ISEIS

Journal of Environmental Informatics 28(1) 1-10 (2016)

Journal of Environmental Informatics

www.iseis.org/jei

Mapping the Vulnerability of Asthmatic Allergy Prevalence Based on Environmental Characteristics through Fuzzy Spatial Association Rule Mining

F. Karimipour* and Y. Kanani-Sadat

Department of Surveying and Geomatics Engineering, College of Engineering, University of Tehran 1417614418, Iran

Received April 24, 2014; revised March 10, 2015; accepted March 15, 2015; published online August 15, 2016

ABSTRACT. The prevalence of allergic diseases has highly increased in recent decades due to contamination of the environment with the allergy stimuli. A common treat is identifying the allergy stimulus, and then avoiding the patient to be exposed with it. There are, however, many unknown allergic diseases stimuli that are related to the characteristics of the living environment. In this article, we focus on the effect of air pollution on asthmatic allergies and investigate the association between prevalence of such allergies with those characteristics of the environment that may affect the air pollution. This investigation, eventually, leads to map the vulnerability of asthmatic allergy prevalence based on environmental characteristics. For this, spatial association rule mining has been deployed to mine the association between spatial distribution of allergy prevalence and the air pollution parameters such as CO, SO_2, NO_2, PM_{10}, $PM_{2.5}$, and O_3 (compiled by the air pollution monitoring stations) as well as living distance to parks and roads. The categories of attributes have been defined as fuzzy sets in order to handle the data uncertainty. The results for the case study (i.e., Tehran metropolitan area) indicates that distance to parks and roads as well as CO, NO_2, PM_{10}, and $PM_{2.5}$ is related to the allergy prevalence in December (the most polluted month of the year in Tehran), while SO_2 and O_3 have no effect on that. In June, however, the distance to parks and roads as well as NO_2, PM_{10}, and $PM_{2.5}$ affect the allergy prevalence, but CO, SO_2 and O_3 are ineffective.

Keywords: fuzzy spatial association rule mining, vulnerability mapping, asthmatic allergy, air pollution, Apriori

1. Introduction

Prevalence of allergic diseases has highly increased in recent decades, especially among children, due to modern living conditions resulted in contamination of the environment with the allergy stimuli, called allergen (Zöllner et al., 2005; Ng et al., 2009). Allergic patients have hypersensitive immune systems that abnormally react to harmless substances. Several factors cause allergic reactions, which depend on the gene, living style and habits, foods, as well as the geography and conditions of the environment (Asher et al., 1995).

A common treat to allergic diseases is identifying the allergen, and then avoiding the patient to be exposed with it (Douglass and O'Hehir, 2006). There are, however, several unknown stimuli that may cause allergic diseases, many of which are related to the characteristics of the living environment. Therefore, analyzing the data collected about the living environment of allergic patients may lead to identifying the role of environmental parameters in prevalence of allergies. As the patients are distributed in the space, the spatial data mining techniques seems very efficient in this regards (Mohan, 2014). Especially, association rule mining is capable to extract the associations between allergic asthma and environmental parameters. This article deploys the spatial data mining, and especially association rule mining to investigate the relation between prevalence of asthmatic allergies with characteristics of the environment.

1.1. Spatial Data Mining

Spatial data mining concerns development and application of novel computational techniques to analyze very large spatial databases (Koperski et al., 1996; Buttenfield et al., 2001; Mennis and Guo, 2009; Yadav and Rizvi, 2014). A major distinction of spatial data mining is that attributes of the neighboring objects influence each other and thus must be taken in to account. Furthermore, the location and extension of spatial objects define implicit relations of spatial neighborhoods (such as topological, distance and directional relations), which are used by spatial data mining algorithms (Karimipour et al., 2005; Miller and Han, 2009).

Data mining techniques is a common approach to study prevalence of allergic disease. Ng et al. (2009) used data mining techniques to predict allergy symptoms among children in Taiwan. They used the allergy data of children under the age of 12 and considered 30 predictor variables including personal factors, health behavior factors, living condition factors, family factors, and allergy-inducing factors. Akinbami et

* Corresponding author. Tel: +98 2161114376; Fax: +98 2188008837.
E-mail address: fkarimipr@ut.ac.ir (F. Karimipour).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.201600338

al. (2010) assessed the association between chronic outdoor air pollution exposure and childhood asthma in metropolitan areas across the US. They compiled 12-month average air pollutant levels for SO_2, NO_2, O_3 and PM and linked eligible children to pollutant levels for the previous 12 months for their county of residence. Finally, logistic regression models were used to estimate asthma attack. YoussefAgha et al. (2013) studied the application of data mining techniques to predict allergy outbreaks among elementary school children. They used the binary logistic regression to determine if there is any relation between prevalence of allergies among elementary school children and daily upper-air observations (i.e., temperature, relative humidity, dew point, and mixing ratio) and daily air pollution (CO, SO_2, NO_2, PM_{10}, $PM_{2.5}$ and O_3). Gasana et al. (2012) conducted a meta-analysis to clarify the potential relationship between motor vehicle emissions and the development of childhood asthma. They concluded that living or attending school near high traffic density roads exposes children to higher levels of motor vehicle air pollutants, and that this increases the incidence and prevalence of childhood asthma and wheezing. Ayres-Sampaio et al. (2014) evaluated the relationship between asthma hospital admissions and several environmental variables in mainland Portugal using spatial data from remote sensing and spatial modeling. Their results suggest that asthmatic people living in highly urbanized and sparsely vegetated areas are at a greater risk of suffering severe asthma attacks that lead to hospital admissions. The results of all of these researches are plausible. Nevertheless, none of them considered spatio-temporal characteristics of data to study prevalence of allergy.

1.2. Spatial Association Rule Mining

Spatial association rule mining is a class of spatial data mining techniques. It seeks interesting association or correlation relationships among a large set of data items, i.e., certain data items that often occur together (Agrawal et al., 1993; Han et al., 2011). An association rule is an implication of the form A → B where A (the antecedent) and B (the consequent) are sets of predicates. For example, in a supermarket transactions database, this process analyzes customer buying habits through seeking associations among different items bought by customers and discovers rules like "older than 50 years customers that purchase cheese, also purchase milk", which is expressed as:

$$age\ (X, greaterThan\ 50)\ \wedge\ purchase\ (X, cheese) \rightarrow purchase\ (X, milk) \quad (1)$$

The association rule mining has been proposed as a non-model based (rule-based) prediction method (Kamei et al., 2008). Unlike the regression models which require the dependent and independent variables to be clearly defined and consequently force the analyst to priorly well-defines the questions and hypotheses the association rule mining can extract relations among the variables for which no questions have been formulated. On the other hand, association rule mining can provide the relation among multiple variables, which is not always possible through the conventional statistical methods (Nembhard et al., 2012).

