Int. J. Environ. Sci. Technol. (2015) 12:3551–3562
DOI 10.1007/s13762-015-0880-4

ORIGINAL PAPER

Identification of air pollution patterns using a modified fuzzy co-occurrence pattern mining method

M. Akbari[1] · **F. Samadzadegan**[1]

Received: 28 January 2015 / Revised: 6 July 2015 / Accepted: 17 August 2015 / Published online: 4 September 2015

Abstract Spatio-temporal co-occurrence patterns represent subsets of object types which are located together in both space and time. Discovering spatio-temporal co-occurrence patterns is an important task having many application domains. There are a number of developed methods to mine co-occurrence patterns; however, using them needs a unique parameter to define the neighborhood. Identification of a unique optimum *k*-value or neighborhood radius is a challenging issue in different application domains. The developed method of this research defines a new fuzzy neighborhood and new fuzzy metrics to be applicable for real applications such as air pollution, especially when the researchers have no comprehensive knowledge regarding the application domain; in addition, it mines patterns based on the fuzzy nature of environmental phenomena. The new method mines patterns locally without localization step to speed up the mining process and considers all feature types (point, line and polygon) to handle all applications. Subsequently, it is applied to a real data set of Tehran city for air pollution to discover significant co-occurrence patterns of air pollution and influencing environmental parameters such as meteorological, topography and traffic. The case study results showed seven meaningful patterns among air pollution classes 2 and 3 and wind speed class 1, topography class 1 and traffic classes 1 and 2. The evaluation confirmed the accuracy and applicability of the new developed method for air pollution case. Furthermore, two performance tests for the method itself and a performance test against a crisp method were done, where the results exhibited an efficient computational performance.

Keywords Air pollution · Data mining · Co-occurrence pattern mining · Fuzzy · Tehran

Introduction

Air pollution leads to instability, harmful and undesirable effects in the environment (Goel et al. 2012). In the present century, urban areas deprive environment to have a healthy air quality due to the increase in concentration of tropospheric air pollutants such as SOx, NOx, CO and O3 (Venkanna et al. 2014). With the rapid growth of industrialization, population increase and disorderly urbanization, environmental pollution has become a significant area of concern (Dursun et al. 2015; Sakarde et al. 2014). Studies have shown that the air pollution in an urban area has a complex spatial pattern and levels can vary significantly over small distances (Sheng and Tang 2013). Next, it is important to consider air pollution, affecting parameters and also their interactive relation. Several techniques can be used to monitor air pollution data; one such technique is spatial data mining (Lanjewar and Shah 2012). The spatial data mining has been introduced to discover interesting and previously unknown, but potentially useful patterns from large spatial databases (Miller and Han 2009; Yoo and Bow 2011). The spatial data mining provides different methods including clustering, classification, co-location and association rules. This research is focused on co-location pattern mining. Spatial co-location patterns describe subsets of spatial features which are usually placed in close geographic proximity (Manikandan and Srinivasan 2012). In advance, spatio-temporal co-

✉ M. Akbari
moakbari@ut.ac.ir

[1] GIS Division, Department of Surveying and Geomatics Engineering, University College of Engineering, University of Tehran, North Kargar Ave., Tehran, Iran

occurrence pattern mining considers both space and time in the mining process (Celik et al. 2008; Huang et al. 2008; Qian et al. 2009a, b). Spatio-temporal co-occurrence patterns are useful to discover special characteristics behind co-located phenomena (Qian et al. 2009a). Discovering useful knowledge and information is a difficult task in a co-occurrence pattern mining process due to complexity of spatial data types, spatial relationships as well as time dependence of events (Wan and Zhou 2008).

Considering spatio-temporal co-occurrence pattern literature, it can be categorized in two classes: Some studies such as (Gudmundsson and Kreveld 2006; Vieira et al. 2009) have considered uniform groups of moving objects, and their methods discover flock patterns. These methods are not applied to general cases especially for applications with different feature types. The other researchers work with a mixed group of moving objects. Our case belongs to this category. Celik et al. (2008) proposed a new monotonic composite interest measure to mine mixed-drove spatio-temporal co-occurrence patterns. In addition, he proposed a new method (Celik 2011) to consider the presence period of the objects in the database for extracting partial spatio-temporal co-occurrence patterns. To cascade spatio-temporal patterns, Mohan et al. (2010) developed a new method to mine partially ordered subsets of event types whose instances are located together and occur in stages. Furthermore, there are a number of studies that generalize co-location patterns to the spatio-temporal domain (Huang et al. 2004; Shekhar et al. 2001) or the other studies that considered time factor as an alternative spatial dimension to mine spatio-temporal sequential patterns (Huang et al. 2008). These literatures do not consider a fully spatio-temporal problem. All the aforementioned studies tried to detect time-prevalent patterns, but none of these methods allows identifying how a spatio-temporal co-occurrence pattern evolves over time. Recently, authors have developed a crisp method (Akbari et al. 2015) to handle this lack. They proposed a new spatio-temporal measure to consider evolution of patterns simultaneously in space and time. However, there is still another problem that finding co-occurrence patterns occur in a spatial neighborhood of Pattern Core Element (PCEs) and it is usually difficult to find an optimum neighborhood radius in different applications. To neglect using a neighborhood radius in co-location mining, Wan et al. (2008) and Wan and Zhou (2008) proposed *k*-nearest feature technique to extract patterns. Although these methods need no unique neighborhood radius, there is another problem to find an optimal *k*-value. Based on this problem, Akbari and Samadzadegan (2014) proposed a new fuzzy neighborhood for pattern core elements which does not need to find an optimum unique neighborhood radius or *k*-value. However, this research only works for point data and mines only spatial co-location patterns without considering time.

In summary, to handle the aforementioned shortcomings of the literatures, this research presents a fuzzy spatio-temporal co-occurrence pattern mining method. Briefly, the contributions to this study can be explained as follows:

- As there are different feature types in real applications and also most of environmental phenomena have a fuzzy nature, moreover, the developed method defines a fuzzy neighborhood for different PCE's feature types.
- Using this fuzzy definition of neighborhood, this method reduces a localization step, and instead, it uses upper bound of neighborhood to index features. By eliminating a step in mining process, it reduces execution time.
- The method proposes new fuzzy metrics based on membership degree of spatial objects to PCE's neighborhood to identify prevalent co-occurrence patterns based on fuzzy definition of neighborhood. Against the existing metric, participation ratio, that is based on the number of objects in neighborhood of the PCEs, the new metric is based on the degree of belonging to neighborhood of the PCEs which can be more logical.
- As Tehran's air pollution is a big challenge, the proposed method was applied for a real case study, air pollution in Tehran, to evaluate its effectiveness and extract useful patterns to help urban decision makers.

