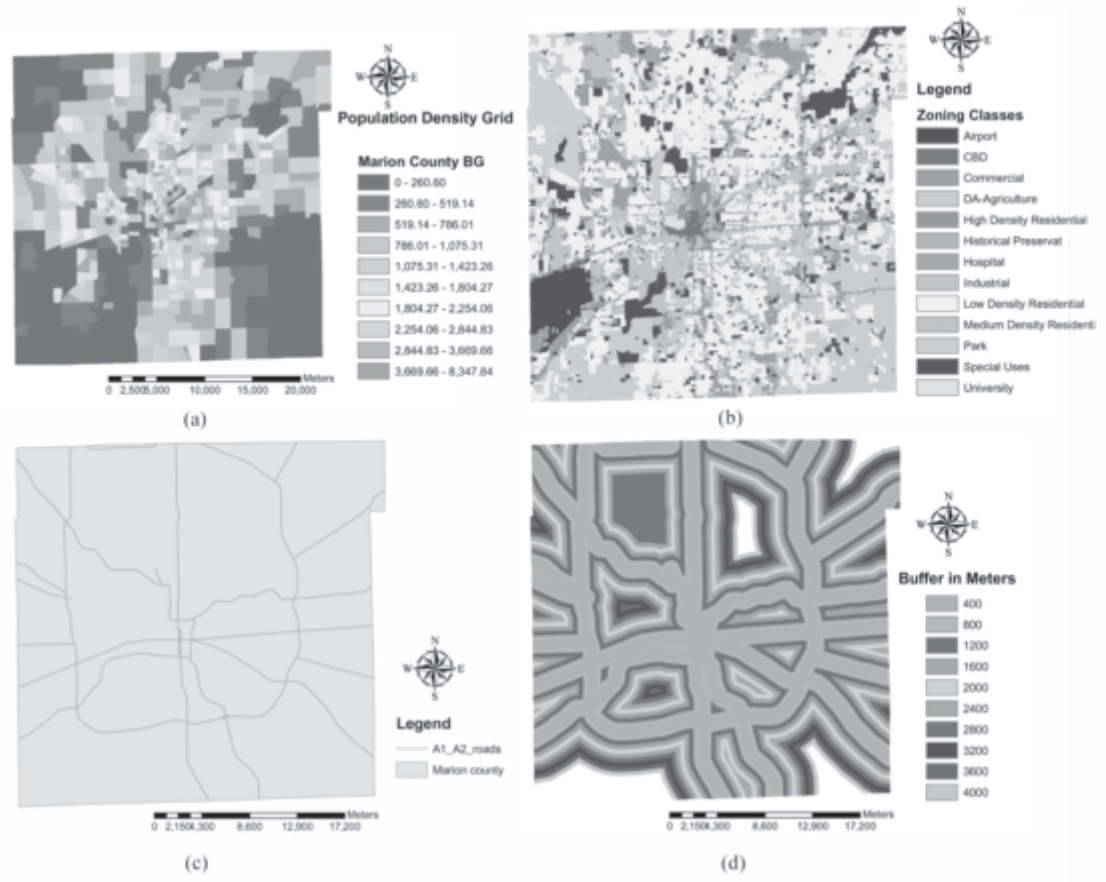

Figure 2. (a) the population density dataset, (b) the zoning dataset, (c) the major roads which were used in the creation of the transportation buffer zones, and (d) the transportation buffer zones and their corresponding distances in various colors. A color version of this figure is available at the ASPRS website: **www.asprs.org** .

TABLE 3. LULC CLASSIFICATION AND THEIR CORRESPONDING PERCENTAGES OF PIXELS WITHIN EACH IMAGE FOR THE TIME PERIOD 2000, 2001, AND 2004

LULC Class	Description	% Pixels 2000	% Pixels 2002	% Pixels 2004
Impervious Surfaces	Industrial lands, roads and rails, commercial, right-of-way, all building and built structures > 15 square meter within golf courses, soccer and recreation areas, towers, and so on	31.61	32.71	31.96
Barren Land	Active mine-lands, active quarries, bare dunes, and so on	0.64	0.61	0.63
Grassland	Prairies, pasture, savannas, historic grasslands, farm bill program lands, caves, and subterranean features, and so on	28.76	28.35	29.51
Agricultural Land	Row crop by type, cereal grains, vineyards, feedlots, residue management, and confined operations, and so on	6.56	6.64	6.68
Forest	Successional stage, like pre-forest stage and mature or high canopy stage, and so on	29.6	28.7	28.24
Water	Lakes, rivers and streams by order and watershed, miles of unimpounded rivers and streams, and so on	2.83	2.98	2.98

Methodology

Data Mining is a field which has been developed by encompassing principles and techniques from statistics, machine learning, pattern recognition, numeric search, and scientific visualization to accommodate new data types and data volumes being generated (Miller and Han, 2001). The tasks of data mining might vary, but the premise about discovering unknown information from large databases remains the same. In short, data mining can be defined as "Analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owners" (Hand *et al.*, 2001). Over the last few years, the techniques of data mining have been pushed by three major technological factors which have advanced in parallel. First, the growth in the amount of data has led to the development of mass storage devices. Second, the problem of accessing this information has led to the development of advanced and improved processors. Third, the need for automating the tasks involved in data retrieval and processing led to the advancements in statistic and machine learning algorithms.

In this research, the technique of association rule mining was utilized to study the quantitative relationship between LST patterns and other environmental, physical, and demographic data. Even though the very nature of urban LST is dependent upon various influencing factors, the nature of dependence between LULC and LST has proved to be strong (Weng *et al.*, 2004). Therefore, the main rational of this research was to model, analyze, and quantify this relationship. The results of this analysis in terms of the variables which strongly associate with the LULC of Marion county (Indianapolis) for the year 2000 was then used to analyze the association of the same variables for different time periods (2001 and 2004) to detect and analyze the seasonal effect and the effect of changing landscapes.

