GEODESY AND CARTOGRAPHY
ISSN 2029-6991 / eISSN 2029-7009
2015 VOLUME 41(2): 101–112
http://dx.doi.org/10.3846/20296991.2015.1051339

UDK 528.946

FUZZY SPATIAL ASSOCIATION RULE MINING TO ANALYZE THE EFFECT OF ENVIRONMENTAL VARIABLES ON THE RISK OF ALLERGIC ASTHMA PREVALENCE

Yousef KANANI SADAT[1], Tina NIKAEIN[2], Farid KARIMIPOUR[3]

Department of Surveying and Geomatics Engineering, College of Engineering, University of Tehran, North Kargar Ave., Jalal-Al-Ahmad Crossing, Tehran, Iran

E-mails: [1]yousefkanani@ut.ac.ir (corresponding author); [2]tina_nikaein@ut.ac.ir; [3]fkarmipr@ut.ac.ir

Received 19 March 2015; accepted 20 May 2015

Abstract. The prevalence of allergic diseases has greatly increased in recent decades, likely due to contamination of the environment with allergy irritants. One common treatment is identifying that allergy irritant, and then avoiding exposure to it. This article studies the relation between the prevalence of allergic asthma and certain allergy irritants that are related to environmental variables. To that end, we use spatial association rule mining to determine the association between the spatial distribution of allergic asthma prevalence and air pollutants such as CO, SO_2, NO_2, PM_{10}, $PM_{2.5}$, and O_3 (from data compiled by air pollution monitoring stations), as well as other factors, such as the distance of residence from parks and roads. In order to clear up the uncertainties inherent in the attributes linked to the spatial data, the dimensions in question have been defined as fuzzy sets. Results for the case study (i.e. Tehran metropolitan area) indicate that distance to parks and roads, as well as CO, NO_2, PM_{10}, and $PM_{2.5}$ levels are related to allergic asthma prevalence, while SO_2 and O_3 are not. Finally, we use the extracted association rules in fuzzy inference system to produce the spatial risk map of allergic asthma prevalence, which shows how much is the risk of allergic asthma prevalence at each point of the city.

Keywords: spatial data mining, fuzzy spatial association rule mining, risk analysis, allergic asthma, air pollution; GIS.

Introduction

Over the last few decades, the prevalence of allergic diseases has greatly increased, especially among children. This is widely assumed to be due to modern living conditions in environments contaminated with allergy irritants (Zöllner *et al.* 2005, Ng *et al.* 2009). Allergy patients have hypersensitive immune systems that react abnormally to usually harmless substances. Allergic reactions can be caused by a variety of factors, depending on genetic conditions, lifestyle and habits, foods, as well as geography and environmental conditions (Asher *et al.* 1995).

Allergic asthma, which causes airway obstruction and inflammation, is a type of asthma triggered by allergens. Unlike non-allergic asthma, the symptoms of this type of asthma are associated with an allergic reaction involving the immune system. Many of the symptoms of allergic and non-allergic asthma are the same (coughing, wheezing, shortness of breath or rapid breathing, chest tightness); however, allergic asthma is triggered by inhaled irritants such as pollen etc. which then result in asthma symptoms (Rackemann 1947, Romanet Manent *et al.* 2002).

The most commonly used approach for the treatment of allergic diseases is simply identifying the irritant that exacerbates the allergy, and then having the patient avoid exposure to it (Douglass, O Hehir 2006). Many of these irritants are likely to be related to environmental variables. Analyzing data about the environment of allergy patients may therefore lead to identifying the roles of certain environmental variables in the prevalence of allergic diseases.

The databases in question store huge amounts of data. To determine valid, novel, useful, and understandable data patterns, data mining techniques (Miller, Han 2001) are being widely used. Ng *et al.* (2009) used data mining techniques to predict allergy symptoms among children in Taiwan by collecting allergy data

Copyright © 2015 Vilnius Gediminas Technical University (VGTU) Press
http://www.tandfonline.com/TGAC

from children under the age of 12. These data came in the form of 30 predictor variables, including personal factors, health behavior factors, living condition factors, family factors, and allergy-inducing factors. Akinbami *et al.* (2010) assessed the relationship 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. Similarly, the impact of air pollution on respiratory diseases in children with asthma was studied by Esposito *et al.* (2014). YoussefAgha *et al.* (2012) studied the application of data mining techniques in determining if there is any relation between the prevalence of allergies among elementary school children and daily upper-air observations (i.e. temperature, relative humidity, dew point, and mixing ratio) as well as daily air pollution (CO, SO_2, NO_2, PM_{10}, $PM_{2.5}$, and O_3).

As the patients are distributed in space, results are improved if the spatial characteristics of the data are taken into account when studying the prevalence of allergies. 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. 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.

In this article, we use spatial data mining to study the effects of environmental variables on the prevalence of allergic asthma. Spatial data mining involves the development and application of novel computational techniques to analyze very large spatial databases (Buttenfield *et al.* 2001). A major distinction of spatial data mining is that attributes of the neighboring objects influence each other, and thus must be taken into 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 (Miller, Han 2001). Instead of a purely statistical approach, we take the spatial characteristics of the environment into account. Furthermore, rather than plain regression models, we use association rule mining as a more powerful tool to extract the associations between allergic asthma and air pollution (Karimipour, Kanani-Sadat 2015).

Discovering association rules from data stored in spatial databases has been practiced in many research projects. Mennis and Liu (2003) explored spatiotemporal association rules among a set of variables characterizing the socioeconomic and land-cover changes in the Denver, Colorado region from 1970 to 1990. Shua *et al.* (2008) produced 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, and directional and geometric relationships through fuzzy spatial data mining in order 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 spatiotemporal data cubes and the Apriori algorithm in order to generate fuzzy association rules.

In our approach, the places of residence of a group of allergic asthma patients, all from the Tehran metropolitan area, as well as the spatial characteristics of the environment (e.g. the location of parks, roads, and air pollution monitoring stations) were placed on the map. We then utilized spatial association rule mining to extract the associations between allergic asthma prevalence and air pollutants 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 $<10\mu m$ and $<2.5\mu m$, respectively), and O_3 (ozone), as well as the distance of the respective place of residence from roads and the parks effect. With the "park effect", we mean the effect of park in increasing the risk of allergy prevalence due to vegetation, pollens, etc. We used fuzzy multi-dimensional association rule mining in order to involve many parameters, and to handle the uncertainty inherent in the attributes linked to the spatial data. Finally, these mined associations were used to analyze the effect of environmental variables on the risk of allergic asthma prevalence in the case study area.