The remaining sections of this paper are organized as follows: The materials and methods are described in Sect. "Materials and methods". The results and discussion are presented in Sect. "Results and discussion". Finally, the conclusions and future works as well as recommendations are summarized in Sect. "Conclusion and future work".

Materials and methods

Mining spatio-temporal co-occurrence patterns is an important geoprocessing task since it can be used by an extensive range of applications in different fields such as geographic information systems, geomarketing, traffic control, database exploration, image processing, environmental studies and other related fields (Priya et al. 2011). These extensive ranges of applications need methods which can handle all conditions. As mentioned before, some literatures consider only point data, some other do not consider time dimension and especially most of them develop crisp methods which are not suitable for fuzzy environment and its applications. Next, this research proposes a new method to handle the aforementioned

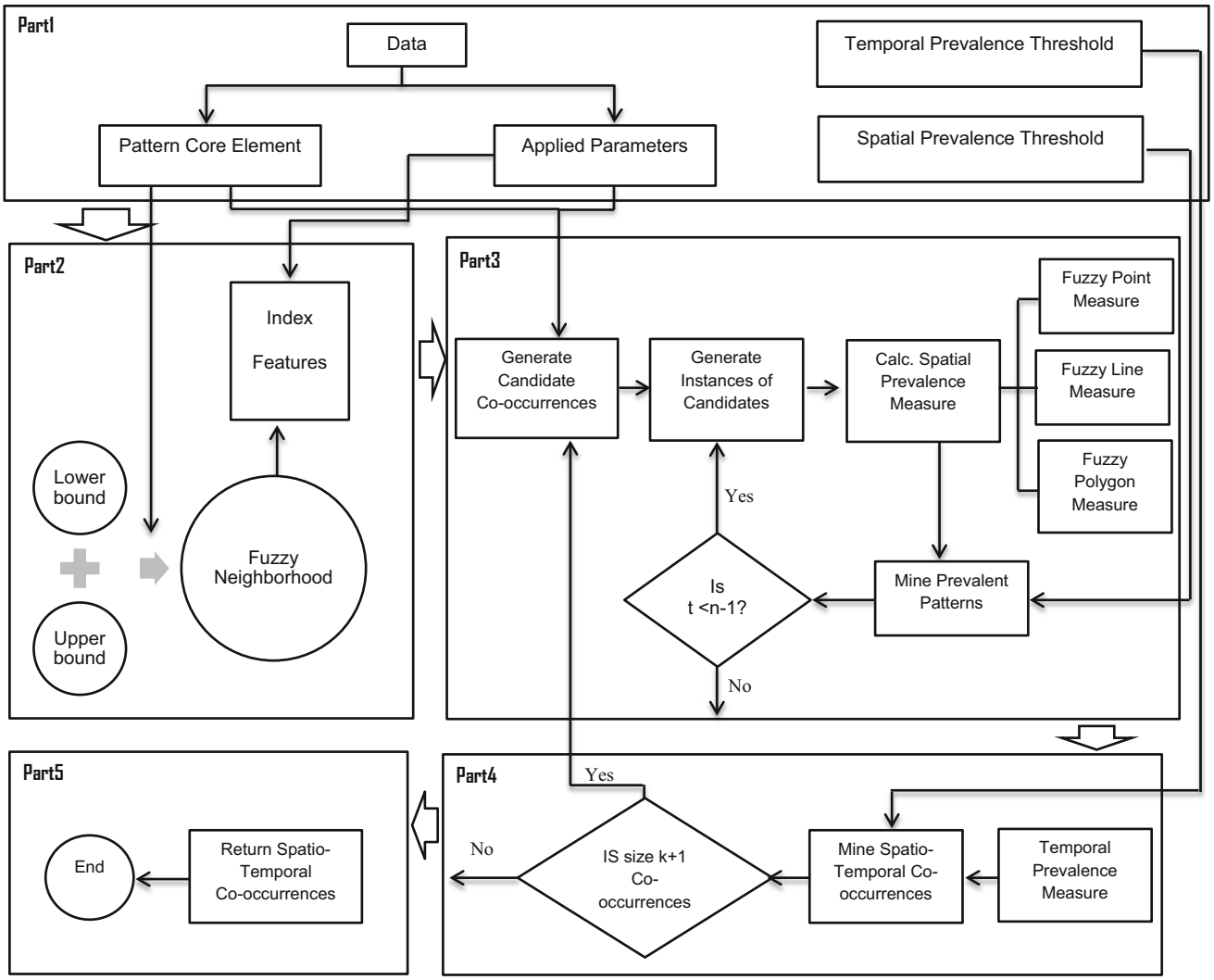

Fig. 1 The proposed flowchart of fuzzy spatio-temporal co-occurrence pattern mining method

problems. The proposed flowchart of spatio-temporal co-occurrence pattern mining method is shown in Fig. 1. It includes five main parts that each of them has some steps. The main steps in each of these parts are explained as follows.

Part 1 To start the mining process, it initializes parameters such as different thresholds and determines pattern core element.

Part 2 To eliminate the necessity of finding an optimal and unique neighborhood radius or k-value, this research considers a fuzzy neighborhood which can be defined based on a lower bound (R_1) and an upper bound (R_2). Figure 2 shows the neighborhood of different feature type of the PCEs. This step defines PCE's neighborhood by using initialized R_1 and R_2 values and then indexes all features participating in mining process to the buffer which was created by the upper bound. As mentioned in (Akbari et al. 2015), it is necessary to find co-occurrence patterns with a local view; in this regard, a Voronoi diagram as a spatial indexing structure was used. However, in this method, to increase execution efficiency, buffer of upper bound was used instead of Voronoi diagram. It is noteworthy that for the purposes of this paper, based on our case study, only the case of point PCEs will be considered.