Association Rule Mining

An association rule is a relationship of the form $X \geq Y$, where X and Y are sets of items. X is called antecedent and Y the consequence. There are two primary measures, support and confidence, used in assessing the quality of the rules. The goal of association rule mining is to find all the rules with support and confidence exceeding user specified thresholds (Ding *et al.*, 2003). The brief explanation of the terms support and confidence are as follows (modified from Borgelt and Kruse, 2002).

Support of an Item Set

Let T be the set of all transactions under consideration, e.g., let T be the set of all the attributes (in our case LULC, population density, LST, SNDVI, zoning and distance to the major roads) which were recorded on an image, ν. Let LC be a subset of certain type which contains a particular set of attributes, then the support of type LC is the percentage of pixels in T which cover similar attributes. For example, LC is a subset which has the attributes: LULC = urban, LST = 20°C and distance = 400 m from the road network. Then if ε is the set of all pixels within the image T which contains all the attributes of LC, then:

$$\text{Support (LC)} = |\varepsilon|/|T|) \ *100\% \tag{4}$$

where $|\varepsilon|$ and $|T|$ are the number of pixels in ε and T, respectively.

Confidence of an Association Rule

This is the measure used to evaluate association rules. The confidence of a rule $R = A$ and $B \Rightarrow C$ is the support of the set of all items that appear in the rule divided by the support of the antecedent of the rule, i.e.,:

$$\text{confidence } (R) = (\text{support}(\{A, B, C\})/ \text{support}(\{A, B\}))*100\% \tag{5}$$

The confidence of a rule is the number of cases in which the rule is correct relative to the number of cases in which it is applicable. For example, let R = land-cover urban and distance from highway 400 m $\Rightarrow$ 20°C temperature. It means that for a given pixel, if LULC is urban and if its location is around 0 to 400 m from the major highways then the temperature value recorded for such region is 20°C. It also means that for a given pixel, if LULC is not urban or the location is not around 0 to 400 m or neither, than the rule is not applicable, and does not say anything about the resulting temperature values. If the rule is applicable, it means that the resulting temperature can be expected to be 20°C. At the same time, if the rule is found to be applicable, but the temperature is not as expected then, the rule may not be correct. Since we are interested in how good the rule is, i.e., how often its prediction that for a given attribute type that a constant temperature turns out. The rule confidence measures this: It states the percentage of cases in which the rule is correct. It computes the percentage relative to the number of cases in which the antecedent holds, since these are the cases in which the rule makes a prediction that can be true or false. If the antecedent does not hold, then the rule does not make a prediction, so these cases are excluded.

With this measure a rule is selected, if its confidence exceeds or is equal to a given lower limit. That is, we look for rules that have a high probability of being true, i.e., we look for "good" rules, which make correct (or very often correct) predictions. Care should be taken that, within any rule if $LC \Rightarrow T$ is true with a confidence of certain percentage then it does not mean that $T \Rightarrow LC$ is also true with the same percentage. For example in our case if Land-cover type urban $\Rightarrow$ Temperature of range 20°C to 30°C with a confidence level of 90 percent, it means that if a LULC is found to be of type urban than the chance that its temperature being within the range 20°C to 30°C is 90 percent, but at the same time it does not mean that if there is a pixel with a temperature of 20°C to30°C then the probability of it being of LULC type urban is 90 percent. Therefore, it can also be said that the rules are unidirectional.

Support of an Association Rule

In this research, the support of a rule is the same as the support of the antecedent of the rule. For example, if the antecedent is LULC type agriculture results in temperature above 30°C, then the support of the rule within this model is calculated from the antecedent of the rule, i.e., the number of occurrence of LULC type agriculture and not the number of cases in which the rule is correct. In mathematical terms, one can represent the support of the rule $X \Rightarrow Y$ as the support of X; and the confidence of a rule $X \Rightarrow Y$ the conf $(X \Rightarrow Y)$ as the ratio supp $(X \cup Y)$/*supp* (X) According to the support–confidence framework (Zhang and Zhang, 2002), and rules of association by Silverstein *et al.*, (1997), support of a rule can be enumerated as follows:

1. $X \cap Y = \phi$,
2. $p(X \cup Y) \geq$ minsupp
3. $p(Y \mid X) \geq$ minconf (e.g., conf$(X \rightarrow Y) \geq$ minconf), and
4. $|\, p(X \cup Y)/(p(X)p(Y) - 1|$.

The valid association rule $X \Rightarrow Y$ can be extracted to a valid rule of interest. The above set of rules describe the method of extracting association rules, further conditions

could be added to generate positive association rules within which the confidence of the rules are higher and negative association rules within which the confidence of the rules are lower. In initial scenario would be to find the positive relationship between objects (for e.g., if LULC type A is present then the temperature should be within range "*X*") the later would help in finding the negative relationship between objects (for e.g., if LULC type A is present then the temperature would most likely not be within range "*X*"). Since our research was aimed at extracting the parameters which have impact on the LST the concentration was mainly given toward extracting positive association rules.

The Model

Within the paradigm of data mining, a model is a high-level, global description of a data set (Hand *et al.*, 2001). It takes a large sample perspective. The models can be categorized into two major divisions: one is a descriptive model, which is used for summarizing the data in a convenient and concise way, and second is inferential or predictive models, which allow one to make some statements about the population from which the data were drawn or about likely future data values (Hand *et al.*, 2001). In order to arrive at an appropriate class of models, one needs to understand the data and the study at hand. Since the objective of this study was to quantitatively estimate the relationship between variables such as LULC, temperature, population density, transportation network, SNDVI, and zoning details, the predictive model (association rule mining) was applied.