The rest of the article is organized as follows: Section 1 introduces association rule mining and its related concepts, as used in our research methodology and implementation. In Section 2, the components of our research methodology are described in detail. The results for the case study are presented and discussed

in Section 3. Finally, conclusion section contains concluding remarks and ideas for future research in this direction.

1. Preliminaries

This section briefly introduces association rule mining and its related concepts that are referred to in the rest of the article. Readers familiar with these concepts can skip this section.

1.1. Association rule mining

Association rule mining seeks interesting association or correlation relationships among a large set of data items, i.e. certain data items that often occur together (Han *et al.* 2011). An association rule is an implication of the form $A \rightarrow B$ where A (the antecedent) and B (the consequent) are sets of predicates. For example, a rule like "the person who lives in area with very high amount of NO_2 and very high park effect, suffers from allergic asthma" is an association rule, and is expressed like this:

$$(NO_2, \text{very high}), (\text{park\_effect, very high}) \rightarrow (\text{allergic\_asthma, yes}) . \quad (1)$$

If there is only one type of predicate (e.g. park_effect), the association rule is one-dimensional. Multi-dimensional association rules on the other hand involve more than one type of predicate.

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

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

$$confidence\ (A \rightarrow B) = prob\{B \mid A\} = \frac{prob\{A \cap B\}}{prob\{A\}}. \quad (2)$$

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

$$A \rightarrow B \text{ [support, confidence, correlation]} . \quad (3)$$

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

$$Kulc\ (A, B) = \frac{1}{2}\left(P(A|B) + P(B|A)\right). \quad (4)$$

It always has a value between 0 and 1. A larger *Kulc* indicates stronger relation between A and B. Those association rules with a certain minimum significant support, confidence, and correlation are called strong association rules, and they are usually the ones considered in the decision-making process. A common influential algorithm for association rule mining is the so-called Apriori algorithm (Agrawal, Srikant 1994).

A spatial association rule contains at least one spatial relationship in an antecedent or consequent predicate (Koperski, Han 1995). For example, *distance_to (road, near)* is a spatial predicate that results in a spatial association rule.

1.2. Fuzzy association rule mining

Fuzzy association rule mining utilizes fuzzy sets to mine association rules in a given attribute data set. It 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). For example, a membership function for fuzzy "nearness" could be defined over distance as (Intan *et al.* 2009):

$$very\_near(x) = \begin{cases} 1 & x \le 75 \\ \dfrac{200 - x}{125} & 75 \le x \le 200 \\ 0 & x \ge 200 \end{cases}. \quad (5)$$

A fuzzy association rule consists of fuzzy data items. Using the previous definition of fuzzy nearness, an example of a multi-dimensional fuzzy association rule with the predicates distance_to, NO_2 and disease is (Intan *et al.* 2009):

$$\text{distance\_to (road, very near)} \wedge NO_2\ (X, \text{high}) \rightarrow \text{disease (X, allergic\_asthma)}. \quad (6)$$

If A and B are fuzzy data sets and d is a fuzzy data item, then the values of support, confidence, and correlation are extended to fuzzy association rule mining as (Intan 2007):

$$\text{support}(A \rightarrow B) = \text{support}(A \cup B) = \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \left\{\mu_{C_j}(d_{ij})\right\}}{r},$$

$$\text{confidence}(A \rightarrow B) = \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cap B} \left\{\mu_{C_j}(d_{ij})\right\}}{\sum_{i=1}^{r} \inf_{A_j \in A} \left\{\mu_{A_j}(d_{ij})\right\}}, \quad (7)$$

$$\text{correlation}(A \rightarrow B) = \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \left\{\mu_{C_j}(d_{ij})\right\}}{\sum_{i=1}^{r} \inf_{A_j \in A} \left\{\mu_{A_j}(d_{ij})\right\} \times \sum_{i=1}^{r} \inf_{B_k \in B} \left\{\mu_{B_k}(d_{ik})\right\}}$$

where: μ_x is the membership value of x, *inf* (S) is the infimum (i.e. the greatest lower band) of the fuzzy set S, and r is the number of items in the dataset.

1.3. Fuzzy spatial association rule mining

The need to handle imprecise and uncertain information concerning spatial data has been widely recognized (Goodchild and Gopal 1990). Spatial uncertainty relates either to a lack of knowledge about the position and shape of an object with an existing, real boundary (positional uncertainty), or to the inability of measuring such an object precisely (measurement uncertainty) (Schneider 2008). Many operations are applied to spatial data under the assumption that features, attributes, and their relationships have been specified in a precise and exact manner. However, inexactness often exists in the positions of features and the assignment of attribute values and may be introduced at various stages of data compilation and database development (Ladner *et al.* 2003). A number of researchers using spatial databases have previously utilized fuzzy set approaches for their modeling of spatial data. Some early work by geographers in the 1970s utilized fuzzy sets in topics such as behavioral geography and geographical decision-making (Gale 1972; Leung 1979; Pipkin 1978; Burrough, Frank 1996). However, the first consistent approach to the use of fuzzy set theory in regards to spatial data was developed by Robinson (1988).

In spatial association rule mining, uncertainty is involved in almost every step of data processing: from data pre-processing, through data conceptualization, to association rules extraction, which then leads to the final results. Intrinsically, all sources of uncertainty in data mining are reduced to interest measures (i.e. support, confidence, and correlation) for the final association rules. Using fuzzy sets in the mining of association rules from spatial databases is useful because fuzzy sets are able to model the uncertainty embedded in the meaning of the data (Calargun, Yazici 2008). There are many problems in association rule mining, such as sharp boundaries, for which fuzzy association mining provides clear solutions (Jain *et al.* 2013).