Part 3 To mine different patterns, it is necessary to generate candidates. The method generates size-k+1 candidate co-occurrence patterns C_{k+1} based on all size-k patterns using an a priori-based method (Agarwal and Srikant 1994). To apply candidate patterns, the method similar to (Huang et al. 2004) generates the instances of candidate C_{k+1} by joining neighbor instances of size-k spatio-temporal co-occurrence patterns. Afterward, to discover important co-occurrence patterns, it is necessary to calculate some measures for pattern desirability evaluation such as participation ratio and participation index (Huang et al. 2004). These measures are only based on the number of features participate in a pattern; however, it will be more logical if these measures consider the dependency degree

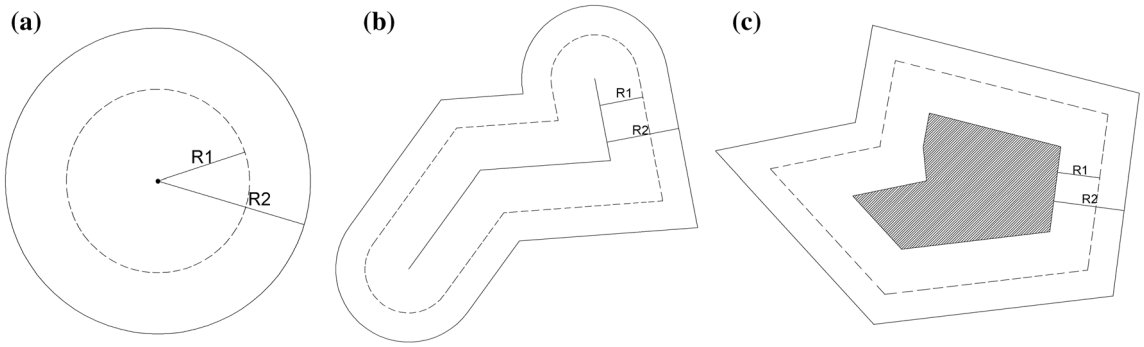

Fig. 2 Fuzzy neighborhood of different PCE feature types. **a** Point PCE, **b** Line PCE, **c** Polygon PCE

of a feature type to a pattern as well. Therefore, this research proposes the participation ratio in a new approach in order to consider not only the number of features participate in a co-occurrence pattern but also considers their dependency degree to a co-occurrence pattern. Figure 3 shows a general case of feature distribution around a point PCE and its neighborhoods.

As Fig. 3 shows, different feature types (point, line and polygon) can participate in a PCE neighborhood and co-occurrence. Therefore, the proposed participation ratio of this research is presented as follows:

- For point data:

$$\Pr(C,f_i) = \frac{\sum MF(f_i)}{N(f_i)} \tag{1}$$

where $MF(f_i)$ is the membership function of f_i feature instances in co-location instance neighborhoods of C and $N(f_i)$ is the total number of f_i feature instances (Akbari and Samadzadegan 2014). $MF(f_i)$ can be calculated using Eq. (2).

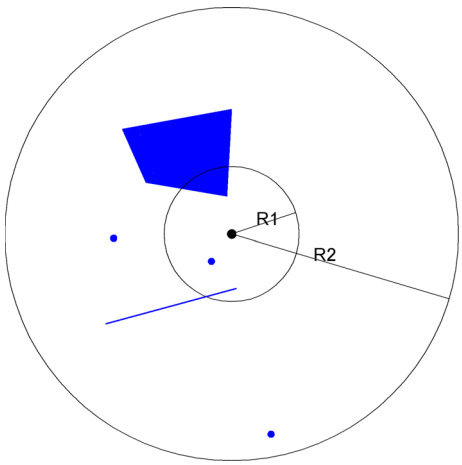

Fig. 3 Feature distribution around a point PCE

$$MF(f_i)\begin{cases} 1\ (\text{compeletly in neighborhood}) & \text{if } 0<X\le R_1 \\ 1-\dfrac{X-R_1}{R_2-R_1}\ (\text{partialy in neighborhood}) & \text{if } R_1<X\le R_2 \\ 0\ (\text{out of neighborhood}) & \text{if } R_2<X \end{cases} \tag{2}$$

where $MF(f_i)$ is the membership function which shows neighborhood relation value, R_1 and R_2 are lower and upper bounds of neighborhood and X is radial distance of a feature to PCE.

However, for other features besides point data, it is necessary to have some modification in the above equations. Considering Fig. 3, when the data are line or polygon, then it is complicated and computationally expensive to use Eq. (2) continuously. Next, to simplify it for line and polygon data, it is discretized. The neighborhood space between R_1 and R_2 is divided into five equidistant pieces. At last, $MF(f_i)$ is proposed as Fig. 4 and Eq. (3).

$$MF = \begin{cases} 1\ (\text{completely in neighborhood}) & \text{if } 0<X\le R_1 \\ 0.9\ (\text{partialy in neighborhood}) & \text{if } R_1<X\le d_1 \\ 0.7\ (\text{partialy in neighborhood}) & \text{if } d_1<X\le d_2 \\ 0.5\ (\text{partialy in neighborhood}) & \text{if } d_3<X\le d_3 \\ 0.3\ (\text{partialy in neighborhood}) & \text{if } d_4<X\le d_4 \\ 0.1\ (\text{partialy in neighborhood}) & \text{if } d_4<X\le R_2 \\ 0\ (\text{out of neighborhood}) & \text{if } R_2<X \end{cases} \tag{3}$$

where MF is the membership function which shows neighborhood relation value, R_1 and R_2 are lower and upper bounds of neighborhood, d_1, d_2, d_3, d_4 are radiuses for equidistance division of lower and upper distance and X is radial distance of a feature to PCE.

- For line data:

$$\mathrm{TMF}(f_i) = \sum\left(MF_j * LF_j\right) \tag{4}$$

$$\Pr(C,f_i) = \frac{\sum \mathrm{TMF}(f_i)}{N(f_i)} \tag{5}$$

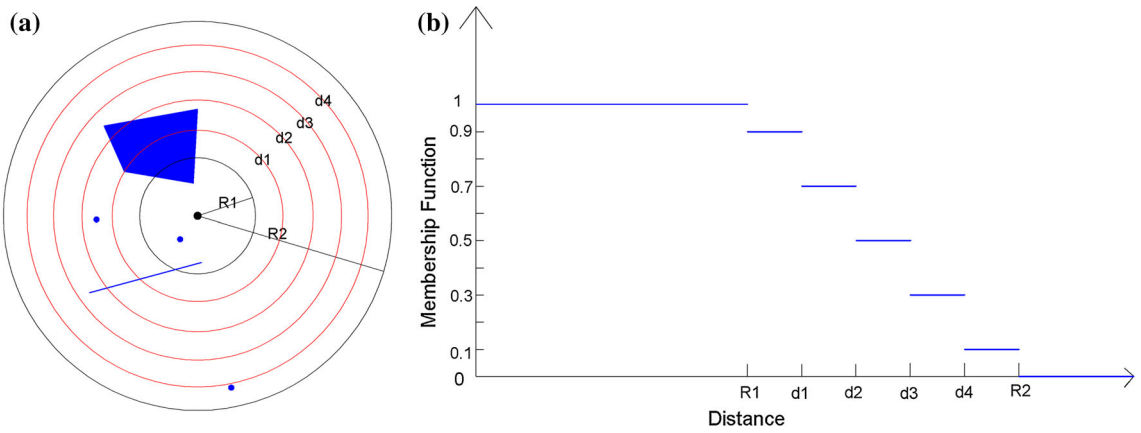

Fig. 4 Fuzzy definition of neighborhood **a** its illustration **b** its function

where MF_j is the partial membership function of f_i instances in neighborhoods of co-occurrence C, LF_j is the fraction of line feature length with a membership value MF_j, TMF(f_i) is the summation of fuzzy membership values of line feature parts and $N(f_i)$ is the total number of f_i feature instances.

