INVESTIGATING THE RELATION BETWEEN PREVALENCE OF ASTHMATIC ALLERGY WITH THE CHARACTERISTICS OF THE ENVIRONMENT USING ASSOCIATION RULE MINING

Y. Kanani Sadat [a,*], F. Karimipour [a], A. Kanani Sadat[b]

[a] Dept. of Surveying and Geomatics Engineering, College of Engineering, University of Tehran, Iran - (yousefkanani, fkarimipr)@ut.ac.ir
[b] Dept. of Civil Engineering, Islamic Azad University, Ahar branch, Iran - kanani.ali66@yahoo.com

KEY WORDS: Spatial Association Rule Mining; Environmental Characteristics; Asthmatic Allergy; Air Pollution; Apriori

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 paper, 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. 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 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.

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 (Ng et al., 2009; Zöllner et al., 2005). 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, and the relation varies with time, the spatio-temporal data mining techniques seems very efficient in this regards.

Spatial data mining concerns development and application of novel computational techniques to analyze very large spatial databases (Buttenfield et al., 2001; Koperski et al., 1996). 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 (Miller and Han, 2001).

In this paper, we focus on the effect of air pollution on asthmatic allergies and investigate the relation between prevalence of such allergies with those characteristics of the environment that may affect the air pollution. The reside location 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), SO2 (sulfur dioxide), NO2 (nitrogen dioxide), PM10 and PM2.5 (particulate matter with a diameter of <10µm and <2.5µm, respectively), and O3 (ozone) as well as living distance to parks and roads, as major sources of asthmatic allergy irritants.

The rest of the paper is organized as follow: Section 2 surveys some of the previous work related to the topic of this paper, including data mining to study allergy prevalence; and spatial association rule mining to discover relations between spatially related parameters. Sections 3, introduce spatial association rule mining. In section 4, the components of the research methodology are described in details. The results for the case study are presented and discussed in Section 5. Finally, Section 6 contains concluding remarks and ideas for future research in this direction.

2. RELATED WORK

This section reviews the previous researches related to the topic of this paper, which are classified into: (i) using data mining techniques to study prevalence of allergic disease; and (ii) deploying spatial association rule mining to discover relations between spatially related parameters.

Data mining is an approach to determine the valid, novel, useful and understandable data patterns from huge amount of data stored in a database (Miller and Han, 2001). To the best of our

[*] Corresponding author.

This contribution has been peer-reviewed.
doi:10.5194/isprsarchives-XL-2-W3-169-2014

knowledge, three researches used data mining to study allergy outbreaks: 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. They deployed three predictive models: neural networks, decision trees and support vector machines (SVM). Akinbami et 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 SO2, NO2, O3 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. (2012) 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, SO2, NO2, PM10, PM2.5 and O3). The results of all of these researches are plausible. Nevertheless, none of them considered neither spatial nor temporal characteristics of data to study prevalence of allergy.

On the other hand, 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 north-eastern 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.

3. SPATIAL 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 (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, the rule like "the person who live in area with very high amount of NO2 and very high park effect, is infected with asthmatic allergy ", which is expressed as:

$$(NO_2, \text{very high}), (\text{park\_efct}, \text{very high}) \rightarrow (\text{asthmatic\_allergy}, \text{yes}) \quad (1)$$

If there is only one type of predicate (e.g., park_efct), the association rule is one-dimensional. Whereas, in multi-dimensional association rules, more than one type of predicate involves.

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:

$$\text{Support } (A \rightarrow B) = \text{prob}\{A \cup B\} \quad (2)$$

$$\text{Confidence } (A \rightarrow B) = \text{prob}\{B \mid A\} = (\text{prob}\{A \cup B\})/(\text{prob}\{A\}) \quad (3)$$

The association rules that have the minimum significant support and confidence are called strong association rules and are considered in decision making process. A common influential algorithm for the association rule mining is the so called Apriori algorithm (Agrawal and Srikant, 1994).