2. Research methodology

This article analyzes the effect of environmental variables on the prevalence of allergic asthma by deploying fuzzy spatial association rule mining. Our case study is the Tehran metropolitan area. The steps of our research methodology are as follows (Fig. 1):

2.1. Data pre-processing

The air pollutants we are analyzing are CO, SO_2, NO_2, PM_{10}, $PM_{2.5}$, and O_3. Their presence in the air was measured every hour for all of December 2013 by Tehran's air pollution monitoring stations (Fig. 2). This data is cleaned by filling the gaps (using interpolation by a Fourier series), and by filtering the noises (using statistical methods). To reduce the massive volume of hourly data to a monthly representation of air pollutants, the monthly average of maximum values observed for each parameter on each day is computed. These values

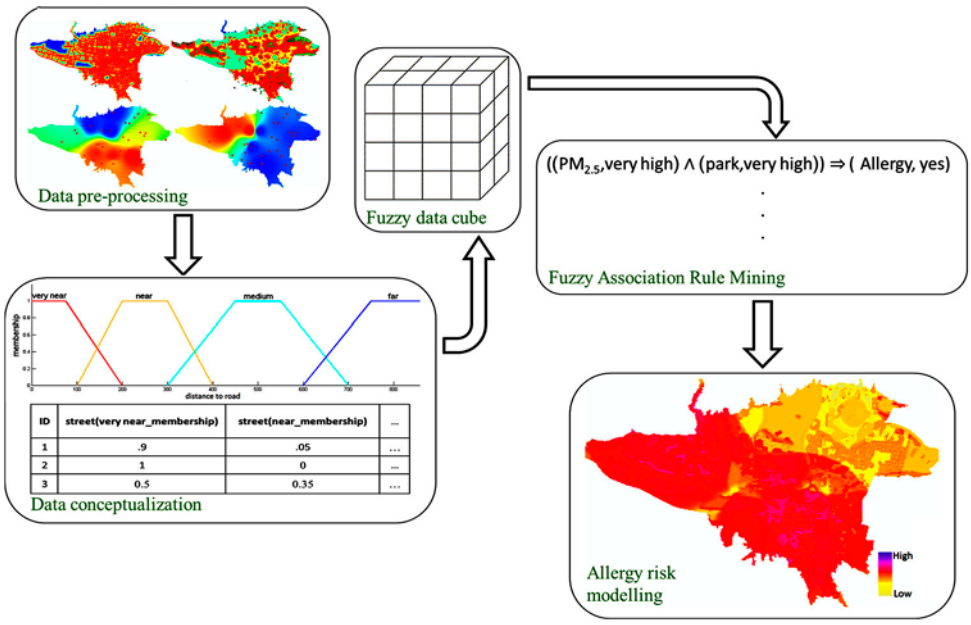


Fig. 1. Research methodology

are then used to produce a monthly pollution distribution map of Tehran for each air pollutant through Kriging spatial interpolation (Fig. 3) (Wackernagel 2003).

To model the effect of distance to roads, we produced a map (Fig. 4a) in which each point is assigned the distance to the nearest road. The same process was used to model the effect of parks (Fig. 4b), using the following equation to quantify the effect of nearby parks:

$$T_j = \sum \frac{A_i}{d_{ij}^2} , \quad (8)$$

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.

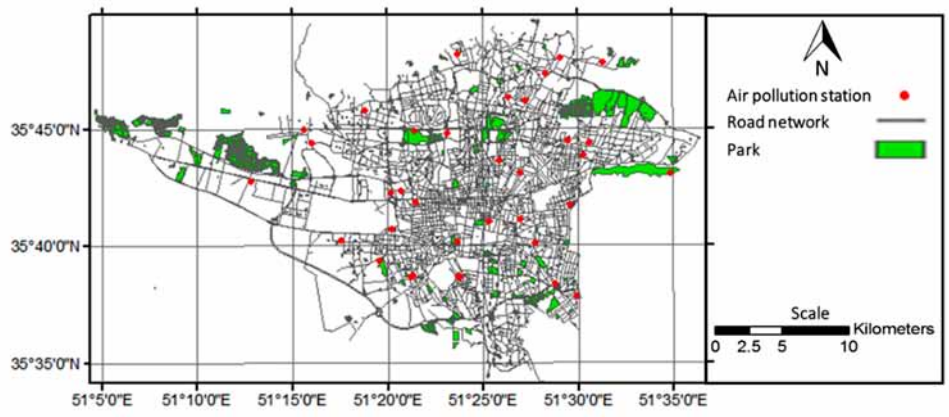


Fig. 2. The map of Tehran's roads, parks and air pollution monitoring stations

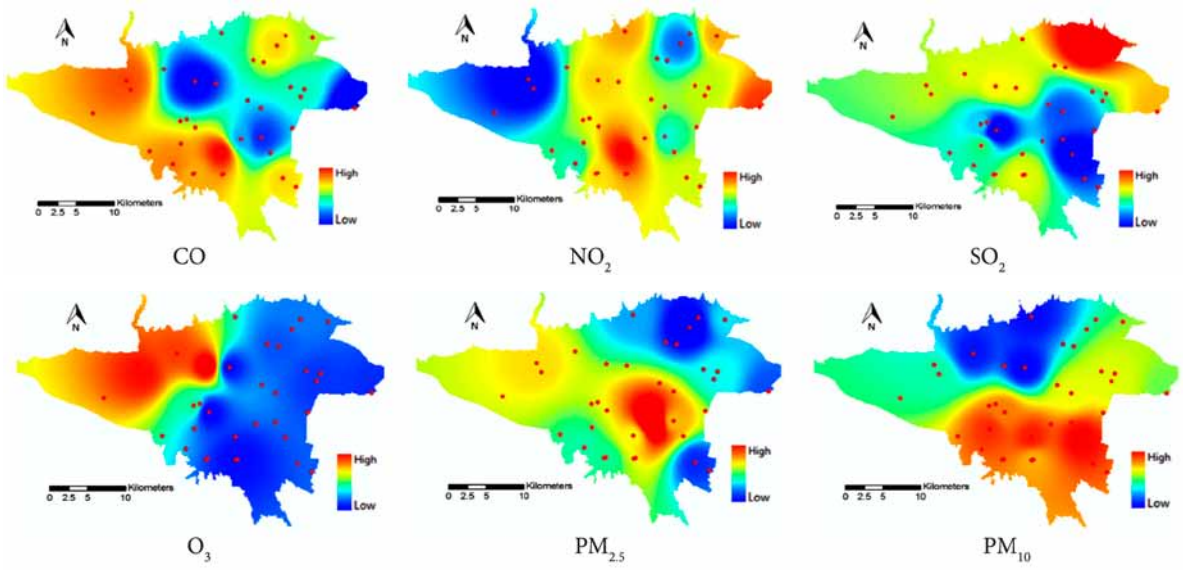
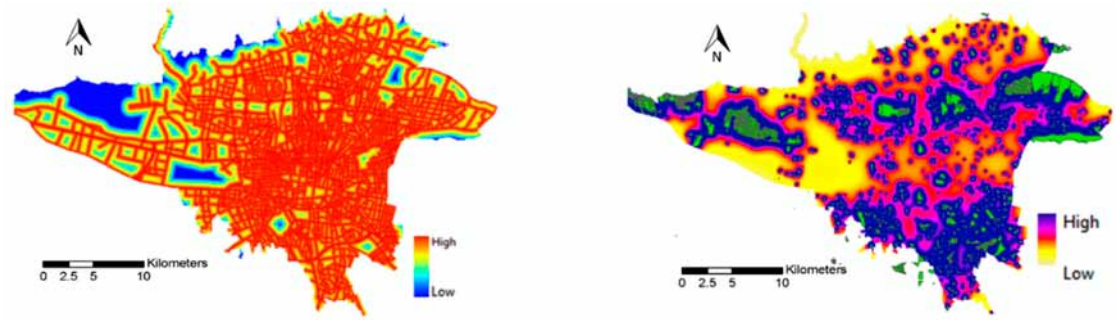