- For polygon data:

$$TMF(f_i) = \sum (MF_j * AF_j) \tag{6}$$

$$\Pr(C, f_i) = \frac{\sum \mathrm{TMF}(f_i)}{N(f_i)} \tag{7}$$

where MF_j is the partial membership function of f_i instances in neighborhoods of co-occurrence C, AF_j is the fraction of polygon feature area with a membership value MF_j, TMF(f_i) is the summation of fuzzy membership values of polygon feature parts and $N(f_i)$ is the total number of f_i feature instances.

Then the participation index can be defined as the minimum of the participation ratios of all constitute feature types in a co-occurrence (Huang et al. 2004). In this step, the proposed method evaluates the calculated participation indices of candidate patterns against a spatial prevalence threshold (θ_s). Those patterns which their prevalence criteria are greater than the predefined threshold are considered as spatial prevalent patterns or important patterns and will be used to generate higher-level candidate patterns. *Part 4* To evaluate the discovered patterns of part 3 temporally, it is necessary to calculate a measure. This research uses temporal prevalence index (TPI) criterion as defined in (Akbari et al. 2015) in order to mine temporal prevalent patterns. In this step, the calculated TPI measure for spatial prevalent co-occurrence patterns is compared with a temporal prevalence threshold (θ_t). Next, those patterns which their TPI measure meets the temporal threshold are considered as spatio-temporal prevalent co-occurrence patterns. In addition, a classification for co-occurrence patterns has been presented in (Akbari et al. 2015), which categorizes different patterns based on their TPI to six classes: sustained, emerging, dispersing, time prevalent, time variant and no-pattern.

Results and discussion

To evaluate the proposed method, it was applied for a real case study to mine co-occurrence patterns of air pollution and the other environmental influencing parameters including road traffic, wind and topography.

Case study

The study area of this research is in city of Tehran, which is located in northern part of Iran (between 35.56–35.83 N and 51.20–51.61E). Tehran is bordered by the Alborz mountain range to the north, and it lacks permanent winds. As a result, smoke and other particulate materials cannot escape from the city. Air pollution in Tehran nearly similar to the other air polluted cities is primarily due to motor vehicles and heavily polluting industries (Dursun et al. 2015). Therefore, this area is affected by anthropogenic emissions, and a thick layer of particulate matter which is usually found in the atmosphere (Shad et al. 2009). Tehran's air pollution has been a significant challenge for several years, where this problem makes Tehran as one of the worst areas in the world for atmospheric pollution having many days exceeding air quality standards during each year (WHO 2011). Tehran's air pollution is responsible for thousands of deaths, and it costs millions of dollars each year (Environmental Protection Organization 2005). The data sets used include air pollution,

meteorology, traffic and topography of part of Tehran. The study area can be seen in Fig. 5. The study area consists of regions number 1–8, 21 and 22 of Tehran city. These regions were selected because first, data were available for these units and second, the data were up-to-date and had sufficient spatial and temporal overlap. The data span 12 days of different months between March 21, 2011, and March 19, 2012 (a Solar Hijri Year), one day per month (Akbari et al. 2015).

To use the case study data, they preprocessed to obtain the input data for the proposed method. Different data, as shown in Fig. 6, were classified into three classes to use directly in the mining process. The classification process for different data sets was conducted based on some existing knowledge and rules. We obtained the traffic data from Tehran Traffic Control Center; afterward, we used their classification for this data. Similarly, the rules for air pollution classification by Tehran Air Quality Control Center were accepted. Regarding meteorological parameters, since wind speed has an effective impact on air pollution, we used it and to classify it; moreover, the Beaufort scale was used similar to the Iran Meteorological Organization. Finally, for topography classification, due to nearly linear change in elevation from north to south of the city, three meaningful classes were formed. The last column in Fig. 6 presents the label and break points used for each class of data in the co-occurrence pattern mining process.

Experimental results

To use the proposed method of this research for co-occurrence pattern mining of Tehran air pollution, it needs some initial parameters as follows:

- Neighborhood lower bound R_1= 1500 m and neighborhood upper bound R_2= 2500 m; we used R_1 and R_2 to define a fuzzy neighborhood. In fact, as stated before, a fuzzy neighborhood was used to neglect determination of a precise neighborhood radius or k-value. These values were selected based on the average distance between air pollution measurement stations as well as a general knowledge about spatially meaningful variation in air pollution affecting parameters in a neighborhood.
- Spatial Prevalence Threshold: 0.5; as this threshold controls a pattern spatial importance criterion. Next, it was selected to supply at least for half of the participant features.
- Time Prevalence Threshold: 0.6; this threshold is about temporal importance; next, it is selected somewhat higher than 50 % such that only the most important patterns are found.

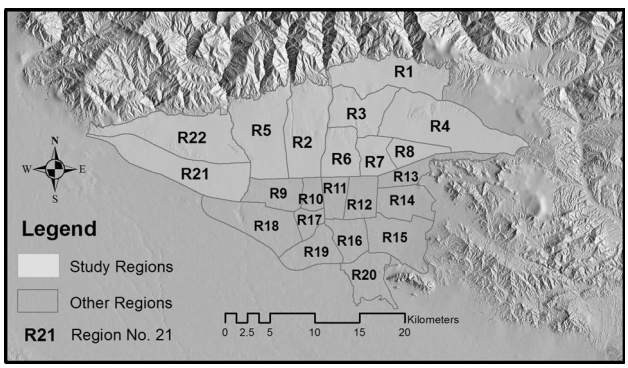

Fig. 5 Case study

Figure 7 shows the results of this evaluation. We used a similar representation method for the discovered patterns as proposed in (Akbari et al. 2015).