Figure 3 outlines the conceptual model used in this research. The first part of this study involved the selection of appropriate variables which associate strongly within the datasets using the 2000 ASTER derived LST, SNDVI, and LULC. The second part involved utilizing the selected variables on the remaining (2001 and 2004) ASTER derived databases to remove any season effects and urban LULC change over time which might be associated with the data used. The third part involved in extracting knowledge in terms of generalized rules from these results.

Results and Discussion

The results are divided into two parts. One deals with the results obtained for the association within the year 2000 LULC classification and its relevant data sets such as the SNDVI, LST, population density, zoning, and distance from the major transportation network (in this case the highways). The other is the comparison of the results of the same data mining model between the years 2000, 2001, and 2004.

After experimenting with various confidence intervals, a confidence limit of 80 percent and above was chosen as the threshold for the data mining model. This means that the rules discussed in this study have a probability of 80 percent or more confidence. The initial rule mining was carried out with the year 2000 data which included the ASTER-derived LULC information, SNDVI, LST, and GIS datasets such as zoning information, population density, and transportation buffer zones. The results of the simulation are discussed below:

- There existed no strong association between the LULC variable and the transportation buffer zones. This might be due to main spatial arrangement of the city. The main reason for including the transportation buffer zoned areas was to identify any strongly associated LULC types occupying the area around the major highways. Since Marion County had a mixed LULC around the highways, there were no association rules that can be derived with the confidence of 80 percent or higher.
- Population density below 261 people per square kilometer had association with the LULC types barren and agriculture with a confidence of 90 percent and 81 percent, respectively. It was also found that the LULC type impervious surfaces had a strong association with the population density greater than 4,435 people per kilometer square. From these rules, we can summarize that in Marion County, high population led to high impervious surfaces. But the derived rules addressed to only a small portion of the data sets (less than 40 percent support) and the major part of the population density had no direct relationship with the LULC classes. This result is similar to that of an earlier study by Weber and Hirish (1992) which found that remote sensing derived LULC data had a more obvious relationship with the housing structure than with the population structure (see Figure 4).
- An unexpected result was the correlation of pixels with high surface temperatures and hospital zoned areas irrespective of the type of LULC. This association implies a lack of biomass which helps in ameliorating elevated surface temperature (see Figure 5).
- LULC type impervious surface had strong associations with medium to high LST and low to medium SNDVI, i.e., less than or equal to 0.81 (see Figure 6).

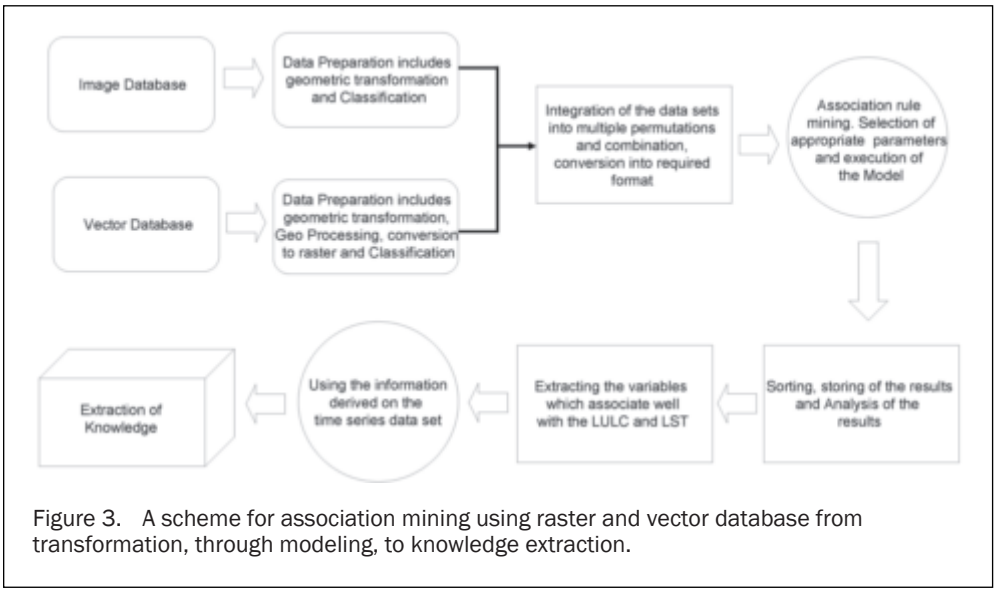

Figure 3. A scheme for association mining using raster and vector database from transformation, through modeling, to knowledge extraction.