A spatial association rule contains at least one spatial relationship in an antecedent or consequent predicate (Koperski and Han, 1995). For example distance_to (police_station, between 0 to 100 m) is a spatial predicate that results in a spatial association rule. There are two important issues in dealing with spatial association rules: (1) Unlike non-spatial association rules – which are explicitly encoded transactions – spatial relationships are typically embedded within the spatial framework of the geo-referenced data. Therefore, the seeking patterns are implicit and the "spatial relationships must be extracted from the data prior to the actual association rule mining" (Shekhar and Chawla, 2003). Nevertheless, pre-pro cessing and storing all combinations of the relations among massive volume of spatial data is not practically possible. Therefore, there must be a trade-off between pre- and on-demand processing of spatial relationships among geographic objects (Klosgen and May, 2002). (2) Spatial predicates usually contain numeric data (e.g. metric distance), while the conventional association rule mining can only deal with categorical (classified) data. A solution to this problem is that we, first classify numeric data into ordinal categories and then mine these ordinal data for association rules (Piatetsky-Shapiro, 1991; Srikant and Agrawal, 1996). For example, metric distance may be categorized into ‘very near’, ‘near’, ‘medium, and ‘far’.

Discovering association rules from data stored in spatial databases has been considered in many researches. Mennis and Liu (2003) explored the spatio-temporal association rules among a set of variables characterizing the socioeconomic and land cover changes in Denver, Colorado region from 1970 to 1990. Shua et al. (2008) used Apriori algorithm to produce association rules in vegetation and climate changing data of northeastern China. Ladner et al. (2003) studied the correlations of spatially related data such as soil types, directional and geometric relationships. They combined spatial and fuzzy data mining to handle the spatial uncertainty of data. Finally, Calargun and Yazici (2008) analyzed the real meteorological data for Turkey recorded between 1970 and 2007 using spatio-temporal data cube and Apriori algorithm in order to generate fuzzy association rules. The results of the two approaches were then compared according to interpretability, precision, utility, novelty, direct-to-the-point, performance and visualization. They also visualized the association rules based on their significance and support values in order to provide a complete analysis tool for a decision support system in meteorology domain.

In this article, we focus on the effect of air pollution on asthmatic allergies and deploy the association rule mining to investigate the relation between prevalence of such allergies with those characteristics of the environment that may affect the air pollution. The places of residence of a group of asthmatic allergic patients, live in Tehran metropolitan area, as

well as spatial characteristics of the environment (e.g., location of parks, roads and air pollution monitoring stations) were placed on the map. We then deployed spatial association rule mining (as one of the spatial data mining analyses) to extract the association between asthmatic allergy prevalence and the air pollution parameters such as CO (carbon monoxide), SO_2 (sulfur dioxide), NO_2 (nitrogen dioxide), PM_{10} and $PM_{2.5}$ (particulate matter with a diameter of < 10 μm and < 2.5 μm, respectively), and O_3 (ozone) as well as living distance to parks and roads, as major sources of asthmatic allergens. The fuzzy multi-dimensional spatial association rule mining was deployed, in order to investigate many parameters to be examined, and to handle the uncertainty exists in the attributes linked to the spatial data. Finally, the results (i.e., the discovered association rules between prevalence of asthmatic allergies and the characteristics of the environment) were used to map the vulnerability of asthmatic allergy prevalence in the study area based on environmental characteristics.

The rest of the article is organized as follow: Sections 2 describes the components of the research methodology in details. The results for the case study are presented, discussed and evaluated in Section 3. Finally, Section 4 contains concluding remarks and ideas for future research in this direction.

2. Research Methodology

This article maps the vulnerability of asthmatic allergy prevalence based on environmental characteristics through deploying the fuzzy spatial association rule mining to extract the association between prevalence of asthmatic allergies with those characteristics of the environment that may affect the air pollution. Figure 1 illustrates the research methodology:

2.1. Data Pre-processing

This research investigates the relation between spatial distribution of allergy prevalence and the air pollution parameters as well as living distance to parks and roads. The concentration of the studied air pollution parameters (i.e., CO, SO_2, NO_2, PM_{10}, $PM_{2.5}$, and O_3) observed at the monitoring stations are used to produce the distribution maps of these parameters. To model the effect of distance to roads, a map is needed in which each point is assigned the distance to its nearest road. The same process is applied to model the effect of parks using the following equation, which quantifies the effect of nearby parks:

$$T_j = \sum \frac{A_i}{d_{ij}^2} \quad (2)$$

Where T_j is the effect of nearby parks for the point j, A_i is the area of the park i, and d_{ij} is the distance of the park i from the point j.

Finally, the places of residence of the sample patients are placed on the map. For each patient, a data item is stored that shows if he/she has asthmatic allergy. Moreover, having overlaid this map with the air pollution, the "effect of parks" and "distance to roads" maps, the air pollution parameters, effect of parks and distance to roads are assigned to each point as data items (attributes).

2.2. Data Conceptualization

The inputs of association rule mining must be categorical values. Therefore, the data items assigned to the patients must be categorized. Furthermore, in order to deal with sharp break