As a brief statistics of this experimental evaluation, it can be mentioned that there were 53 different pattern candidates. They produced 430 pattern combinations for 12 time slots. Among the assessed patterns, there were four spatio-temporal co-occurrence prevalent patterns {(Ap2, Tr1), (Ap3, Tp1), (Ap3, Wn1) and (Ap3, Tp1, Wn1)} and three time-variant co-occurrence patterns (Akbari et al. 2015) {(Ap2, Tr2), (Ap2, Wn1) and (Ap2, Wn1, Tr1)}. The results of this proposed method evaluation confirm the previous findings by a crisp method (Akbari et al. 2015). As it can be seen from Fig. 7, there is a combination of studied parameters in the discovered patterns. In the following section, we try to discuss the results' similarities and dissimilarities of this proposed method against the crisp method (Akbari et al. 2015). First, the similar trends between these two developed method results have been explained:

- The extracted patterns are only regarding the air pollution classes 2 and 3. This was expected to have co-occurrence patterns with higher levels of air pollution class 2 and 3 because (Azizi 2011; Kavousi et al. 2013; Saadatabadi et al. 2012) have mentioned high air pollution as a major problem for Tehran.
- Prevalent patterns between air pollution and traffic were expected because in this research CO was used as air pollutant and as (Azizi 2011; Safavi and Alijani 2006) stated, CO is a traffic-related pollutant and air quality deterioration is frequently related to traffic emissions (Silva et al. 2014).
- Tehran that has been placed in the southern parts of Alborz mountain range (Safavi and Alijani 2006) due to its physical conditions and topography lies in a valley and a semi-closed area (Fig. 5). These mountains prevent the flow of humid wind through the capital and trap air pollution over the city (Naddafi et al. 2012). Next, as Saadatabadi et al. (2012) explained, low wind speed is one of the most important meteorological

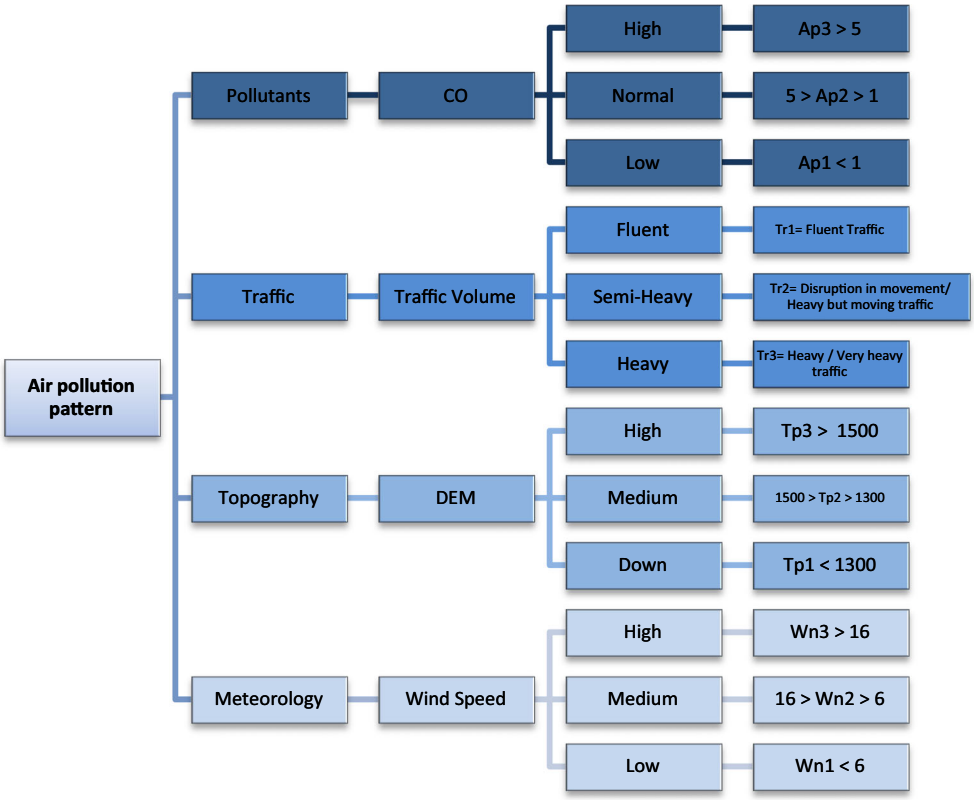

Fig. 6 Applied parameters of this research

factors causing high levels of CO concentration; therefore, the extracted patterns between high air pollution and low wind speed are meaningful.

- As temperature inversion is a usual case for Tehran, its occurrences further trap pollutants in the city, especially as (Rahimi Ghoroghi 2012; Safavi and Alijani 2006) stated in central and southern parts of Tehran, where the topography class is 1 (i.e., low topography). Therefore, the extracted pattern of air pollution class 3 (high) and topography class 1 (low) is a valid pattern.

However, the achieved results of the proposed method of this research have a difference with the crisp method (Akbari et al. 2015) in the following case:

- As it can be seen in Fig. 7 and also the result explanations in Sect. "Experimental results", the number of prevalent co-occurrence patterns against the crisp method (Akbari et al. 2015) has been reduced. In this experiment, four prevalent patterns and three time-variant patterns were extracted, while with the same data, the crisp method (Akbari et al. 2015) mined six prevalent patterns and 11 time-variant patterns. The reason of this reduction can be explained by the new proposed method. In this research, a new participation ratio concept based on fuzzy membership of features to the neighborhood has been developed that against the existing methods, it calculates participation ratio precisely based on the degree of fuzzy membership of features. This method causes less value of participation ratio and thus fewer numbers of discovered patterns. In other words, using this new developed criterion, a deep filtration of candidate patterns has been done and only the most important patterns can be discovered.

Performance results

To assess the proposed method performance and also to check its sensitivity, some experiments have been done.

First, we evaluated it for different neighborhood radiuses. Regarding this experiment, we changed R_1 from 500 to 2500 m and R_2 from 1000 to 3000 m and test different combinations of neighborhood radiuses to check processor time during execution, while we left all other parameters unchanged (number of time slots = 12, spatial prevalence threshold = 0.5, temporal prevalence threshold = 0.6). Figure 8 shows the results of this evaluation.

Next, two different regression models were fitted. A quadratic model is shown in Fig. 8a, and a linear model is shown in Fig. 8b, c, d. Figure 8 shows a quadratic regression if we consider all different cases of changing neighborhood radiuses R_1 and R_2 together. However, it will be more logical