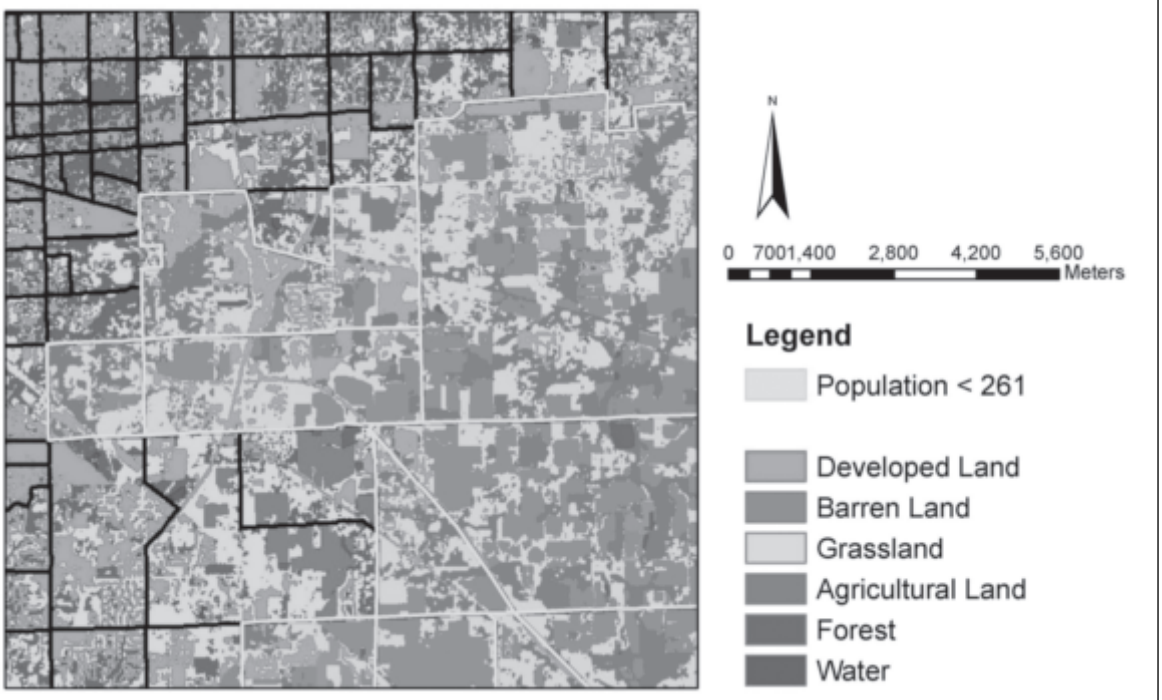

Figure 4. A sample area showing the relationship between population density and LULC. A color version of this figure is available at the ASPRS website: **www.asprs.org**.

- The model indicates that LULC types of water bodies, agriculture, and forest had negative association with university zoned areas. This implies that, the percentage area of these LULC types, which have been demonstrated to aid in the reduction of surface temperatures were not positively associated with university zoned areas which upon preliminary investigation seemed counter intuitive.
- 90 percent of the central business district was occupied by the LULC type impervious surfaces.
- An interesting result with a wide area of implications was the association of water temperatures. In zoned areas of low density housing, special uses, and parks, the water temperature demonstrated similar low temperatures with the surrounding zones. However, in other zoned areas the water temperature was highly variable with no strong associations with its surrounding. This might be due to the shadow effect and/or the depth of the water within those zones. Further site investigation is needed to find more about this anomaly exhibition by the skin surface temperature measurement (see Figure 7).

Well established relationships were verified by the model which aids in communicating already explored phenomenon. This relationship between forest LULC and low temperatures aids in articulating the accuracy of the model.

- The LULC type forest was found in the areas where temperature was low, the SNDVI was high, and the zoning is either commercial or residential (which includes medium density and low-density housing).
- The lowest temperatures were associated with the LULC water bodies. The highest temperature was associated with the LULC type impervious surfaces.

Based on the rules derived from this model, two of the variables, i.e., population and transportation buffer zones, were removed due to its lack of association with the remaining variables. The main reason for the two variables (population and transportation) not establishing association with the remaining variables might be due to the modifiable areal unit problem (MAUP) (Openshaw 1983; Quattrochi *et al.*, 1997; Jenerette *et al.*, 2007). Since the rest of data excluding the zoning are at 15 m resolution, the population and transportation data set were resampled to fit the scale. One of the inferences might be that, these two datasets if available at relatively higher or micro scale their effect could have been realized. As with the case of zoning dataset, even though it was used at a macro scale as in comparison with the rest of the variables, the rules utilized for their creation should have been based upon LULC characteristics making it fit the model well. Even though there might be a scale problem (where the same set of areal data is aggregated or divided into several sets of areal units (Jelinski and Wu, 1996)), we assume that since the dataset also included the LULC at a lower aggregation level, the results would give us better understanding of the urban pattern at both scales. Therefore, the remaining variables were retained in the model. This model was then tested for three time periods: 2000, 2001, and 2004. The results were contrasted for comparison. The confidence limit for these models was kept at 80 percent or higher. Each time period generated a number of association rules totaling to a support of more than 40 percent and similar to the earlier model, the confidence of 80 percent or higher was selected for the analysis. The results obtained from the comparison are discussed below:

- The rule which communicated most efficiently between the three temporal datasets was associated with the impervious surfaces. This also coincides with the derived information that there was little change in the impervious surfaces during the time period from 2000 to 2004
- The airport zoned areas exhibited a temperature range varying from class 6 to class 7 (refer to Table 2) in all the three images. The SNDVI over these classes ranged from class 1 to 3 (refer to Table 2). This relationship demonstrates the presence of heat around the airports was closely related to