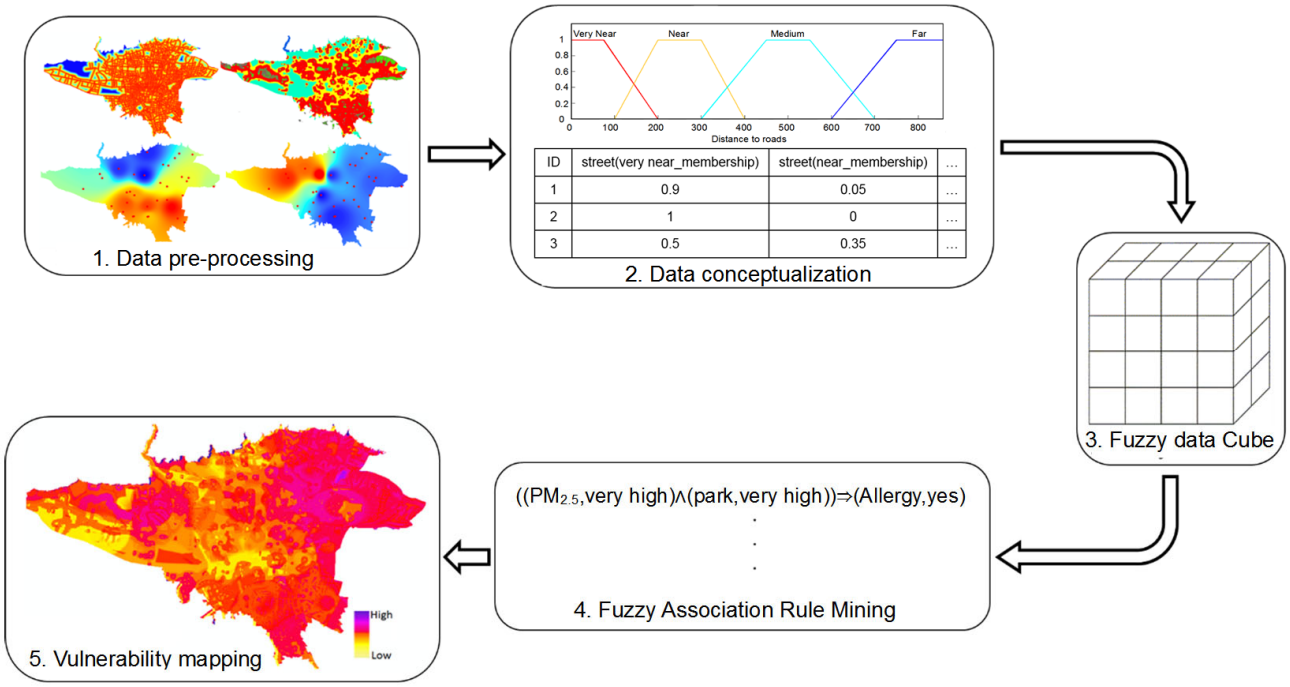

Figure 1. Research methodology.

points between categories, fuzzy labels are assigned to data items.

To categorize the air pollution parameters, the air quality index (AQI) is used. As the categorization breakpoints used by AQI varies from an air pollution parameter to another (Table 1), the following equation is used to normalize the measured values (Mintz, 2012):

$$I_P = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}}\left(C_P - BP_{Lo}\right) + I_{Lo} \quad (3)$$

where

I_p = the air quality index for the air pollution parameter p

C_p = the value measured for the air pollution parameter p

BP_{Hi} = the first break point greater than C_p

BP_{Lo} = the first break point less than C_p

I_{Hi} = the air quality index for BP_{Hi}

I_{Lo} = the air quality index for BP_{Lo}

We merge the above air pollution categories to "very high", "high", "moderate", and "low" using the relevant membership functions. The membership function illustrated in Figure 2.a is also used to classify the distance to roads into "very near", "near", "medium" and "far". The same process classifies the effect of parks into "very highly affected", "highly affected", "moderately affected" and "lowly affected" (Figure 2.b).

2.3. Fuzzy Data Cube Construction

Multidimensional data mining searches for interesting patterns by exploring the data in multidimensional space. Using of the data cube and a multidimensional data model provides flexible access to summarized data and facilitates the processing of multidimensional data. The multidimensional data cube is a common organization form of data for data mining in data warehouse structures (Han et al., 2011; Ladner et al., 2003). An *n*-dimensional data cube is an *n*-dimensional database where each dimension illustrates an attribute, and the "aggregate measure", is computed for each possible combination of the dimensions. (Han et al., 2011; Wang, 2010). Figure 3 shows a 3D data cube for dimensions A, B, and C, and the aggregate measure *count* for data presented in Table 2. Each cell in the data cube stores the number of tuples that have the corresponding attribute values. For example, there are three rows in Table 2 (#5, #12, and #15) whose values are (a_0, b_1, c_0), thus the cell $a_0b_1c_0$ in the data cube will be assigned 3.

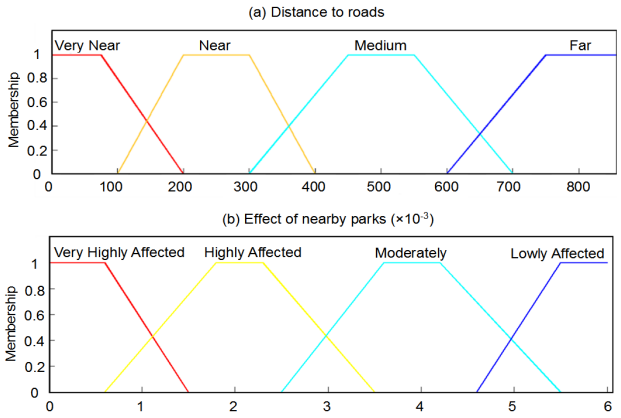

Figure 2. The membership functions used to classify (a) distance to roads and (b) effect of nearby parks.

Table 2. Tuples Value in Dimensions A, B and C

DI	A	B	C
1	a_0	b_1	c_1
2	a_2	b_0	c_0
3	a_1	b_2	c_2
4	a_3	b_0	c_0
5	a_0	b_1	c_0
6	a_0	b_1	c_2
7	a_1	b_2	c_1
8	a_3	b_0	c_2
9	a_0	b_2	c_0
10	a_3	b_2	c_1
11	a_2	b_0	c_0
12	a_0	b_1	c_0
13	a_1	b_2	c_1
14	a_3	b_0	c_2
15	a_0	b_1	c_0

Table 1. Breakpoints for the AQI

Category	AQI	Breakpoints					
		NO_2 (ppb)	SO_2 (ppb)	CO (ppm)	$PM_{2.5}$ (µg/m^3)	PM_{10} (µg/m^3)	O_3 (ppm)
Good	0 - 50	0 - 53	0 - 35	0.0 - 4.4	0.0 - 15.4	0 - 54	0.000 - 0.059
Moderate	51 - 100	54 - 100	36 - 75	4.5 - 9.4	15.5 - 40.4	55 - 154	0.060 - 0.075
Unhealthy for Sensitive Groups	101 - 150	101 - 360	76 - 185	9.5 - 12.4	40.5 - 65.4	155 - 254	0.076 - 0.095
Unhealthy	151 - 200	361 - 649	186 - 304	12.5 - 15.4	65.5 - 150.4	255 - 354	0.096 - 0.115
Very Unhealthy	201 - 300	650 - 1249	305 - 604	15.5 - 30.4	150.5 - 250.4	355 - 424	0.116 - 0.374
Hazardous	301 - 500	1250 - 2049	605 - 1004	30.5 - 50.4	250.5 - 500.4	425 - 604	-

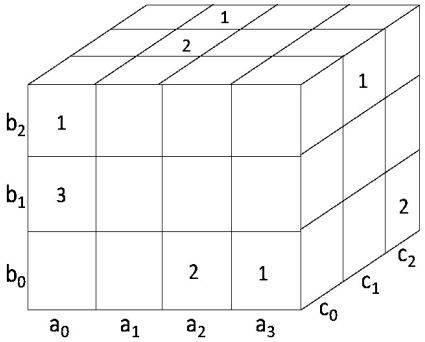

Figure 3. The 3D data cube constructed for Table 2.