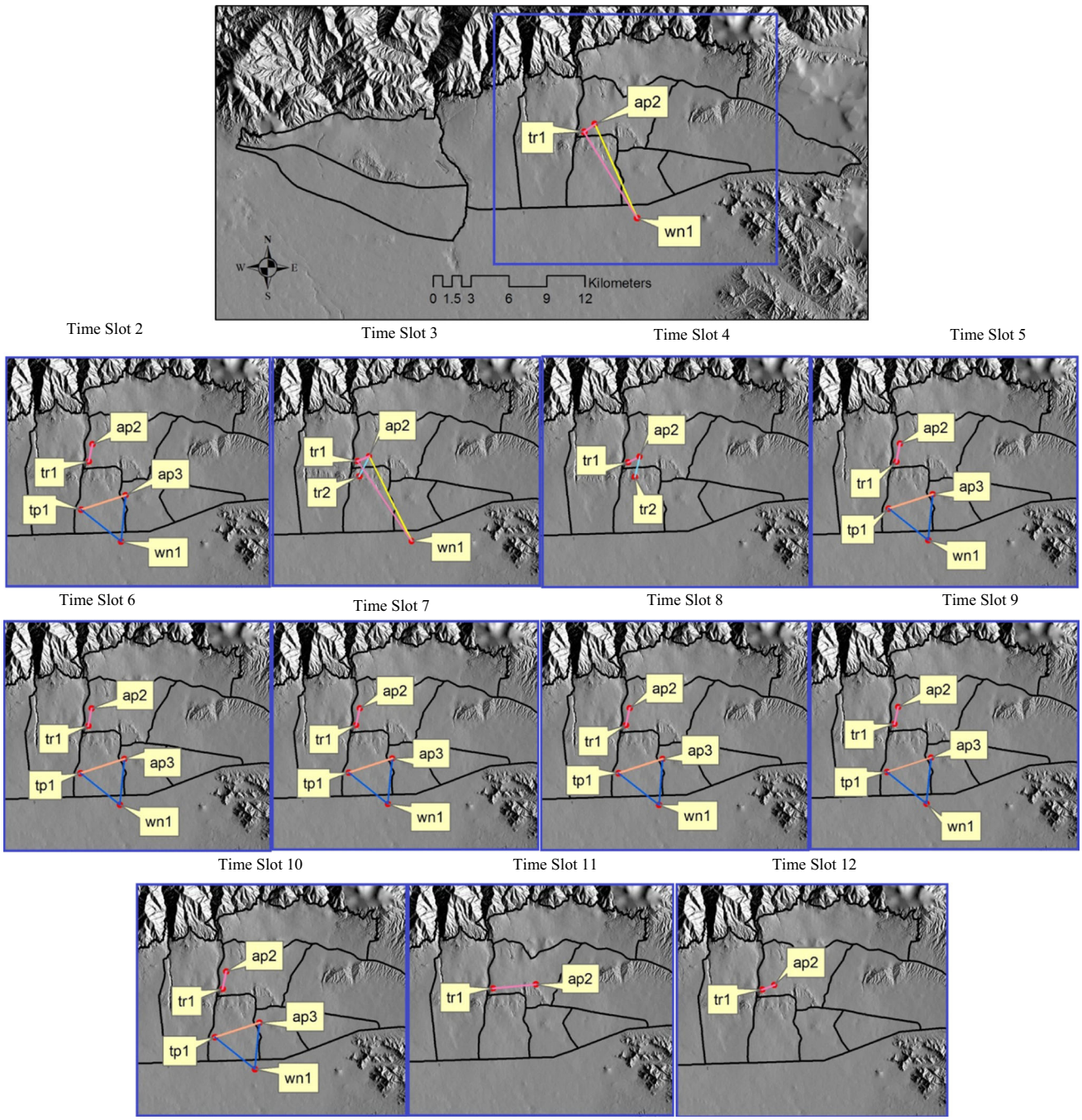

Fig. 7 Evaluation results of developed method for case study

if we consider performance trend of the proposed method when one of neighborhood radiuses is changing and the other one is unchanged. Therefore, we presented it by parts of (b), (c) and (d) of Fig. 8 when the performance trend of the method shows a linear regression. Next, it can be concluded that the method has an $O(n^2)$ performance trend if both R_1 and R_2 change simultaneously that is a special case, but it has an $O(n)$ performance trend if one of neighborhood radiuses changes each time. Afterward, we compared the efficiency of the proposed method of this research with a crisp method (Akbari et al. 2015). Figure 9 shows processor time for different cases, where "Crisp method" stands for the proposed method in (Akbari et al. 2015) which has neighborhood radius $R = 2500$ for this test; moreover, "Case 1" is the developed method of this research for $R_1 = 500$ and $R_2 = 2500$, "Case 2" is for $R_1 = 1000$ and $R_2 = 2500$, "Case 3" is for $R_1 = 1500$ and $R_2 = 2500$ and "Case 4" is for $R_1 = 2000$ and $R_2 = 2500$.