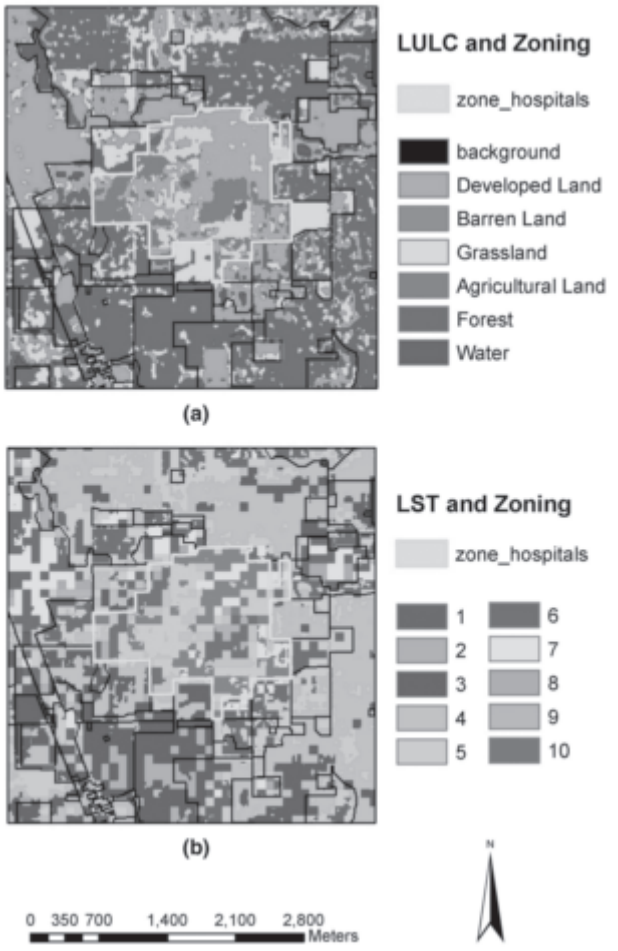

Figure 5. A sample area showing the relationship between (a) LULC hospital zoned area, and (b) LST hospital zoned area. A color version of this figure is available at the ASPRS website: **www.asprs.org**.

the lack of vegetation and this relative heat was constant throughout the year. The airports in Marion County did not experience much change in LULC over the study period (see Figure 8).

- A major part of the LULC type impervious areas were within the temperature range varying from class 6 to 7 (see Table 2) and for the grassland the temperature varied between class 4 and class 7 (see Table 2). Therefore, we can infer that the grassland around the city center exhibited more heat than the grassland at the rural areas contributing to the effect of urban heat islands. This variation in the temperature may be attributed to the type and density of grass, its growth, and the effect of other land-cover in its vicinity.
- The special-uses zoned areas was comprised of two different LULC classes; impervious surface areas with temperature varying from class 6 to 7, SNDVI ranging from class 2 to 4 and grasslands with temperatures varying from class 5 to 7, and the SNDVI ranging from class 9 to 10 (see Table 2).
- Some parts of water bodies located in low density housing, special uses and parks the demonstrated to have LST of range 1 to 2 (with majority of area around 1) within all three time periods. For the rest of the locations, the water bodies exhibited temperature of class range 2 to 3 (see table 2). There was not much clear distinction on visual analysis of the ASTER images. As discussed earlier this might be due to the varying water depth within these zones, but further site investigation is needed in order to verify.

The major commonalities between all the three time periods were restricted to the LULC impervious surfaces, and the results of the model did not demonstrate any strong association with the LULC types of barren land, agriculture, and forest land. The explanation for this is the seasonal changes within the images used. Regardless that the images were from different years, the major impact to the analysis was due to the varying seasons with regards to the images used. Irrespective of the temporal (both season and annual change) differences between the images, the certain rules remained constant for the given area. Since the UHI effect has demonstrated to behave in varying fashion over varying landscapes (Arnfield, 2003), it would be interesting to study the varying association derived over varying regions (such as equatorial, tropical, subtropical, etc.). The flexibility of this model to incorporate both GIS and a remote sensing database might be extended further by incorporating atmospheric variables (Grimmond, 2006) which would constitute a new study.

Conclusions

The association rule mining model discussed in this study demonstrated a new approach to extracting quantitative information about the relationships among the urban landscape parameters. Remote sensing datasets were created from ASTER images at three times for different years. The datasets were then processed to identify three variables of LULC, LST, and SNDVI. These temporal datasets were used in synergy with GIS datasets including population density, zoning, and transportation buffer zones to analyze their association. With minimum combination of simulations, the model predicted both interesting and evident rules. This study also demonstrates the application and extension of the association rule mining, which is conventionally used for market basket or tabular data structure to remote sensing datasets. Further research is needed along the direction of the MAUP and the selection of appropriate image classification techniques.

The rules obtained in this research are applicable only for Indianapolis, since the pixel-based approach used for rule generation has a strong dependence toward the spatial structure of the city. Nevertheless, the research has proved to be promising in the utilization of spatial data sets in the generation of association rules which could help researchers, planners, and environmental managers to understand better the spatial variables and their relationship in any given city (for instance, Mennis and Liu (2005), Mennis (2006) used similar technique to analyze urban socioeconomic and land-cover change in other cities). Given the dataset, the same model with minor or no modifications could be applied to any city or region for extracting information about the relationship between LST and its influencing parameters. A most important conclusion of this study is that the role of LULC in the UHI could be quantitatively measured and analyzed using the association model. The scope of this research could be extended further by implementing the same for varying combinations and also by using additional data sets and finding their association with respect to LST and UHI. It should be noted that in this study, only positive association rules were discussed since we were interested in finding the parameters which would aid in UHI study leaving the possibility of extracting negative association rules.