In order to flexibly access to the data and facilitate the association rule mining process, we use data cube to organize data after the data conceptualization step. Here, we deal with the data cube as a fuzzy object in order to optimally discover the knowledge (Ladner et al., 2003). As the data items are defined as fuzzy sets, instead of the aggregate measure *count*, each cell contains the sum of the minimum of the membership values of the corresponding fuzzy labels.

2.4. Fuzzy Spatial Association Rule Mining

The fuzzy association rule mining utilize fuzzy sets to mine the association rules in a given attribute dataset, which provides more reliable associations rules (Intan, 2007; Intan et al., 2009). A membership function defined for a fuzzy set is used to assign fuzzy values to each member (attribute).

Having the fuzzy data cube constructed, the association rules between allergy prevalence and spatial characteristics of the environment (i.e., air pollution and distance to parks and roads) are extracted, along their supports and confidences. As we are interested in antecedents that result in allergy, we only keep those rules whose consequence is "(*allergy, yes*)", such as:

[($PM_{2.5}$,*very high*), (*park_efct, very high*)] → (*allergy*, *yes*) (4)

2.5. Vulnerability Mapping

The extracted rules, which associate the asthmatic allergy prevalence to spatial characteristics of the environment, are now used for spatial modeling of the vulnerability of asthmatic allergy prevalence based on environment characteristics. For this, the GIS-fuzzy integration is used as follows (Dodge et al., 2008; Ross, 2009).

The support and confidence thresholds are set to zero in order to collect all the association rules (no matter how much supportive and confident they are). The Kulczynski correlation factor of each rule is normalized to [0, 100] and fuzzified using the function shown in Figure 4. Table 3 illustrates three example association rules along their corresponding fuzzy *if-then* rules and Kulczynski categories.

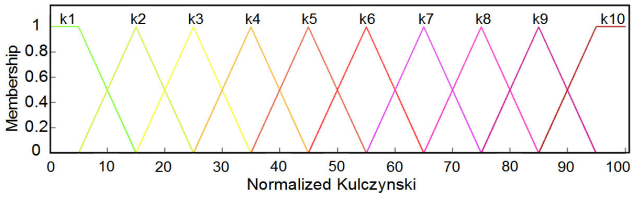

Figure 4. The function to fuzzify the Kulczynski correlation factors.

Fuzzy inference is the process of mapping a given input to an output using fuzzy logic. The Mamdani fuzzy inference method (Akgun et al., 2012; Mamdani and Assilian, 1975) is utilized to calculate the risk of asthmatic allergy prevalence. Having defined the fuzzy rules, the maps of those air pollutants that affect allergic asthma, induced by our association rule mining, as well as the maps of park effects and distance to road are combined through the fuzzy inference system (Figure 5) to compute the Kulczynski relation factor for each point of the city, from which the risk map of allergic asthma prevalence is produced.

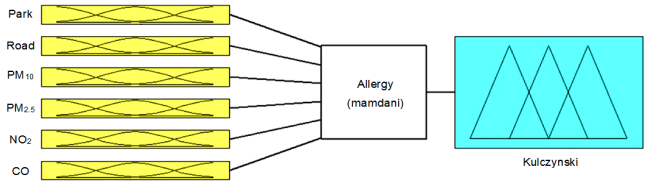

Figure 5. Combining the fuzzy rules to compute the Kulczynski relation factor for each point.

3. Results and Discussion

This section presents the outputs of applying the procedure described in Section 2 to Tehran metropolitan area, a case study, and discusses and evaluates the results. In order to study the effect of time in the results, the process is separately performed on data collected in June and December.

Table 3. Three of the Association Rules and the Corresponding Fuzzy Rules

	ID	Rule	Normalized Kulc
Association rules	1	[(park_efct, high)] → (allergy, yes)	39.35
	2	[($PM_{2.5}$, moderate), (CO, low)] → (allergy, yes)	6.73
	3	[(NO_2, very high), (CO, very high), (park_efct, very high)] → (allergy, yes)	97.94
Fuzzy rules	1	IF park_efct is high THEN Kulc is k4	
	2	IF $PM_{2.5}$ is moderate AND CO is low THEN Kulc is k1	
	3	IF NO_2 is very high AND CO is very high AND park_efct is very high THEN Kulc is k10	

The air pollution parameters consist of CO, SO_2, NO_2, PM_{10}, $PM_{2.5}$, and O_3 compiled hourly in June and December 2013 by Tehran's air pollution monitoring stations are used (Figure 6). This data is cleaned by filling the gaps (through interpolation) and filtering the noises. To reduce this voluminous data to monthly air pollution parameters, the monthly average of maximum values observed for each parameter in a day is calculated. These values are used to produce a monthly pollution map for each air pollution parameter through Kriging spatial interpolation (Wackernagel, 2003). Figure 7 illustrates the pollution maps of December 2013.

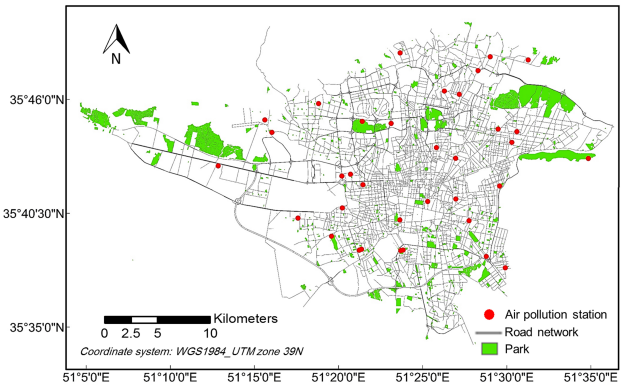

Figure 6. Tehran's roads, parks and air pollution monitoring stations.

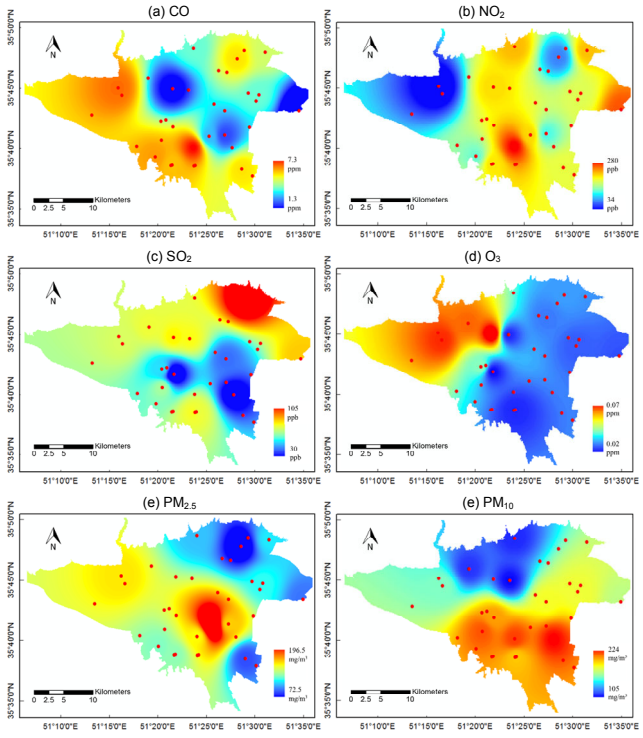

Figure 7. Air Pollution map of December 2013 for air pollution parameters.