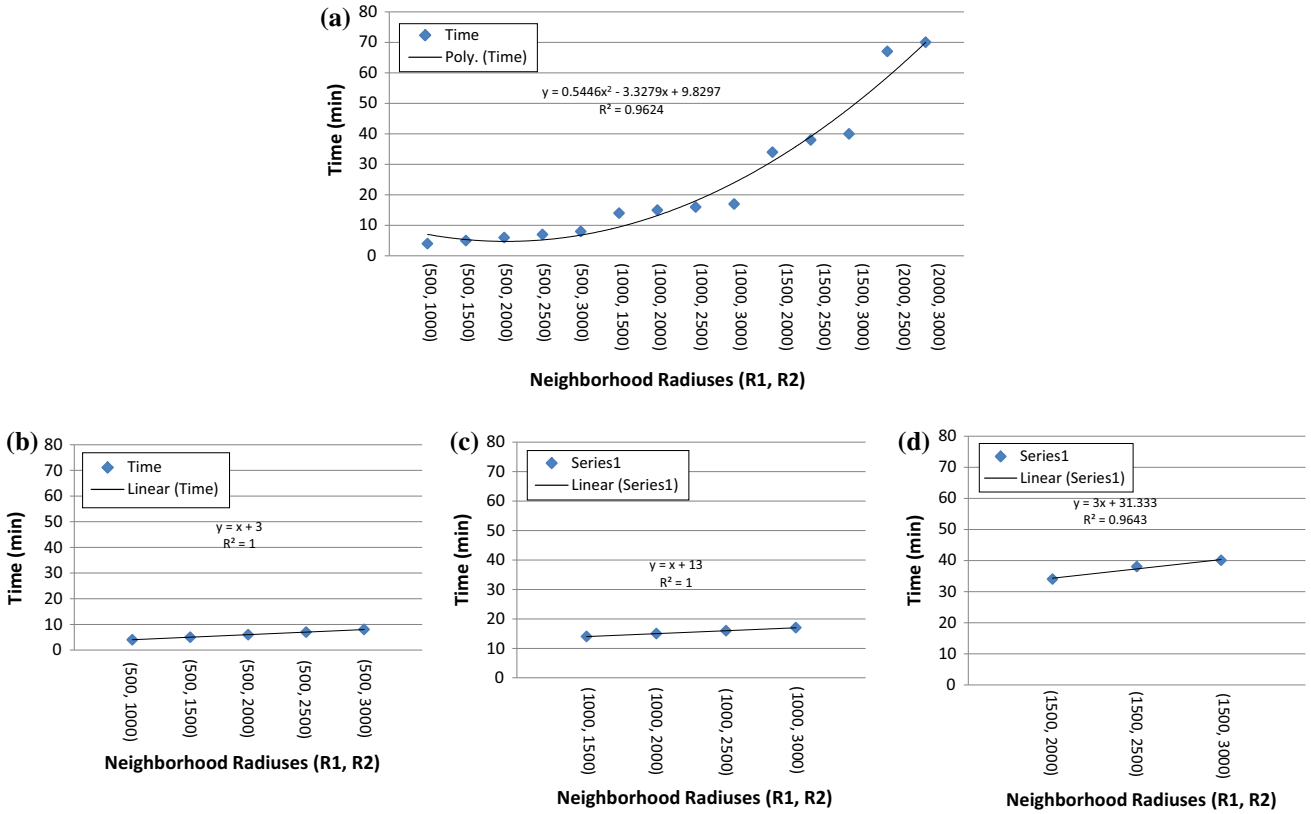

Fig. 8 Execution performance with different neighborhood radiuses

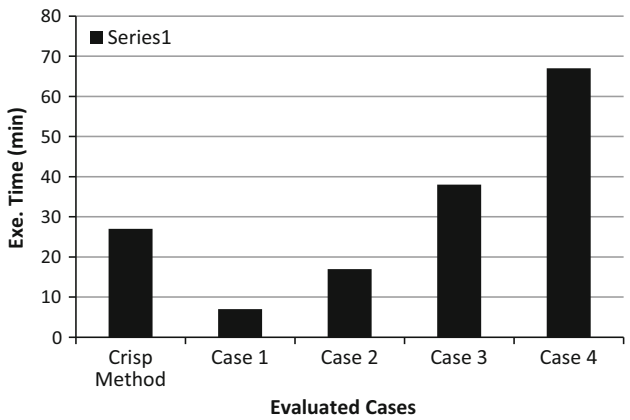

Fig. 9 Performance tests of the proposed fuzzy method against the crisp method of (Akbari et al. 2015)

As it is evident in Fig. 9, it can be concluded that the performance of different cases of the developed method of this research varies based on different neighborhood radiuses although the $R_2 = 2500$ is the same for all of them. Perhaps, it is not meaningful to compare the crisp method performance against different cases of the proposed method of this research since there are some differences in neighborhood definition and parameters, where in the crisp method, there is only one neighborhood radius, but in the new proposed method, we have two neighborhood radiuses defining a fuzzy neighborhood. The proposed fuzzy method exhibits a better performance compared to the crisp method where its lower bound is {500, 1000} and it exhibits worse performance where the lower bound is {1500, 2000}. In these experiments, the upper bound is fixed and is 2500 similar to the neighborhood radius of the crisp method. It means that as the inner neighborhood region increases, the performance of the proposed method decreases. In addition, by assessing the results of this figure, it can be concluded although the new proposed method uses a more extensive computational participation ratio metric based on the new definition of neighborhood, but it does not show a significant change in execution time. It can be explained by this fact that though the participation ratio computation is more expensive compared to the crisp method (Akbari et al. 2015), but in the new developed method, the localization step has been eliminated by creating Voronoi diagram and indexing features to it which can speed up the mining process. Therefore, it could make a balance in execution time for the new developed method against the crisp one. Next, to evaluate the model results against different break points for analyzed parameters of this research, eight different scenarios have been defined as presented in Table 1, so that in each scenario, parameter classification break points have been changed and the others

Table 1 Different scenario conditions

Scenarios	Changed parameter	Changed values	No. of prevalent patterns	Prevalent patterns
Main	–	–	4	(ap2, tr1), (ap3, tp1), (ap3, wn1), (ap3, tp1, wn1)
Scenario 1	Air pollution	Ap1 < 0.5 0.5 < Ap2 < 2.5 Ap3 > 2.5	1	(ap3; tr1)
Scenario 2	Air pollution	Ap1 < 2.5 2.5 < Ap2 < 6 Ap3 > 6	1	(ap1; tr1)
Scenario 3	Topography	Tp1 < 1150 1150 < Tp2 < 1350 Tp3 > 1350	3	(ap2; tr1), (ap3, tp3), (ap3; wn1)
Scenario 4	Topography	Tp1 < 1450 1450 < Tp2 < 1650 Tp3 > 1650	2	(ap2; tr1), (ap3; wn1)
Scenario 5	Wind speed	Wn1 < 2 2 < Wn2 < 8 Wn3 > 8	3	(ap2; tr1), (ap3; tp1), (ap2; wn2)
Scenario 6	Wind speed	Wn1 < 9 9 < Wn2 < 20 Wn3 > 20	5	(ap2; tr1), (ap3; tp1), (ap3; wn1), (ap2; wn1), (ap3; tp1; wn1)
Scenario 7	Traffic volume	Tr1 = Fluent traffic/disruption in movement Tr2 = Heavy but moving traffic/heavy Tr3 = Very heavy traffic	5	(ap2; tr1), (ap3; tr1), (ap3; tp1), (ap3; wn1), (ap3; tp1; wn1)
Scenario 8	Traffic volume	Tr1 = Fluent traffic Tr2 = Disruption in movement Tr3 = Heavy but moving/heavy/very heavy traffic	4	(ap2; tr3), (ap3; tp1), (ap3; wn1), (ap3; tp1; wn1)

remained unchanged. As stated, in this study, only the break points of each parameter for different scenarios were changed, but the number of classes is the same as the main scenario as showed in Fig. 6, since changing the number of classes can make a complicated problem and comparing scenarios with the main scenario is not possible.

Based on Table 1, when we change the air pollution parameter, then the discovered patterns show the highest degree of change against the main scenario. This change is expected because air pollution is the core element parameter in the developed co-occurrence pattern mining method, and therefore, it can create more changes in the discovered patterns. In addition, the other point in this table is that whenever we change each of parameter break points, then the extracted prevalent patterns related to that parameter will change against the main scenario prevalent patterns, which is comprehensible as well. In addition, a performance test for different scenarios against the main scenario has been done as shown in Fig. 10. For this test, processor time is measured as a performance measure.