Acknowledgments

This research is supported by National Science Foundation (BCS-0521734) for a project entitled "Role of urban canopy composition and structure in determining heat islands: A

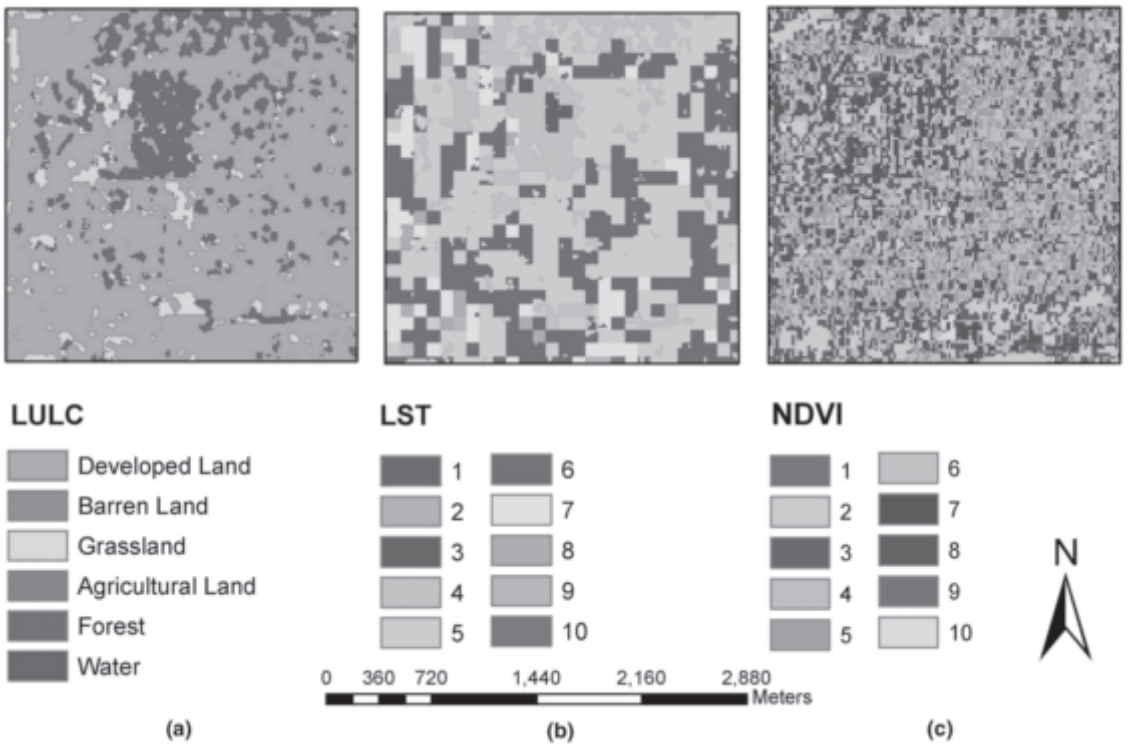

Figure 6. A sample area showing the relationship between (a) LULC, (b) SNDVI, and (c) LST. A color version of this figure is available at the ASPRS website: **www.asprs.org**.

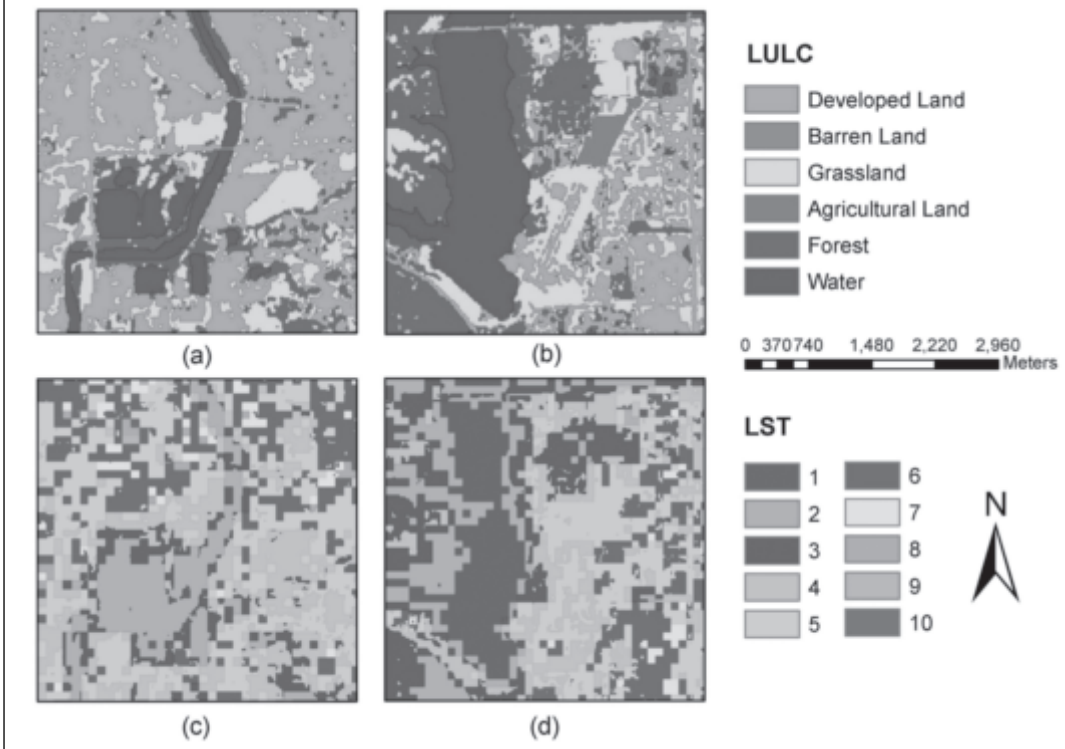

Figure 7. A sample area showing the relationship between LULC and LST around the water-bodies: (a) and (c) represent an area around the industrial zoning area, and (b) and (d) represent an area around low-density residential. A color version of this figure is available at the ASPRS website: **www.asprs.org**.