The maps of the effect of distance to roads and parks were produced (Figure 8) based on the procedure described in Section 2.1 using ArcGIS 10.1.

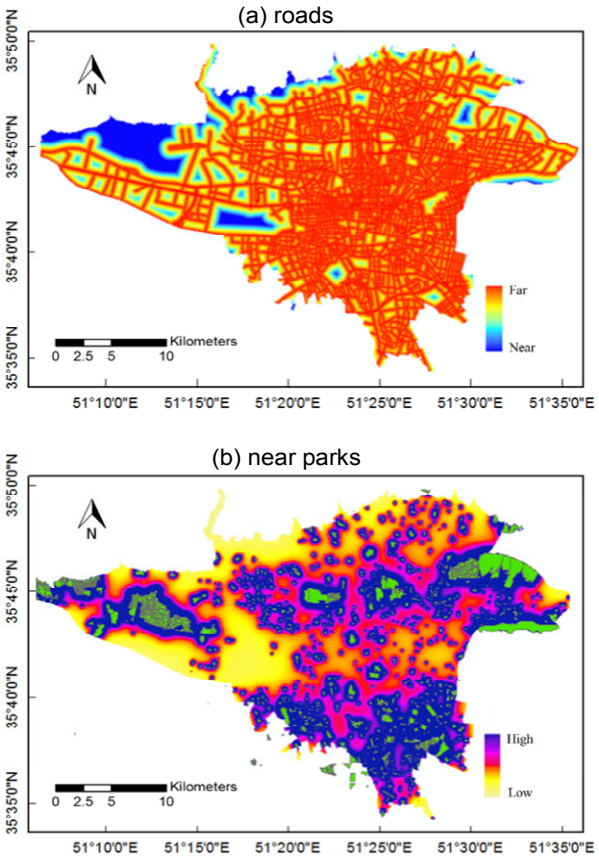

Figure 8. The maps classified the effect of distance to (a) roads and (b) nearby parks.

Finally, the places of residence of 1000 patients referred to the "Tehran Children's Medical Clinic" in June 2013 and 1000 patients in December 2013 are placed on the map; and the air pollution parameters, effect of parks and distance to roads are assigned to each point as data items (attributes). The extracted attributes for some example people is shown in Table 4.

Based on the defined membership functions, the attribute table values will be described with fuzzy labels and corresponding membership value. For example the membership value of Table 4 tuples in “distance to road” categories is shown in Table 5.

Then for our case study data, we constructed two 9-D data cubes for June and December 2013 whose dimensions are spatial characteristics of the residence location of the patients (i.e. air pollution and distance to parks and roads), as well as the allergy status. The dimensions and their categories are shown in Table 6.

As an example, the value for the data cube cell [park_effect(very high), distance_to_road(very near), O_3(low), PM_{10}(high), SO_2(low), $PM_{2.5}$(very high), NO_2(very high), CO(high), Allergy(yes)] is calculated in Table 7 based on the data presented in Table 4. Applying the same process to all data cube cell provides us with the final data cube.

Table 5. Membership Values for "Distance to Road" Category

ID	Crisp value for distance to road	Fuzzy sets membership value			
		Very near	Near	Medium	Far
1	101.98	0.78	0.02	0	0
2	480.10	0	0	1	0
3	180.27	0.16	0.80	0	0
4	90	0.88	0	0	0
5	278.92	0	1	0	0

Table 6. Data Cube Dimensions and Their Categories

Dimension	Categories
Park_effect	"very highly", "highly", "moderately" and "lowly" affected
Distance_to_road	"very near", "near", "medium" and "far"
O_3	"very high", "high", "moderate" and "low"
PM_{10}	"very high", "high", "moderate" and "low"
SO_2	"very high", "high", "moderate" and "low"
$PM_{2.5}$	"very high", "high", "moderate" and "low"
NO_2	"very high", "high", "moderate" and "low"
CO	"very high", "high", "moderate" and "low"
Allergy	"yes" and "no"

As the air pollution parameters varies from time to time, the data cubes constructed for June and December 2013 are separately involved in the rule mining procedure expecting that different rules are extracted. Applying the rule mining procedure to the dataset of December 2013, provided 60 association rules between prevalence of allergy with characteristics of the environment, some of which are illustrated in Table 8.

In order to determine if a rule is significant, reliable and interesting, the concepts of *support* and *confidence* are used. The support is the probability of an item in the database satisfying the set of predicates contained in both the antecedent and consequent; and the confidence is the probability that an item that contains the antecedent also contains the consequent:

$$support\ (A \rightarrow B) = \ prob\{A \cup B\} \tag{5}$$

$$confidence\ (A \rightarrow B) = \ prob\{B|A\} = \frac{prob\{A \cup B\}}{prob\{A\}} \tag{6}$$

The association rules that have the minimum significant support and confidence are called strong association rules and are considered in decision making process (Agrawal and Srikant, 1994).

In our case, the minimum support and confidence thresholds are respectively defined as 5% and 30%. For example, rule #5 in Table 8 with 6.55% support, and 75.52% confidence says that 6.55% of the statistical population lives in locations where the amount of NO_2 and $PM_{2.5}$, and the effect of nearby parks are very high and are suffering from asthmatic allergy; and this is 75.52% of the statistical population who live in such areas;

On the other hand, to reliably eliminate the weak associations, *correlation* factor is defined to measure the degree of relation between *A* and *B* (Han et al., 2011). Therefore, the extracted rules are evaluated as:

$$A \rightarrow B\ [support, confidence, correlation] \tag{7}$$

The *Kulczynski*, a measure to evaluate the correlation, is defined as (Kulczynski, 1927):

$$Kul\ (A, B) = \frac{1}{2}\left(P(A|B) + P(B|A)\right) \tag{8}$$

which is a value between 0 and 1. A larger *Kulc* indicates stronger relation between *A* and *B*. For example, The *Kulczynski*'s correlation measure between the antecedent and the consequence in the rule #5 of Table 8 is 64%.