Fig. 3. The distribution maps of air pollutants in Tehran in December 2013



a) b)

Fig. 4. The maps for the effect of (a) distance to roads and (b) parks

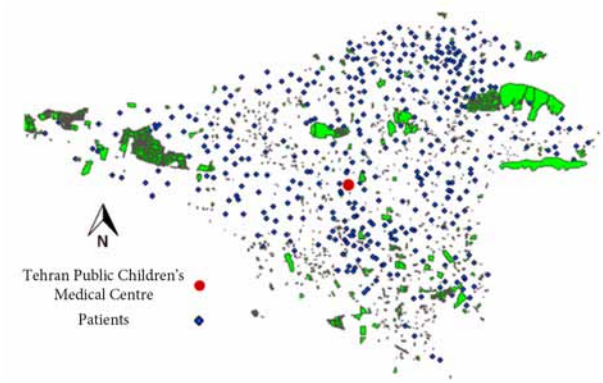


Fig. 5. The map of patients' place of residence

Finally, the place of residence for the randomly selected 1000 patients who visited the "Tehran Public Children's Medical Clinic" in December 2013 are placed on the map (Fig. 5), 284 number of which had allergic asthma. For each patient, a data item is stored that shows if he/she is suffering from allergic asthma. Moreover, having overlaid this map with the distribution maps of the air pollutants, the maps of parks effect and distance to roads, the estimated values for these variables are assigned to each point as data items (attributes).

2.2. Data conceptualization

Association rule mining can only deal with categorical (classified) data. Therefore, the data items assigned to the patients must be categorized. For the air pollutants, this is achieved through the air quality index (AQI), which is an indicator that characterizes air quality as "very high", "high", "moderate" and "low". As the categorization breakpoints used by AQI vary from one air pollutant to another (Table 1), the following Equation is used to normalize the measured values (Mintz 2012):

$$I_P = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}}\left(C_P - BP_{Lo}\right) + I_{Lo}, \qquad (9)$$

where:

I_p = the air quality index for the air pollutant p
C_p = the value measured for the air pollutant 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}

In order to deal with sharp break points between the categories, fuzzy labels are assigned to the data items. Using the concept of fuzzy sets, we consider the uncertainty and errors that may exist in the data. At the first stage of raw data pre-processing, uncertainty is introduced by the positions of the air pollution monitoring stations, the values measured by the stations, and the interpolation function used for estimated lost data. At the stage of data conceptualization, uncertainty is introduced by applying the spatial interpolation function (Kriging). Additionally, uncertain boundaries may occur in the decomposition of numerical data into different categories.

The membership function illustrated in Figure 6a classifies the distance to roads into "very near", "near", "medium", and "far". For instance, based on Figure 6a, if the distance to a road is 60 *m*, the membership of this value to the class "very near" is 1, and to the other classes it is 0. For a distance of 175 *m*, the membership of this value to the class "very near" is 0.2, to the "near" class it is 0.75, and to the other classes it is 0. The same process classifies the effect of parks into "very highly affected", "highly affected", "moderately affected", and "lowly affected" (Fig. 6b). The membership functions are defined based on the study data, i.e., each class has approximately equal area.

2.3. Fuzzy data cube construction

The multi-dimensional data cube is a common organization form of data for data mining in data warehouse

Table 1. Breakpoints for the AQI (Mintz 2012)

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	–

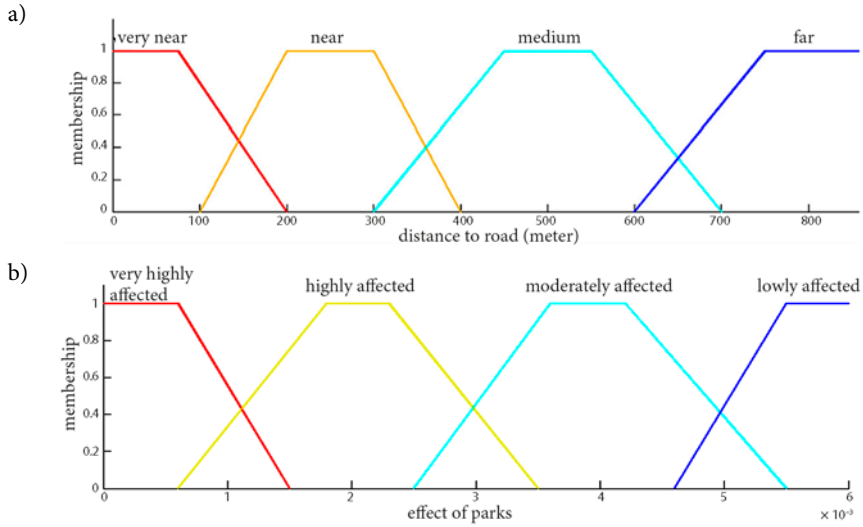


Fig. 6. The membership functions to classify (a) distance to roads and (b) effect of nearby parks

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 each cell stores the number of tuples that have that attribute value (Han *et al.* 2011).