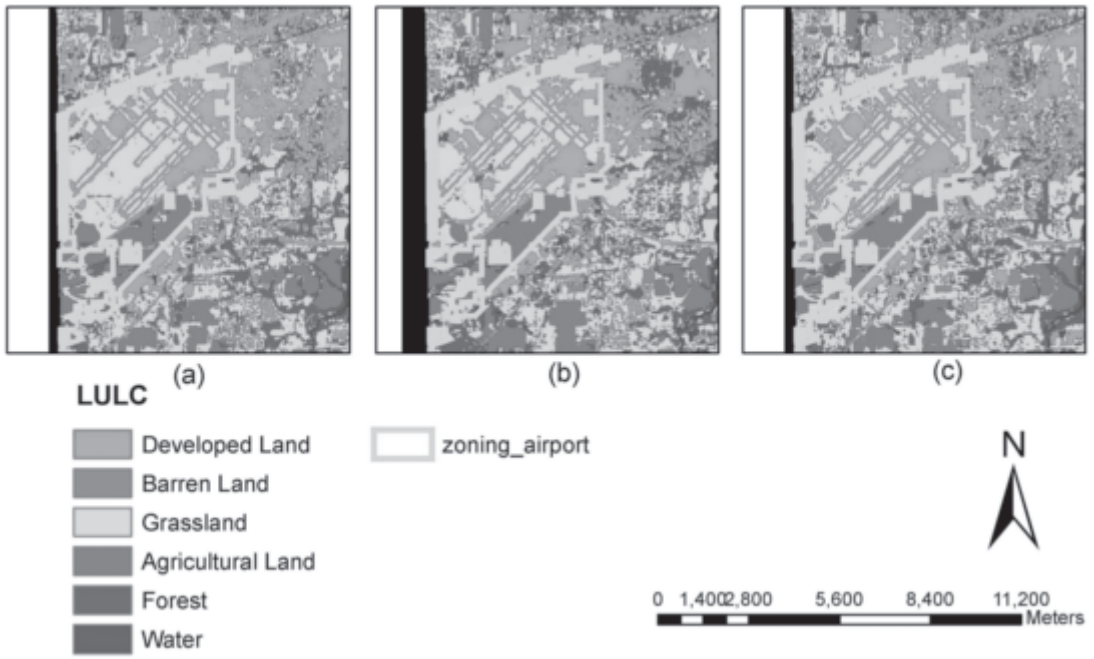

Figure 8. A sample area showing the LULC from the zoning type Airport: (a) 2000, (b) 2002, and (c) 2004. A color version of this figure is available at the ASPRS website: **www.asprs.org**.

synthesis of remote sensing and landscape ecology approach." Dr. Hua Liu assisted in ASTER image acquisition and processing for land-use and land-cover classification. We would also like to thank the three anonymous reviewers for their constructive comments and suggestions.

References

Agrawal, R., and R. Srikant, 1994. Fast Algorithms for mining association rules, *Proceedings of the International Conference of Very Large Databases*, pp. 487–499.

Arnfield, A.J., 2003. Two decades of urban climate research: a review of turbulence, exchanges of energy and water, and the urban heat island, *International Journal of Climatology*, 23:1–26.

Borgelt, C., and R. Kruse, 2002. *Graphical Models: Methods for Data Analysis and Mining*, John Wiley & Sons, Ltd., 358 p.

Cabena, P., 1998. *Discovering Data Mining from Concept to Implementation*, Prentice Hall, 224 p.

Chudnovsky, A., E. Ben-Dor, and H. Saaroni, 2004. Diurnal thermal behavior of selected objects using remote sensing measurements, *Energy & Buildings*, 36:1063–1074.

Dash, P., F.M. Göttsche, F.S. Olesen, and H. Fischer, 2002. Land surface temperature and emissivity estimation from passive sensor data: Theory and practice - Current trends, *International Journal of Remote Sensing*, 13:2563–2594.

Ding, Q., Q. Ding, and W. Perrizo, 2003. *Association Rule Mining on Remotely Sensed Image*s, Database Systems Users & Research Group.

Dousset, B., and F. Gourmelon, 2003. Satellite multi-sensor data analysis of urban surface temperatures and landcover, *ISPRS Journal of Photogrammetry and Remote Sensing*, 58:43–54.

Fast, J.D., J.C. Torcolini, and R. Redman, 2005. Pseudovertical temperature profiles and the urban heat island measured by a temperature datalogger network in Phoenix, Arizona, *American Meteorological Society*, 44:3–13.

Gallo, K.P., and T.W. Owen, 1999. Satellite-based adjustments for the urban heat island temperature bias, *Journal of Applied Meteorology*, 38:806–813.

Ghiaus, C., F. Allard, M. Santamouris, C. Georgakis, and F. Nicol, 2006. Urban environment influence on natural ventilation potential, *Building and Environment*, 41:395.

Gillies, R.R., W.P. Kustas, and K.S. Humes, 1997. A verification of the 'triangle' method for obtaining surface soil water content and energy fluxes from remote measurements of the Normalized Difference Vegetation Index (NDVI) and surface, *International Journal of Remote Sensing*, 18:3145–3166.

Golden, J.S., 2004. The built environment induced urban heat island effect in rapidly urbanizing arid regions - A sustainable urban engineering complexity, *Environmental Sciences*, 1:321–333.

Grimmond, C.S.B., 2006. Progress in measuring and observing the urban atmosphere, *Theoretical & Applied Climatology*, 84:3–22.

Hand, D., H. Mannila, and P. Smyth, 2001. *Principles of Data Mining, Volume 1*, A Bradford Book, The MIT Press, 578 p.

Hawkins, T.W., A.J. Brazel, W.L. Stefanov, W. Bigler, and E.M. Saffell, 2004. The role of rural variability in urban heat island determination for Phoenix, Arizona, *Journal of Applied Meteorology*, 43:476–486.

Liu, H., and Q. Weng, 2008. Seasonal variations in the relationship between landscape pattern and land surface temperature in Indianapolis, U.S.A, *Environmental Monitoring and Assessment*, 44(1–3):199–219.