Table 4. Extracted Attributes for some Example People

ID	Park effect	Distance to road	O_3	PM_{10}	SO_2	PM_{25}	NO_2	CO	Allergy
1	0.0003	101.98	27.86	113.28	61.30	201.45	81.59	59.08	no
2	0.0011	480.10	29.37	102	114.37	157.24	53.40	84.88	no
3	0.0049	180.27	29	99.79	113.84	157.03	53.05	84.98	no
4	0.0062	90	26.69	119.04	59.31	207.16	101.9	73.46	yes
5	0.0002	278.92	44.75	75.56	87.28	176.38	103.11	66.94	no

Table 7. Calculation of One Cell of Fuzzy Data Cube

ID	Park_effect Very high	Distance to road Very near	O_3 low	PM_{10} high	SO_2 low	PM_{25} Very high	NO_2 Very high	CO high	Allergy yes	Minimum membership
1	0	0.78	0.05	1	0.95	1	0.04	0	0	0
2	0	0	0	0	0	0	0	0.06	0	0
3	0.2	0.16	0	0	0	0	0	0.06	0	0
4	1	0.88	0.43	0.31	1	1	1	0.96	1	0.31
5	0	0	0	0	0	0	1	0	0	0
Sum										0.31

Table 8. Some of the Rules Extracted for December through Association Rule Mining

ID	Association rules	Sup	Conf	Kulc
1	[($PM_{2.5}$, very high)] → (allergy, yes)	14.61	33.52	0.63
2	[($PM_{2.5}$, very high), (park_efct, very high)] → (allergy, yes)	9.35	55.43	0.61
3	[($PM_{2.5}$, very high), (PM_{10}, very high)] → (allergy, yes)	9.47	51.62	0.60
4	[($PM_{2.5}$, very high), (PM_{10}, very high), (park_efct, very high)] → (allergy, yes)	6.15	72.60	0.62
5	[($PM_{2.5}$, very high), (NO_2, very high), (park_efct, very high)] → (allergy, yes)	6.55	75.52	0.64
6	[(NO_2, very high), (CO, very high), (park_efct, very high)] → (allergy, yes)	5.32	81.63	0.64
7	[(PM_{10}, very high), (NO_2, very high), (CO, very high)] → (allergy, yes)	7.25	68.55	0.62
8	[($PM_{2.5}$, very high), (PM_{10}, very high), (NO_2, very high), (park_efct, very high)] → (allergy, yes)	5.84	78.64	0.64
9	[($PM_{2.5}$, very high), (PM_{10}, very high), (NO_2, very high), (CO, very high)] → (allergy, yes)	6.69	71.33	0.62
10	[(PM_{10}, very high), (NO_2, very high), (CO, very high), (road, very near)] → (allergy, yes)	5.39	71.84	0.59
11	[($PM_{2.5}$, very high), (PM_{10}, very high), (NO_2, very high), (CO, very high), (road, very near)] → (allergy, yes)	5.39	73.00	0.60

Table 9. Some of the Rules Extracted for June through Association Rule Mining

ID	Association rules	Sup	Conf	Kulc
1	[($PM_{2.5}$, very high)] → (allergy, yes)	15.38	52.38	0.69
2	[(park_efct, very high)] → (allergy, yes)	21.22	59.56	0.64
3	[(NO_2, very high)] → (allergy, yes)	16.19	52.51	0.62
4	[($PM_{2.5}$, very high), (PM_{10}, very high)] → (allergy, yes)	12.21	75.00	0.62
5	[(PM_{10}, very high), (NO_2, very high)] → (allergy, yes)	10.27	71.95	0.61
6	[(PM_{10}, very high), (park_efct, very high)] → (allergy, yes)	13.79	74.39	0.66
7	[($PM_{2.5}$, very high), (PM_{10}, very high), (park_efct, very high)] → (allergy, yes)	8.60	83.64	0.68
8	[(PM_{10}, very high), (NO_2, very high), (park_efct, very high)] → (allergy, yes)	7.55	81.36	0.67
9	[($PM_{2.5}$, very high), (PM_{10}, very high), (road, very near)] → (allergy, yes)	9.66	80.00	0.68
10	[($PM_{2.5}$, very high), (PM_{10}, very high), (NO_2, very high), (park_efct, very high)] → (allergy, yes)	7.04	86.64	0.74

Based on the extracted rules, distance to parks and roads as well as CO, NO_2, PM_{10} and $PM_{2.5}$ affect allergy prevalence in December, while SO_2 and O_3 has no significant relation. On the other hand, the rules that include "*(park_efct, very high)*" and for which at least one of the air pollution parameters is high (e.g., rules #2, #4 and #6) has greater confidences compare to those that only have one of these components (e.g., rules #1, #3 and #7). The research on air pollution and asthmatic allergy certifies this: Air pollution may itself outbreak the allergy, but it also facilitates the pollen to get into the respiratory system (Bartra et al., 2007). On the other hand, the rules #10 and #11, which contain "(road, very near)" has no significant increase in confidence as the effect of this parameter already manifested in increase of air pollution parameters.

For June, 42 rules were extracted (Table 9). The rules show that the effect of distance to parks has been increased, which seems true because most plants pollen in spring. On the other hand, despite December, the amount of CO is ineffective in June. It is interpreted as the amount of CO is often high in December due to inversion phenomena in Tehran, but the AQI of CO in June is always "good" and therefore this parameter does not significantly affect the allergy.

Finally, the process described in subsection 2.5 provided 2148 and 1271 fuzzy rules for June and December, respectively. Figure 9 illustrates the vulnerability maps of asthmatic allergy prevalence based on environmental characteristics in Tehran for June and December. As expected from the extracted rules (Tables 8 and 9) the effect of distance to parks have significantly affected the vulnerability maps.

In order to evaluate the created vulnerability maps, they are first classified into three classes of high, moderate and low

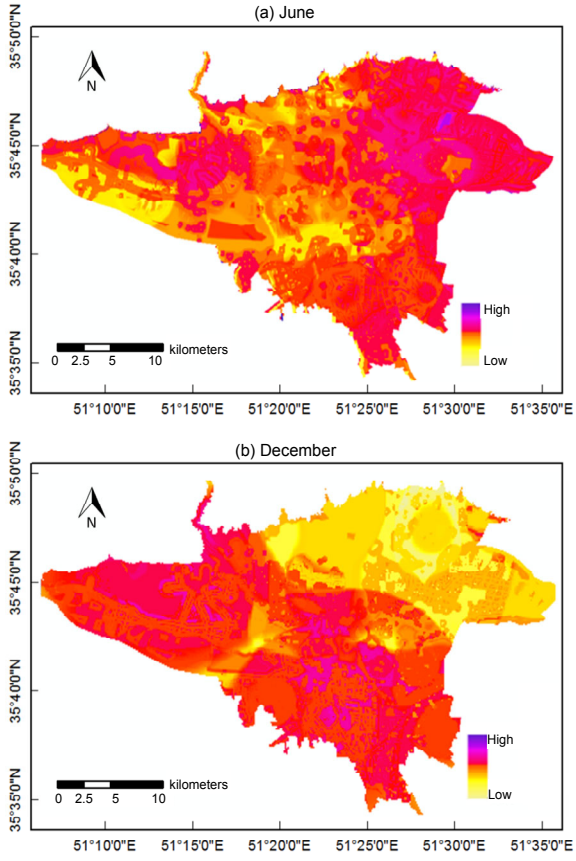

Figure 9. The vulnerability maps of asthmatic allergy prevalence based on environmental characteristics in Tehran for (a) June and (b) December.

risk areas. Then, for each time epoch (i.e., June and December 2013), the places of residence of 100 asthmatic allergy patients who were not involved in the rule mining and risk mapping processes (called check patients) were overlaid on the corresponding maps (Figure 10); and the assigned classes by the map were determined and counted, which confirm our vulnerability maps: as Figure 10 indicates, the number of people suffering from asthmatic allergy is significantly more in high risk areas (69 for June; 75 for December) compare to middle (20 for June; 18 for December) and low (11 for June; 7 for December) risk areas.