Here, we deal with the data cube as a fuzzy object in order to optimally discover the knowledge (Ladner *et al.* 2003). In this case, each cell contains the sum of the minimum of the membership values of the corresponding fuzzy labels. For example, suppose there are two attribute predictors ***a*** and ***b***. Consider Table 2 that shows the membership values of five data items "1" to "5" to ***a*** and ***b***; each may have two fuzzy labels: *a1* and *a2* for ***a*** and *b1* and *b2* for ***b***. Here, each cell shows the membership value of that attribute to the corresponding fuzzy label. Table 3 shows the computation of the ($a1$, $b1$) membership value for each data item. This process is done for all combinations of ***a*** and ***b*** fuzzy labels, which finally results in the data cube illustrated in Table 4.

For our case study data, we constructed a 9-D data cube whose dimensions are spatial characteristics of the residence location of the patients (i.e. air pollution, park effect and distance to roads), as well as a binary value indicating whether he/she is suffering from allergic asthma.

2.4. Fuzzy spatial association rule mining

Having constructed our fuzzy data cube, the association rules between allergy prevalence and spatial characteristics of the residence location are extracted, along with their supports, confidences, and correlations (as described in Section 1). Since we are interested in antecedents that result in allergic asthma, we only keep those rules whose consequence is "*(allergy, yes)*", such as:

$$[(PM_{2.5}, \text{very high}), (\text{park\_effect}, \text{very high})] \rightarrow (\text{allergy}, \text{yes}).$$

Table 2. The membership values for the attributes *a* and *b*

Tuples	$\mu_{a1}(a)$	$\mu_{a2}(a)$	$\mu_{b1}(b)$	$\mu_{b2}(b)$
1	1	0	0	1
2	0.66	0.27	1	0
3	0.53	0.41	0.83	0.11
4	0.87	0.10	0	1
5	0.33	0.61	0.83	0.11

Table 3. Computation of ($a1$, $b1$) membership value

Tuples	$\mu_{a1}(a)$	$\mu_{b1}(b)$	Min
1	1	0	0
2	0.66	1	0.66
3	0.53	0.83	0.53
4	0.87	0	0
5	0.33	0.83	0.33
Sum			1.52

Table 4. The 2D data cube constructed for *a* and *b* (The values are computed as described in Table 2)

	*a*1	*a*2	Sum
*b*1	1.52	1.29	2.81
*b*2	2.09	1.22	3.31
Sum	3.61	2.51	6.12

2.5. Risk analysis

The extracted association rules are now used to analyze the effect of environmental variables on the risk of allergic asthma prevalence and to produce the spatial risk map of allergic asthma prevalence in the entire city of Tehran. For this, the GIS-fuzzy integration is used as follows (Ross 2009).

The support and confidence thresholds are set to zero in order to collect all the association rules (no matter how 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 7. Table 5 illustrates three of the association rules along with their corresponding fuzzy rules and Kulczynski categories.

Fuzzy inference is the process of mapping a given input to an output using fuzzy logic. Given the two following sample rules, we describe the calculation of risk of allergic asthma for a sample point ***P*** on the map based on Mamdani's fuzzy inference method (Akgun *et al.* 2012; Mamdani, Assilian 1975).

(1) **IF** *park_effect* **is** *high* **THEN** *Kulc* **is** *k4*

(2) **IF** *park_effect* **is** very *high* **AND** *road* **is** *near* **THEN** *Kulc* **is** *k6*

The numerical values for "distance to road" and "effect of park" were extracted for point ***P*** from the maps produced in section 2.1, and are 150 *m* and 0.001, respectively. Fuzzy output is then calculated according to the fuzzy rules. Figs 8a and 8b show the output of rules #1 and #2 respectively for point ***P***. Note that in the rules with an AND connector, which is the case for rule #2, the minimum of the membership values is considered (Fig. 8b). The result of the fuzzy inference is a fuzzy subset composed of the slices of normalized Kulczynski: *k*6 (red) and *k*4 (brown). To assign a crisp value, among several methods, the centroid of the area is considered. Finally, the risk of allergic asthma prevalence for point ***P*** according to the above two rules is estimated, which is 46.76% (Fig. 8c).

We utilized the Fuzzy Logic Design toolbox of Matlab R2012b to calculate the risk of allergic asthma prevalence in the entire city of Tehran. Having defined the fuzzy rules, the maps of those air pollutants that affect allergic asthma (Fig. 3), induced by our association rule mining, as well as the maps of park effect and distance to road (Fig. 4) are combined through the Mamdani fuzzy inference method (Fig. 9) to compute the Kulczynski relation factor for each point of the city, from which the risk map of allergic asthma prevalence is produced.

3. Results and discussion

Applying the procedure described in Section 2 to the data set provided 60 association rules between allergic asthma prevalence and the environmental variables of

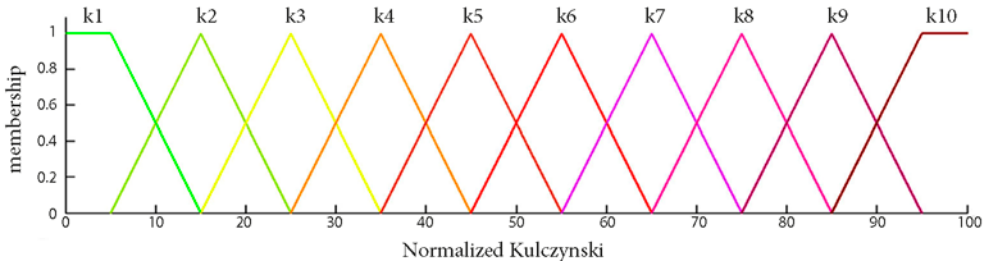


Fig. 7. The function to fuzzify the Kulczynski correlation factors

Table 5. Three of the association rules (rows 2 to 4) along with their corresponding fuzzy rule (rows 5 to 7) and Kulczynski category

	ID	Rule	Normalized Kulc
Association rules	1	[(park_effect, 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_effect, very high)] → (allergy, yes)	97.94
Fuzzy rules	1	**IF** *park_effect* **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_effect* **is** *very high* **THEN** *Kulc* **is** *k10*	