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:

$$A \rightarrow B \text{ [support, confidence, correlation]} \quad (4)$$

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

$$\text{Kulc } (A, B) = 1/2(P(A \mid B)+P(B \mid A)) \quad (5)$$

Which is a value between 0 and 1. A larger Kulc indicates stronger relation between A and B.

A spatial association rule contains at least one spatial relationship in an antecedent or consequent predicate (Koperski and Han, 1995). For example distance_to (road, near) is a spatial predicate that results in a spatial association rule. There are two important issues in dealing with spatial association rules:

- 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-processing 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).
- 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', 'far', and 'very far'.

4. RESEARCH METHODOLOGY

This paper analyze the risk of asthmatic allergy prevalence based on environmental characteristics through deploying the 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.

The case study is Tehran metropolitan area. Figure 1 illustrates the research methodology:

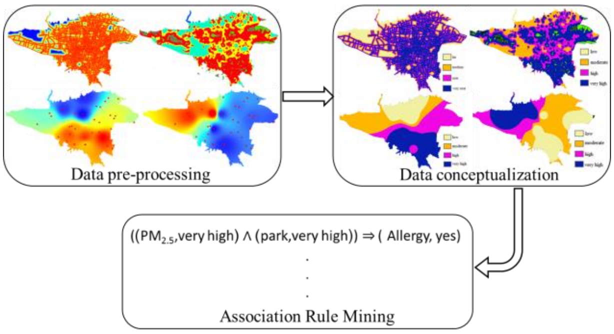


Figure 1. Research methodology

4.1 Data Pre-processing

The air pollution parameters consist of CO, SO_2, NO_2, PM_{10}, $PM_{2.5}$, and O_3 compiled hourly in December 2013 by Tehran's air pollution monitoring stations are used (Figure 2). This data is cleaned by filling the gaps 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).

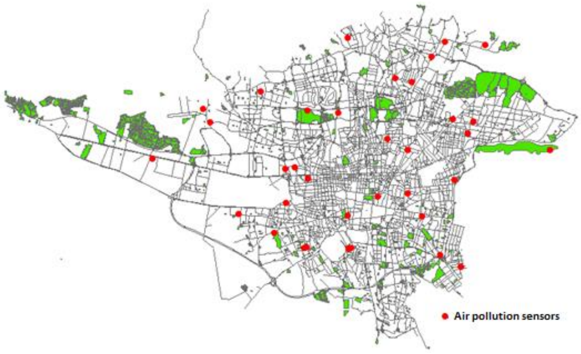


Figure 2. The map of Tehran's roads, parks and air pollution monitoring stations

To model the effect of distance to roads, a map is produced in which the distance to the nearest road is calculated. The same process was applied to model the effect of parks using the following equation, which quantifies the effect of nearby parks:

$$T_j = \sum \frac{A_i}{d_{ij}^2} \qquad (6)$$

where Tj = the effect of nearby parks for the point j
Ai = the area of the park i, and
dij = the distance of the park i from the point j

4.2 Data Conceptualization

As mentioned in Section 3, the inputs of association rule mining must be categorical values. Therefore, the data items assigned to the patients must be categorized. For this purpose, distance to roads was classified into "very near", "near", "medium" and "far" (Figure 3.a). The effect of parks also is classified into "very highly affected", "highly affected", "moderately affected" and "lowly affected" (Figure 3.b).

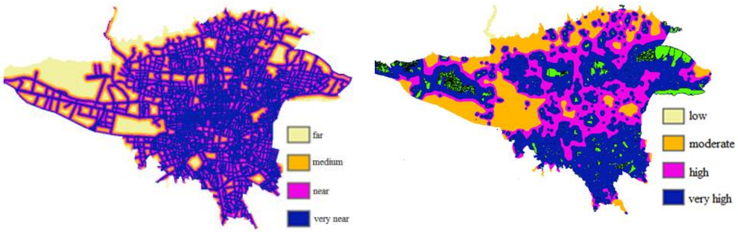
To categorize the air pollution parameters, the air quality index (AQI) – which is an indicator of air quality – is used. As the categorization breakpoints used by AQI varies from an air pollution parameter to another (Table 1), the following equation is used to normalize the measured values (Mintz,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

$$I_P = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}}(C_P - BP_{Lo}) + I_{Lo} \qquad (7)$$

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 AQI air pollution categories to "very high", "high", "moderate" and "low" (Figure 4).