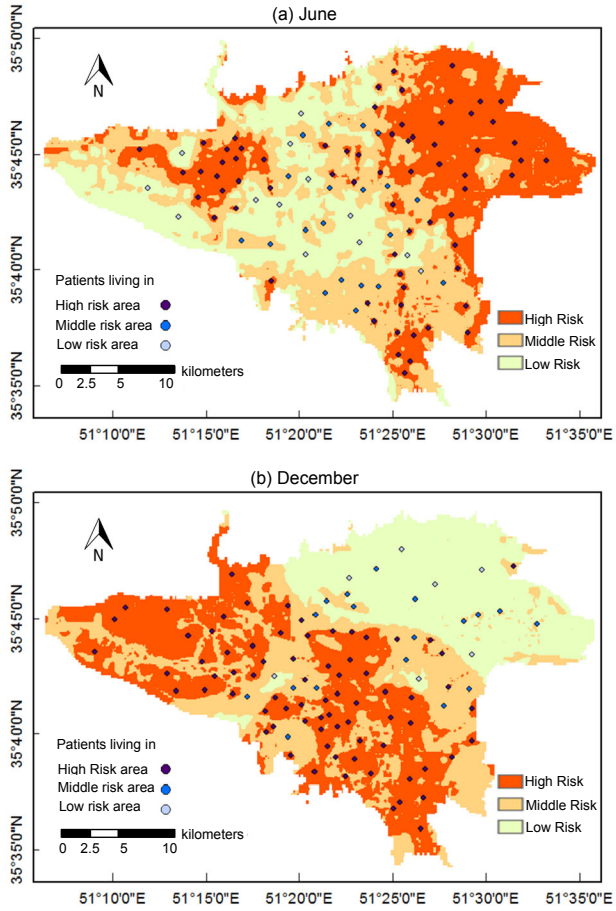

Figure 10. The place of residence of the 100 check patients overlaid on the classified vulnerability maps for (a) June and (b) December.

4. Conclusions

This article deploys the fuzzy spatial association rule mining to investigate the relation between prevalence of asthmatic allergies and those characteristics of the environment that may affect the air pollution, through which maps the vulnerability of asthmatic allergy prevalence based on environmental characteristics. The results for the case study (i.e., Tehran metropolitan area) shows that considering spatial distribution of the patients as well as fuzzy definition of data items (i.e., attributes) enabled to extract more reliable associations, as their interpretation certifies. Furthermore, the rules extracted for two different months (i.e., June and December), for which the air pollution conditions are different in Tehran, showed these relations are not static, as the air pollution parameter and pollen varies with time. Finally, the visualized vulnerability map of Tehran could help to avoid asthmatic allergic patients to be exposed with the allergy stimuli.

This paper focuses on the hypothesis that "there is a relation between environmental parameters and allergy prevalence and such relations can be extracted using the data mining techniques". For this, we only consider distance to parks and roads as parameters that may affect the air pollution and asthmatic allergies. In future, other characteristics of the environment (e.g., buildings, elevation, wind direction, etc.) as well as more accurate data (e.g., vegetation type, allergy type, etc.) will be taken into account in order to achieve more reliable results. For instance, the hourly data provided by the air pollution stations was integrated to only one AQI (air quality index) for each parameter at each month, as we only have the monthly allergy prevalence data; thus finer air pollution data is useless. In future, we will consider finer time intervals for allergy prevalence as well as for air pollution data to achieve more realistic results. Finally, we are going to use more efficient vulnerability assessments, i.e., involving more effective parameters, to produce more reliable vulnerability maps.

References

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases, *Proc. of the ACM SIGMOD Conference on Management of Data*, pp. 207-216. http://dx.doi.org/10.1145/170036.170072

Agrawal, R., and Srikant, R. (1994). Fast algorithms for mining association rules. *Proc. of the 1994 Int. Conf. Very Large Data Bases(VLDB),* Santiago, Chiel, pp. 487-499, 1994.

Akgun, A., Sezer, E.A., Nefeslioglu, H.A., Gokceoglu, C., and Pradhan, B. (2012). An easy-to-use MATLAB program (MamLand) for the assessment of landslide susceptibility using a Mamdani fuzzy algorithm. *Comput. Geosci-UK.,* 38(1), 23-34. http://dx.doi.org/10.1016/j.cageo.2011.04.012

Akinbami, L.J., Lynch, C.D., Parker, J.D., and Woodruff, T.J. (2010). The association between childhood asthma prevalence and monitored air pollutants in metropolitan areas, United States, 2001-2004. *Environ. Res.,* 110(3), 294-301. http://dx.doi.org/10.1016/j.envres.2010.01.001

Asher, M., Keil, U., Anderson, H.R., Beasley, R., Crane, J., Martinez, F., Mitchell, E.A. Pearce, N., Sibbald, B., Stewart, A.W., al.et. (1995). International study of asthma and allergies in childhood (ISAAC): rationale and methods. *Eur. Respir. J,* 8(3), 483-491. http://dx.doi.org/10.1183/09031936.95.08030483

Ayres-Sampaio, D., Teodoro, A.C, Sillero, N., Santos, C., Fonseca, J., and Freitas, A. (2014). An investigation of the environmental determinants of asthma hospitalizations: An applied spatial approach. *Appl. Geogr.,* 47, 10-19. http://dx.doi.org/10.1016/j.apgeog.2013.11.011

Bartra, J., Mullol, J., Cuvillo, A.D., Dávila, I., Ferrer, M., Jáuregui, I., Montoro, J., Sastre, J. Valero, A. (2007). Air pollution and allergens. *J. Investig. Allergol. Clin. Immunol.,* 17(Suppl 2), 3-8.

Buttenfield, B., Gahegan, M., Miller, H. and Yuan, M. (2001). *Geospatial Data Mining and Knowledge Discovery,* University Consortium for Geographic Information Science, Emerging Research Themes, Washington.