December 2013, some of which are illustrated in Table 6 (the minimum support and confidence thresholds were respectively defined as 5% and 30%, by practice). For example, rule #5 with 6.78% support and 75.89% confidence says that 6.78% of the total sampling population lives in locations where the amounts of NO_2 and $PM_{2.5}$ and the effect of nearby parks are very high, and that they are suffering from allergic asthma; this is true

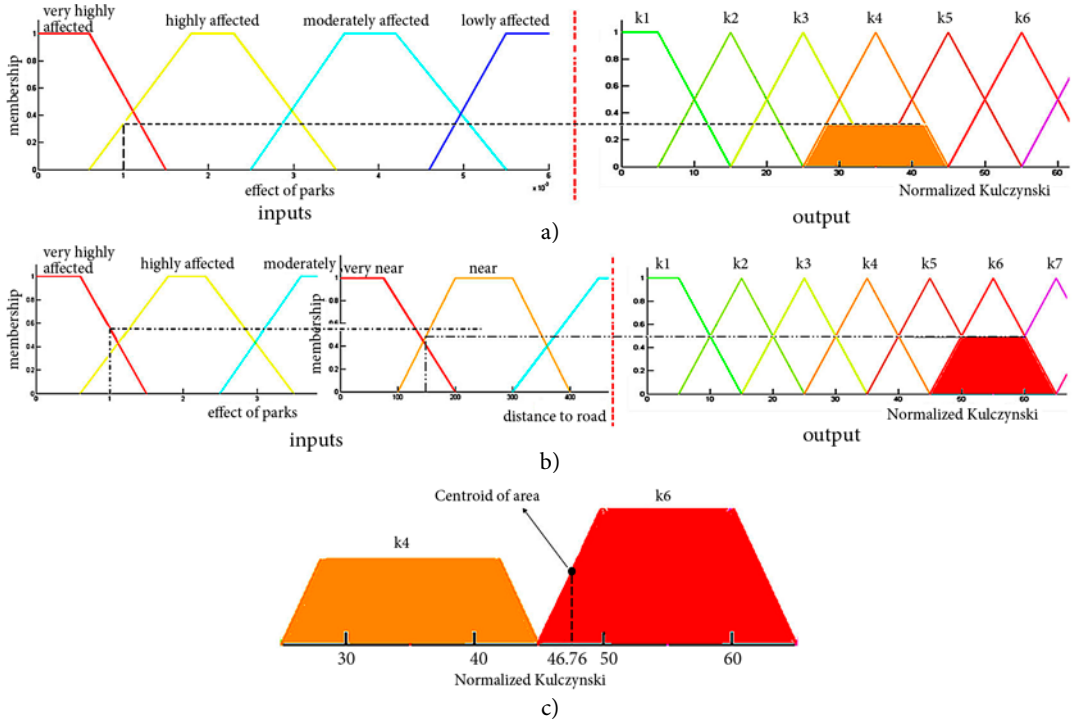


Fig. 8. Calculation of the fuzzy outputs for the rules (a) #1 (b) #2; (c) Assigning a crisp output

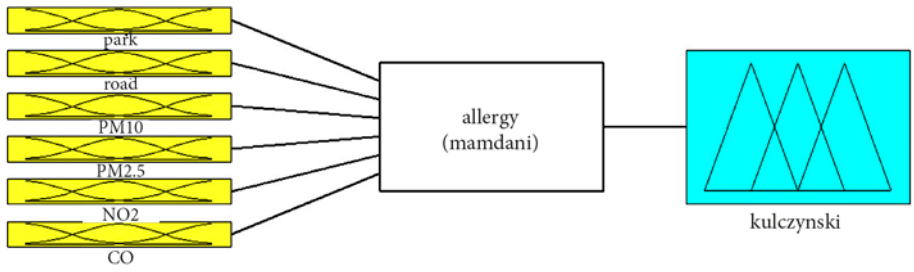


Fig. 9. Combining the fuzzy rules to compute the Kulczynski relation factor for each point

Table 6. 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)	13.12	33.84	0.64
2	[($PM_{2.5}$, very high), (park_effect, very high)] → (allergy, yes)	9.28	55.57	0.62
3	[($PM_{2.5}$, very high), (PM_{10}, very high)] → (allergy, yes)	9.51	53.26	0.60
4	[($PM_{2.5}$, very high), (PM_{10}, very high), (park_effect, very high)] → (allergy, yes)	6.05	71.25	0.62
5	[($PM_{2.5}$, very high), (NO_2, very high), (park_effect, very high)] → (allergy, yes)	6.78	75.89	0.65
6	[(NO_2, very high), (CO, very high), (park_effect, very high)] → (allergy, yes)	5.10	82.58	0.65
7	[(PM_{10}, very high), (NO_2, very high), (CO, very high)] → (allergy, yes)	8.02	67.36	0.62
8	[($PM_{2.5}$, very high), (PM_{10}, very high), (NO_2, very high), (park_effect, very high)] → (allergy, yes)	6.02	79.72	0.64
9	[($PM_{2.5}$, very high), (PM_{10}, very high), (NO_2, very high), (CO, very high)] → (allergy, yes)	6.57	71.23	0.61
10	[(PM_{10}, very high), (NO_2, very high), (CO, very high), (road, very near)] → (allergy, yes)	5.26	71.56	0.58
11	[($PM_{2.5}$, very high), (PM_{10}, very high), (NO_2, very high), (CO, very high), (road, very near)] → (allergy, yes)	5.37	72.51	0.61

for 75.89% of the sampling population that lives in these areas; the *Kulczynski*'s correlation measure between the antecedent and the consequence is 65%.