Finally, the reside location of 1000 patients referred to the "Tehran Children's Medical Clinic" in December 2013 are places 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 classified air pollution and distance to roads and parks maps, the air pollution parameters and distance to roads and parks are assigned to each point as data items (attributes). Table 2 shows some of the recorded data in database.



(a) (b)

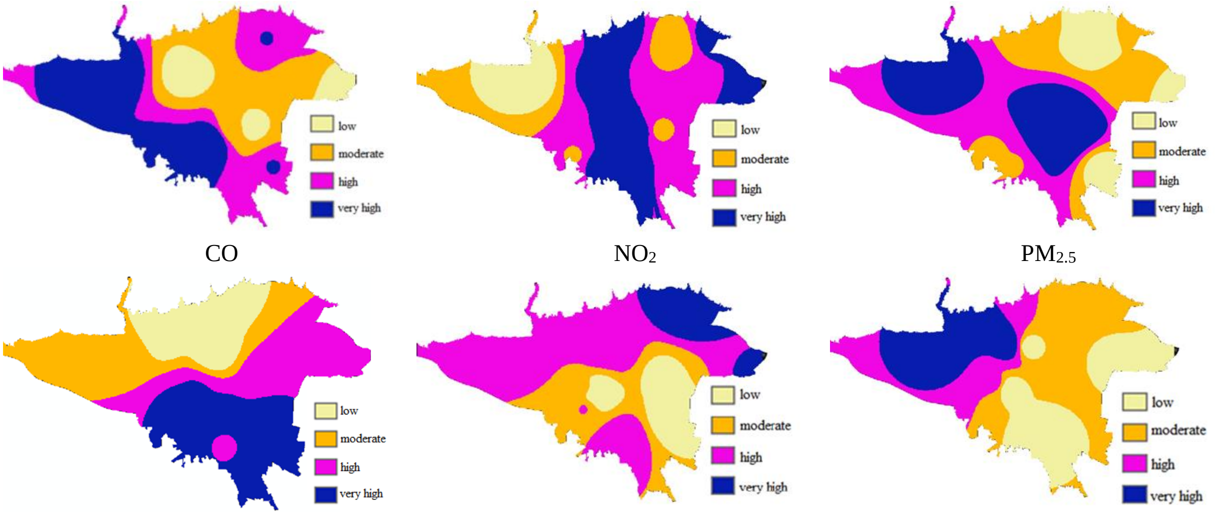
Figure 3. The classified maps to show the effect of distance to (a) roads and (b) nearby parks

Category	AQI	Breakpoints					
		NO2 (ppb)	SO2 (ppb)	CO (ppm)	$PM_{2.5}$ ($\mu g/m^3$)	PM_{10} ($\mu g/m^3$)	O3 (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	-

Table 1. Breakpoints for the AQI

ID	Park effect	Distance to road	O3	$PM_{2.5}$	PM_{10}	NO2	SO2	CO	asthma
1	very high	very near	moderate	moderate	very high	high	moderate	very high	Yes
2	moderate	near	low	very high	high	moderate	low	moderate	No
3	moderate	very near	moderate	moderate	high	moderate	very high	high	No
4	moderate	near	moderate	moderate	moderate	very high	very high	high	No
5	moderate	near	moderate	low	moderate	low	very high	very high	No
6	moderate	near	low	very high	high	very high	low	very high	Yes
7	high	medium	moderate	low	low	high	high	moderate	No
8	moderate	very near	moderate	low	low	low	high	moderate	No
9	moderate	very near	high	moderate	low	very high	high	moderate	No
10	moderate	medium	high	moderate	low	moderate	high	moderate	No
11	very high	very near	low	very high	very high	very high	high	high	Yes
12	moderate	near	high	moderate	low	moderate	high	moderate	No