Calargun, S.U., and Yazici, A. (2008). Fuzzy association rule mining from spatio-temporal data, *Computational Science and Its Applications-ICCSA 2008,* Springer, pp. 631-646, 2008.

Dodge, M., McDerby, M., and Turner, M. (2008). *Geographic Visualization: Concepts, Tools and Applications,* John Wiley & Sons. http://dx.doi.org/10.1002/9780470987643

Douglass, J.A, and O'Hehir, R.E. (2006). Diagnosis, treatment and prevention of allergic disease: the basics. *Med. J. Australia,* 185(4), 228.

Gasana, J., Dillikar, D., Mendy, A., Forno, E., and Vieira, E.R. (2012). Motor vehicle air pollution and asthma in children: a meta-analysis. *Environ. Res.,* 117, 36-45. http://dx.doi.org/10.1016/j.envres.2012.05.001

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques,* Morgan Kaufmann Publishers Inc.

Intan, R. (2007). A proposal of fuzzy multidimensional association rules. *Jurnal Informatika*, 7(2), 85-90.

Intan, R., Yuliana, O.Y., and Handojo, A. (2009). Mining fuzzy multidimensional association rules using Fuzzy Decision Tree Induction approach. *Int.J. Comput. Netw. Secur.,* 1(2), 60-68.

Kamei, Y., Monden, A., Morisaki, S., and Matsumoto, K.I. (2008). A hybrid faulty module prediction using association rule mining and logistic regression analysis. *Proc. of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, 279-281. http://dx.doi.org/10.1145/1414004.1414051

Karimipour, F., Delavar, M.R., and Kinaie, M. (2005). Water quality management using GIS data mining. *J. Environ. Inf.,* 5(2), 61-71. http://dx.doi.org/10.3808/jei.200500047

Klosgen, W., and May, M. (2002). Spatio-temporal subgroup discovery, *Mining Spatio-Temporal Information Systems,* Kluwer Academic Publishers, Boston, pp. 149-168, 2002. http://dx.doi.org/10.1007/978-1-4615-1149-6_8

Koperski, K., Adhikary, J., and Han, J. (1996). Spatial data mining: Progress and challenges survey paper. *Proc. of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery,* pp. 55-75.

Koperski, K., and Han, J. (1995). Discovery of spatial association rules in geographic information databases, *Proc. of the 4th International Symposium on Large Spatial Databases*, Portland, Maine, pp. 47-66, 1995. http://dx.doi.org/10.1007/3-540-60159-7_4

Kulczynski, S. (1927). Die Pflanzenassoziationen der Pieninen, *Bull. Int. Acad. Pol. des Sci. Lett.(Classe Sci. Math. Nat., Sér. B) 3 (supp. 2)*, 57-203.

Ladner, R., Petry, F.E, and Cobb, M.A. (2003). Fuzzy set approaches to spatial data mining of association rules. *Trans. GIS,* 7(1), 123-138. http://dx.doi.org/10.1111/1467-9671.00133

Mamdani, E.H., and Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *Int. J. M-Mach Stud.,* 7(1), 1-13. http://dx.doi.org/10.1016/S0020-7373(75)80002-2

Mennis, J., and Guo, D. (2009). Spatial data mining and geographic knowledge discovery-An introduction. *Comput., Environ. Urban Syst.,* 33(6), 403-408. http://dx.doi.org/10.1016/j.compenvurbsys.2009.11.001

Mennis, J., and Liu, J.W. (2003). Mining association rules in spatio-temporal data, *Proc. of the 7th International Conference on GeoComputation*, 2003.

Miller, H.J., and Han, Jiawei. (2009). *Geographic Data Mining and Knowledge Discovery,* CRC Press.

Mintz, D. (2012). *Technical Assistance Document for the Reporting of Daily Air Quality-the Air Quality Index (AQI),* Office of Air Quality Planning and Standards, US Environmental Protection Agency.

Mohan, A. (2014). *A New Spatio-Temporal Data Mining Method and its Application to Reservoir System Operation,* University of Nebraska, Lincoln, Nebraska, USA.

Nembhard, D.A., Yip, K.K., and Stifter, C.A. (2012). Association rule mining in developmental psychology. International J. Appl. Ind. Eng.(IJAIE), 1(1), 23-37. http://dx.doi.org/10.4018/ijaie.2012010103

Ng, H.F., Fathoni, H., and Chen, I.C. (2009). *Prediction of Allergy Symptoms among Children in Taiwan Using Data Mining,* Deptartment of Computer Science and Information Engineering, Asia University.

Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules, *Knowledge discovery in databases*, pp. 229-248.

Ross, T.J. (2009). *Fuzzy logic with engineering applications, John Wiley & Sons,* 12(11), 78.

Shekhar, S., and Chawla, S. (2003). *Spatial Databases: A Tour (Vol. 2003),* Prentice Hall, Upper Saddle River, NJ, USA.

Shua, H., Zhub, X., and Daic, S. (2008). Mining association rules in geographical spatio-temporal data. *International Archives of the Photogrammetry,* Remote Sensing and Spatial Information Sciences. Beijing, pp. 225-228, 2008.

Srikant, R., and Agrawal, R. (1996). Mining quantitative association rules in large relational tables. *Proc. of the ACM SIGMOD Conference on Management of Data*, pp. 1-12. http://dx.doi.org/10.1145/235968.233311

Wackernagel, H. (2003). *Multivariate Geostatistics-An Introduction with Applications,* Springerr-Verlag, Heidelberg, Berlin. http://dx.doi.org/10.1007/978-3-662-05294-5

Wang, F. (2010). Application of multidimensional association rule techniques in manufacturing resource planning system, *Proc. of the International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 1151-1169, 2010. http://dx.doi.org/10.1109/FSKD.2010.5569197

Yadav, P.K., and Rizvi, S. (2014). An exhaustive study on data mining techniques in mining of Multimedia database, *Proc. of the International Conference on Issues and Challenges in Intelligent Computing Techniques*, pp. 541-545, 2014. http://dx.doi.org/10.1109/ICICICT.2014.6781339

YoussefAgha, A., Jayawardene, W., Lohrmann, D., and Afandi, G.E.. (2013). Application of data mining techniques to predict allergy outbreaks among elementary school children. *J. Commu. Comput,* (4), 451-460.

Zöllner, I.K., Weiland, S.K., Piechotowski, I., Gabrio, T., Mutius, E.V., Link, B., Pfaff, G., Kouros, B., Wuthe, J. (2005). No increase in the prevalence of asthma, allergies, and atopic sensitisation among children in Germany: 1992-2001. *Thorax,* 60(7), 545-548. http://dx.doi.org/10.1136/thx.2004.029561