Based on the extracted rules, the "park effect" and "distance to roads" as well as the levels of CO, NO_2, PM_{10}, and $PM_{2.5}$ affected the prevalence of allergic asthma in December, while SO_2 and O_3 showed no significant relation. According to the extracted associations, the rules that include "*(NO_2, very high)*" or "*(CO, very high)*" have greater confidence compared to those that do not have these pollutants, thus these pollutant specifically have a more adverse impact on allergic asthma. Considering that the rules that include "*(PM_{10}, very high)*" have less confidence than other rules, the adverse impact of this pollutant is less than that of others. On the other hand, the rules that include "*(park_effect, very high)*" and for which at least one of the air pollutants is high (e.g. rules #2, #4 and #6) has greater confidence compared to those that only have one of these components (e.g. rules #1, #3 and #7). Other research on air pollution and allergic asthma cements this assertion: environmental pollution influences pollen allergenicity (Bartra *et al.* 2007). Scientific evidence shows that pollen in heavily polluted zones expresses a larger amount of proteins described as being allergenic, compared to areas characterized by less pollution (Armentia *et al.* 2002, Cortegano *et al.* 2004, D'Amato 2000). In effect, the allergenicity of such aeroallergens may be increased, their transport may be favored, and even their atmospheric concentration may be increased secondary to a rise in their production or exposure time (Bartra *et al.* 2007). On the other hand, rules # 10 and #11, which contain *"(road, very near)"*, have no significant increase in confidence as the effect of this parameter already manifested in an increase of air pollutants. Considering that the best approach for treatment of allergic disease is avoiding the allergic irritant, to improve allergic asthma, patients are recommended to avoid living near the park and polluted areas. According to the discovered rules, people who live in areas with higher amounts of CO, NO_2, PM_{10}, and $PM_{2.5}$, and near to parks, have the highest risk of allergic asthma prevalence.

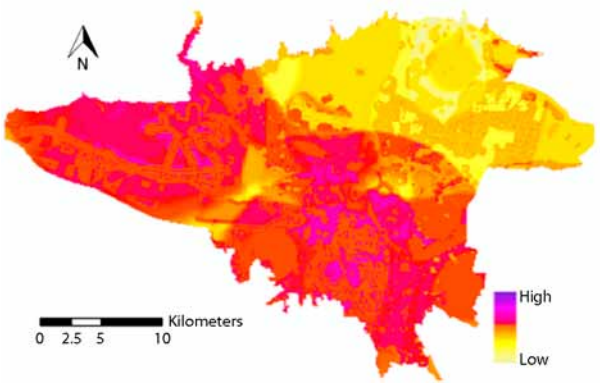


Fig. 10. The risk map for the effect of environmental variables on the prevalence of allergic asthma in Tehran for December

Finally, the process described in subsection 2.5 provided 2148 fuzzy rules for December, which yields the risk map for the effect of environmental variables on the prevalence of allergic asthma in Tehran for December. This map visually agrees with the epidemiology of asthma prevalence in Tehran, which was certified by the clinic's doctors.

Conclusions and future work

This article uses fuzzy spatial association rule mining to investigate the relation between the prevalence of allergic asthma and certain environmental variables. Through that relation, we were able to produce a risk map that shows the effect of those environmental variables on the prevalence of allergic asthma. The results for the case study (i.e. the Tehran metropolitan area) show that by considering the spatial distribution of the patients as well as the fuzzy definition of data items (i.e. attributes) allowed us to extract more reliable associations, and consequently arrive at more valuable interpretations of the data. The visualized risk map of Tehran could help allergic asthma patients avoid exposure to allergy irritants. However, as air pollution conditions (and pollen load as well) vary over time, the extracted rules and the map they produced apply only to December, and may not be applicable to other months.

In addition, for this study we only involved the distance to parks and roads as spatial characteristics that may affect air pollution and allergic asthma. Including other spatial characteristics as well may further improve the results. On the other hand, the hourly data provided by the air pollution stations was integrated to only one AQI (air quality index) for each parameter and each month, in order to avoid heavy computation loads. This integration may result in ignoring the sub-monthly variations of air quality. Considering finer time intervals (e.g. daily) may provide more realistic results. Lastly, more efficient risk assessments, i.e. involving more effective parameters, may be used to produce more reliable vulnerability maps.

References

Agrawal, R.; Srikant, R. 1994. Fast algorithms for mining association rules. Very large data bases, in *Proceedings of 20th*

International Conferrence, 12–15 September 1994, Santiago de Chile, Chile, 487–499.

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

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

Armentia, A.; Lombardero, M.; Callejo, A.; Barber, D.; Martin Gil, F.; Martin-Santos, J.; Vega, J.; Arranz, M. 2002. Is pollen from an urban environment more allergenic than rural pollen?, *Allergologia et immunopathologia* 30: 218–224. http://dx.doi.org/10.1016/S0301-0546(02)79124-6

Asher, M.; Keil, U.; Anderson, H.; Beasley, R.; Crane, J.; Martinez, F. Mitchell, E.; Pearce, N.; Sibbald, B.; Stewart, A. 1995. International study of asthma and allergies in childhood (ISAAC): Rationale and methods, *European Respiratory Journal* 8: 483–491.

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

Bartra, J.; Mullol, J.; Del Cuvillo, A.; Dávila, I.; Ferrer, M.; Jáuregui, I.; Montoro, J.; Sastre, J.; Valero, A. 2007. Air pollution and allergens, *Journal of Investigational Allergology and Clinical Immunology,* 17: 3–8.

Burrough, P. A.; Frank, A. 1996. Geographic objects with indeterminate boundaries. CRC Press.

Buttenfield, B.; Gahegan, M.; Miller, H. J.; Yuan, M. 2001. *Geospatial data mining and knowledge discovery*. Washington: University Consortium for Geographic Information Science.

Calargun, S. U.; Yazici, A. 2008. Fuzzy association rule mining from spatio-temporal data, *Lecture Notes in Computer Science* 5072: 631–646. http://dx.doi.org/10.1007/978-3-540-69839-5_47

Cortegano, I.; Civantos, E.; Aceituno, E.; Del Moral, A.; Lopez, E.; Lombardero, M.; Del Pozo, V.; Lahoz, C. 2004. Cloning and expression of a major allergen from Cupressus arizonica pollen, Cup a 3, a PR-5 protein expressed under polluted environment, *Allergy* 59: 485–490. http://dx.doi.org/10.1046/j.1398-9995.2003.00363.x

D'Amato, G. 2000. Urban air pollution and plant-derived respiratory allergy, *Clinical and Experimental Allergy* 30: 628–636. http://dx.doi.org/10.1046/j.1365-2222.2000.00798.x

Douglass, J. A.; O Hehir, R. E. 2006. Diagnosis, treatment and prevention of allergic disease: the basics, *Medical journal of Australia* 185: 228–233.