Table 2. Part of the multidimensional dataset used in association rule



CO NO$_2$ PM$_{2.5}$

PM$_{10}$ SO$_2$ O$_3$

Figure 4. The categorized maps of different air pollutants in Tehran in December 2013

4.3 Spatial Association Rule Mining

Having the multidimensional dataset constructed, the association rules between asthmatic allergy prevalence and spatial characteristics of the environment (i.e., air pollution and distance to parks and roads) are extracted. 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)

In our case, the minimum support and confidence thresholds are respectively defined as 5% and 40%.

5. IMPLEMENTATION RESULTS

Applying the procedure described in Section 4 to the dataset, provided 60 association rules between prevalence of asthmatic allergy with characteristics of the environment in December 2013, some of which are illustrated in Table 3. For example, rule #5 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; And the *Kulczynski*'s correlation measure between the antecedent and the consequence is 64%.

Based on the extracted rules, distance to parks and roads as well as CO, NO_2, PM_{10} and $PM_{2.5}$ affect asthmatic 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 certify 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. .

ID	Association rules	Sup	Conf	Kulc
1	[($PM_{2.5}$, very high)] → (allergy, yes)	14.61	43.52	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 3. Some of the rules extracted for December through association rule mining

6. CONCLUSION

This paper deploys the 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 risk 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 classified data items (i.e., attributes) enabled to extract more reliable associations, as their interpretation certifies. As the air pollution conditions and pollen vary from time to time, the rules extracted for December may not be applicable to other months for two different months. Here, 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 may be taken into account.

REFERENCES

Agrawal, R., Imieliński, T., Swami, A., 1993. Mining association rules between sets of items in large databases, ACM SIGMOD Record. ACM, pp. 207-216.

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules, Proc. 20th Int. Conf. Very Large Data Bases, VLDB, pp. 487-499.

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.

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.

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

Buttenfield, B., M, M.G., H, H.M., Yuan, M., 2001. Geospatial Data Mining and Knowledge Discovery, Washington D.C., University Consortium for Geographic Information Science, White Paper on Emerging Research Themes.

Calargun, S.U., Yazici, A., 2008. Fuzzy Association Rule Mining from Spatio-temporal Data, Computational Science and Its Applications–ICCSA 2008. Springer, pp. 631-646.

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

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

Klosgen, W., May, M., 2002. Spatio-temporal subgroup discovery. KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE, 149-166.

Koperski, K., Adhikary, J., Han, J., 1996. Spatial data mining: Progress and challenges survey paper, ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 55–70.

Koperski, K., Han, J., 1995. Discovery of spatial association rules in geographic information databases, 4th International Symposium on Large Spatial Databases. Springer-Verlag, Berlin, pp. 47-66.

Kulczynski, S., 1927. Die Pflanzenassoziationen der Pieninen. 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.

Mennis, J., Liu, J., 2003. Mining association rules in spatio-temporal data, Proceedings of the 7th International Conference on GeoComputation.

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. Taylor and Francis, London, pp. 3–32.

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.

Piatetsky-Shapiro, G., 1991. Discovery, analysis and presentation of strong rules. Knowledge discovery in databases, 229-238.

Shekhar, S., Chawla, S., 2003. Spatial databases: a tour. prentice hall Upper Saddle River, NJ.

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 2008.

Srikant, R., Agrawal, R., 1996. Mining quantitative association rules in large relational tables, ACM SIGMOD Record. ACM, pp. 1-12.

Wackernagel, H., 2003. Multivariate geostatistics. Springer.

YoussefAgha, A., Jayawardene, W., Lohrmann, D., El Afandi, G., 2012. Application of Data Mining Techniques to Predict Allergy Outbreaks among Elementary School Children.

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.