Hung, T., D. Uchihama, S. Ochi, and Y. Yasuoka, 2006. Assessment with satellite data of the urban heat island effect in Asian mega cities, *International Journal of Applied Earth Observation and Geoinformation*, 8:34–48.

Jelinski, D.E., and J. Wu, 1996. The modifiable areal unit problem and implications for landscape ecology, *Landscape Ecology*, 11:129–140.

Jenerette, G.D., S.L. Harlan, A. Brazel, N. Jones, L. Larsen, and W.L. Stefanov, 2007. Regional relationships between surface temperature, vegetation, and human settlement in a rapidly urbanizing ecosystem, *Landscape Ecology*, 22:353–365.

Jung, A., P. Kardevan, and L. Tokei, 2005. Detection of urban effect on vegetation in a less built-up Hungarian city by hyperspectral remote sensing, *Physics & Chemistry of the Earth - Parts A/B/C*, 30:255–259.

Kafatos, M., X.S. Wang, Z. Li, R. Yang, and D. Ziskin, 1998. Information technology implementation for a distributed data system serving earth scientists: Seasonal to interannual ESIP, *Proceedings of SSDBM*.

Kalnay, E., and M. Cai, 2003. Impact of urbanization and land use change on climate, *Natur*e, 423:528–531.

Kato, S., and Y. Yamaguchi, 2005. Analysis of urban heat-island effect using ASTER and ETM+ Data: Separation of antropogenic heat discharge and natural heat radiation from sensible heat flux, *Remote Sensing of Environment*, 99:44–54.

Khaikine, M.N., I.N. Kuznetsova, E.N. Kadygrov, and E.A. Miller, 2006. Investigation of temporal-spatial parameters of an urban heat island on the basis of passive microwave remote sensing, *Theoretical and Applied Climatology*, 84:161–169.

Lo, C.P., D.A. Quattrochi, and J.C. Luvall, 1997. Application of high-resolution thermal infrared remote sensing and GIS to assess the urban heat island effect, *International Journal of Remote Sensing*, 18:287–304.

Lu, D., and Q. Weng, 2006. Spectral mixture analysis of ASTER images for examining the relationship between urban thermal features and biophysical descriptors in Indianapolis, Indiana, USA, *Remote Sensing of Environment*, 104:157–167.

Mennis, J., 2006. Socioeconomic-Vegetation relationships in urban, residential land: The case of Denver, Colorado, *Photogrammetric Engineering & Remote Sensin*g, 72(9):911–921.

Mennis, J., and J.W. Liu, 2005. Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change, *Transactions in GIS*, 9:5–17.

Miller, H.J., and J. Han, 2001. *Geographic Data Mining and Knowledge Discovery*, London and New York: Taylor & Francis, 372 p.

Openshaw, S., 1983. The modifiable areal unit problem, No. 38, *Concepts and Techniques in Modern Geography*, Norwich: Geo Books, p 43.

Quattrochi, D.A., N.S.N. Lam, H.L. Qiu, and W. Zhao, 1997. Image Characterization and Modeling System (ICAMS): A geographic information system for the characterization and modeling of multiscale remote sensing data, *Scale in Remote Sensing and GIS* (Dale A. Quattrochi and Michael F. Goodchild, editors), CRC Lewis, Boca Raton, Florida, pp. 295–307.

Shepherd, M.J., and S.J. Burian, 2003. Detection of urban-induced rainfall anomalies in a major coastal city, *Earth Interations*, 7:1–17.

Silverstein, C., S. Brin, and R. Motwani, 1997. Beyond market baskets: Generalizing association rules to dependence rules, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 255–264.

Smith, M.R., 1986. Comparing traditional methods for selecting class intervals on choropleth maps, *Association of American Geographers*, 38:62–67.

Snyder, W.C., Z. Wan, Y. Zhang, and Y.Z. Feng, 1998. Classification-based emissivity for land surface temperature measurement from space, *International Journal of Remote Sensing*, 19:2753–2774.

Stathopoulou, M., C. Cartalis, and I. Keramitsoglou, 2004. Mapping micro-urban heat islands using NOAA/AVHRR images and CORINE land cover: An application to coastal cities of Greece, *International Journal of Remote Sensing*, 25:2301–2316.

Streutker, D.R., 2003. Satellite-measured growth of the urban heat island of Houston, Texas, *Remote Sensing of Environment*, 85:282–289.

Tan, P.N., M. Steinbach, and V. Kumar, 2002. *Finding Spatio-Temporal Patterns in Earth Science Data*, NASA Grant Project, University of Minnesota.

Voogt, J.A., and T.R. Oke, 2003. Thermal remote sensing of urban cities, *Remote Sensing of Environment*, 86:370–384.

Weber, C., and J. Hirish, 1992. Some urban measurements from SPOT data: Urban life quality indices, *International Journal of Remote Sensing*, 13:3251–3261.

Weng, Q., 2001. A remote sensing-GIS evaluation of urban expansion and its impact on surface temperature in the Zhujiang Delta, China, *International Journal of Remote Sensing*, 22:1999–2014.

Weng, Q., 2003. Fractal analysis of satellite-detected urban heat island effect, *Photogrammetric Engineering & Remote Sensing*, 69(5):555–566.

Weng, Q., D. Lu, and J. Schubring, 2004. Estimation of land surface temperature vegetation abundance relationship for urban heat island studies, *Journal of Environmental Management*, 70:145–156.

Zhang, C., and S. Zhang, 2002. *Association Rule Mining: Models and Algorithms*, Springer-Verlag, New York, 238 p.

(Received 23 July 2007; accepted 12 December 2007; revised 18 January 2008)