Esposito, S.; Galeone, C.; Lelii, M.; Longhi, B.; Ascolese, B.; Senatore, L.; Prada, E.; Montinaro, V.; Malerba, S.; Patria, M. F. 2014. Impact of air pollution on respiratory diseases in children with recurrent wheezing or asthma. *BMC Pulmonary Medicine* 14: 130–138. http://dx.doi.org/10.1186/1471-2466-14-130

Gale, S. 1972. Inexactness, fuzzy sets, and the foundations of behavioral geography*, *Geographical Analysis* 4: 337–349. http://dx.doi.org/10.1111/j.1538-4632.1972.tb00480.x

Gasana, J.; Dillikar, D.; Mendy, A.; Forno, E.; Ramos Vieira, E. 2012. Motor vehicle air pollution and asthma in children: a meta-analysis, *Environmental research* 117: 36–45. http://dx.doi.org/10.1016/j.envres.2012.05.001

Goodchild, M.; Gopal, A. 1990. *The accuracy of spatial databases.* London: Taylor and Francis.

Han, J.; Kamber, M.; 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: 85–90.

Intan, R.; Handojo, A.; Yenty, Y. O. 2009. Mining fuzzy multidimensional association rules using fuzzy decision tree induction approach, *International Journal of Computer and Network Security* 1: 60–68.

Jain, S.; Jain, A. P. S.; Jain, A. 2013. An assessment of fuzzy temporal association rule mining, *International Journal of Application or Innovation in Engineering* & *Management* 2(1): 42–45.

Karimipour, F.; Kanani-Sadat. Y. 2015. Investigating the relation between prevalence of asthmatic allergy with the characteristics of the environment using fuzzy association rule mining, *Journal of Geomatics Science and Technology* 4(3): 117–130.

Koperski, K.; Han, J. 1995. Discovery of spatial association rules in geographic information databases, in *4th International Symposium on Large Spatial Databases,* 18–19 January 1995, Berlin, Germany. Berlin: Springer-Verlag. http://dx.doi.org/10.1007/3-540-60159-7_4

Kulczynski, S. 1927. Die Pflanzenassoziationen der Pieninen, in *Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences Mathematiques et Naturelles B*, 57–203.

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

Leung, Y. 1979. Locational choice: A fuzzy set approach, *Geographical Bulletin* 15: 28–34.

Mamdani, E. H.; Assilian, S. 1975. An experiment in linguistic synthesis with a fuzzy logic controller, *International Journal of Man-machine Studies* 7: 1–13. http://dx.doi.org/10.1016/S0020-7373(75)80002-2

Mennis, J.; Liu, J. Mining association rules in spatio-temporal data, *in Proceedings of the 7th International Conference on GeoComputation*, 8–10 September 2003, University of Southampton, United Kingdom.

Miller, H. J.; Han, J. 2001. Geographic data mining and knowledge discovery: An overview. In*:* Miller, H. J. & Han, J. (eds.) *Geographic Data Mining and Knowledge Discovery.* London: Taylor and Francis. http://dx.doi.org/10.4324/9780203468029_chapter_1

Mintz, D. 2012. *Technical assistance document for the reporting of daily air quality-the air quality index (AQI)*. US Environmental Protection Agency, Office of Air Quality Planning and Standards.

Ng, H.-F.; Fathoni, H.; Chen, I.-C. 2009. *Prediction of allergy symptoms among children in Taiwan using data mining*. Taipei Medical University Institutional Repository and ePublications.

Pipkin, J. S. 1978. Fuzzy sets and spatial choice, *Annals of the Association of American Geographers* 68: 196–204. http://dx.doi.org/10.1111/j.1467-8306.1978.tb01190.x

Rackemann, F. M. 1947. A working classification of asthma, *The American Journal of Medicine* 3: 601–606. http://dx.doi.org/10.1016/0002-9343(47)90204-0

Robinson, V. B. 1988. Some implications of fuzzy set theory applied to geographic databases, *Computers, Environment and Urban Systems* 12: 89–97. http://dx.doi.org/10.1016/0198-9715(88)90012-9

Romanet Manent, S.; Charpin, D.; Magnan, A.; Lanteaume, A.; Vervloet, D. 2002. Allergic vs nonallergic asthma: what makes the difference?, *Allergy* 57: 607–613. http://dx.doi.org/10.1034/j.1398-9995.2002.23504.x

Ross, T. J. 2009. *Fuzzy logic with engineering applications*. John Wiley & Sons.

Schneider, M. 2008. Fuzzy spatial data types for spatial uncertainty management in databases, *Handbook of Research on Fuzzy Information Processing in Databases* 2: 490–515.

Shua, H.; Zhub, X.; Daic, S. 2008. *Mining association rules in geographical spatio-temporal data*. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B2. Beijing.

Wackernagel, H. 2003. *Multivariate geostatistics*. Springer. http://dx.doi.org/10.1007/978-3-662-05294-5

YoussefAgha, A.; Jayawardene, W.; Lohrmann, D.; El Afandi, G. 2013. Application of data mining techniques to predict allergy outbreaks among elementary school children, *Journal of Communication and Computer* 10: 451–460.

Zöllner, I.; Weiland, S.; Piechotowski, I.; Gabrio, T.; Von Mutius, E.; 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: 545–548. http://dx.doi.org/10.1136/thx.2004.029561

Yousef KANANI SADAT. M.Sc. in GIS, Department of Surveying and Geomatics Engineering, College of Engineering, University of Tehran, Iran. B.Sc. in Surveying and Geomatics Engineering, College of Engineering, University of Tehran, Tehran, Iran. E-mail: *yousefkanani@ut.ac.ir.*

Research interests: spatial modelling, spatial data mining, volunteered geographic information (VGI), spatial cognition.

Tina NIKAEIN. B.Sc. in Surveying and Geomatics Engineering, Department of Surveying and Geomatics Engineering, College of Engineering, University of Tehran, Iran. E-mail: *tina_nikaein@ut.ac.ir.*

Research interests: spatial modelling, spatial data mining.

Farid KARIMIPOUR. PhD. in GIS, Assistant Professor, Department of Surveying and Geomatics Engineering, College of Engineering, University of Tehran, Iran. M.Sc. in GIS, College of Engineering, University of Tehran, Iran. B.Sc. in Surveying and Geomatics Engineering, College of Engineering, University of Tehran, Tehran, Iran. E-mail: *fkarimipr@ut.ac.ir.*

Research interests: spatial modelling, spatial data mining, volunteered geographic information (VGI), spatial cognition, multi-dimensional spatial analyses, 3D spatial data, moving objects, computational geometry.