The result of this test reveals that most of the scenarios have nearly similar execution time against the main scenario

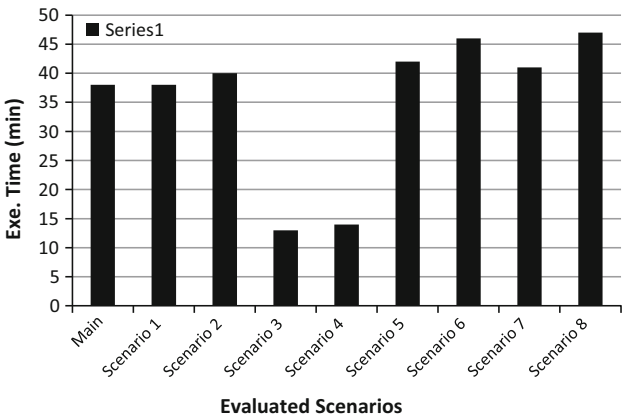

Fig. 10 Performance test of different scenarios

and they have processor time between 38 and 47 min, except scenarios 3 and 4 which have 13- and 14-min execution time, respectively. Scenarios 3 and 4 are related to changing topography break points. As these data are polygon-type data in this study, it means changing the break points can change the number and their area which can have an important impact on execution time of the method.

Conclusion and future work

Air pollution is a significant challenge for urban decision makers and a threat for human health and safety. As air pollution is a multidisciplinary problem, it needs to be assessed and studied from different perspectives. Extracting air pollution patterns regarding environmental affecting parameters can reveal significant information in order to handle this problem. The existing methods due to their shortcomings do not consider fuzzy nature of environmental phenomena, moreover, relying on a unique parameter to define neighborhood and using traditional crisp criteria to mine significant patterns. Therefore, in this research, a new co-occurrence pattern mining method has been developed so that: First, the new proposed method defines a fuzzy neighborhood for pattern core elements, which eliminates the difficulties of finding a unique optimum neighborhood radius or *k*-value. Second, based on the new fuzzy definition of neighborhood, the proposed method eliminates the need of localization step which speeds up the process and increases computation efficiency. Third, to mine prevalent co-occurrence patterns in a fuzzy framework, new fuzzy metrics have been developed to work based on fuzzy nature of environmental phenomena. The proposed method of this research was applied to a real case study in order to extract significant patterns of air pollution and the other influencing parameters for the city of Tehran. The results of evaluation revealed that it is suitable and applicable for real applications such as air pollution, especially when the researchers have no comprehensive knowledge about the application domain of applying pattern mining. Applying the proposed method on case study leads to seven meaningful patterns {(Ap2, Tr1), (Ap3, Tp1), (Ap3, Wn1), (Ap3, Tp1, Wn1), (Ap2, Tr2), (Ap2, Wn1) and (Ap2, Wn1, Tr1)}. To evaluate the method performance, three different experiments have been done. In these experiments, the proposed method is evaluated by changing neighborhood radiuses and by changing parameter breaks and is compared with a crisp method. The results confirmed an acceptable execution performance. For future works, we intend to apply or extend the proposed method of this research for different application domains such as noise pollution, crime, floods or car accidents. In addition, we would like to extend a more comprehensible visualization method to illustrate co-occurrence patterns in a spatio-temporal framework and show the fuzzy concept of environmental parameters in the extracted patterns.

Acknowledgments We are appreciating Prof. Robert Weibel for his helpful comments on this manuscript, and also, we are grateful to the Iran Meteorological Organization, the Tehran Air Quality Control Center, the Tehran Traffic Control Center and the National Cartographic Center of Iran for providing our case study data.

References

Agarwal R, Srikant R (1994) Fast algorithms for Mining Association Rules. In: Proceeding of 20th international conference on very large data bases (VLDB), pp 487–499

Akbari M, Samadzadegan F (2014) New regional co-location pattern mining method using fuzzy definition of neighborhood. ACSIJ 3(9):32–37

Akbari M, Samadzadegan F, Weibel R (2015) A generic regional spatio-temporal co-occurrence pattern mining model: a case study for air pollution. J Geogr syst. 17(3):249–274. doi:10.1007/s10109-015-0216-4

Azizi MH (2011) Impact of traffic-related air pollution on public health: a real challenge. Arch Iran Med 14(2):139–143

Celik M (2011) Discovering partial spatio-temporal co-occurrence patterns. In: Proceeding of 1st international conference on spatial data mining and geographical knowledge services, pp 116–120, Fuzhou, China. doi:10.1109/ICSDM.2011.5969016

Celik M, Shekhar S, Rogers JP, Shine JA (2008) Mixed-drove spatiotemporal co-occurrence pattern mining. IEEE Trans Knowl Data Eng 20(10):1322–1335. doi:10.1109/TKDE.2008.97

Dursun S, Kunt F, Taylan O (2015) Modelling sulphur dioxide levels of Konya city using artificial intelligent related to ozone, nitrogen dioxide and meteorological factors. Int J Environ Sci Technol 1–14. doi:10.1007/s13762-015-0821-2

Environmental Protection Organization (2005) Census air polluters in Tehran, 1999–2003

Goel A, Ray S, Agrawal P, Chandra N (2012) Air pollution detection based on head selection clustering and average method from wireless sensor network. In: Second international conference on advanced computing & communication technologies, pp 434–438

Gudmundsson J, Kreveld MV (2006) Computing longest duration flocks in trajectory data. In: Proceeding of the ACM international symposium on geographic information systems. Virginia, USA, pp 35–42. doi:10.1145/1183471.1183479

Huang Y, Shekhar S, Xiong H (2004) Discovering colocation patterns from spatial datasets: a general approach. IEEE Trans Knowl Data Eng 16(12):1472–1485

Huang Y, Zhang L, Zhang P (2008) A framework for mining sequential patterns from spatio-temporal event datasets. IEEE Trans Knowl Data Eng 20(4):433–448. doi:10.1109/TKDE.2007.190712

Kavousi A, Sefidkar R, Alavimajd H, Rashidi Y, Khonbi ZA (2013) Spatial analysis of CO and PM10 pollutants in Tehran city. JPS 4(3):41–50

Lanjewar UM, Shah JJ (2012) Air pollution monitoring & tracking system using mobile sensors and analysis of data using data mining. Intern J Adv Comput Res 2(4):19

Manikandan G, Srinivasan S (2012) Mining spatially co-located objects from vehicle moving data. Eur J Sci Res 68(3):352–366

Miller HJ, Han J (2009) Geographic data mining and knowledge discovery, 2nd edn. CRC Press, London, p 486

Mohan P, Shekhar S, Shine JA, Rogers JP (2010) Cascading spatiotemporal pattern discovery: a summary of results. In: Proceeding of the SIAM international conference on data mining (SDM), pp 327–338

Naddafi K, Hassanvand MS, Yunesian M, Momeniha F, Nabizadeh R, Faridi S, Gholampour A (2012) Health impact assessment of air pollution in megacity of Tehran, Iran. Iran J Environ Health Sci Eng 9(1):9–28

Priya G, Jaisankar N, Venkatesan M (2011) Mining co-location patterns from spatial data using rulebased approach. Int J Glob Res Comput Sci 2(7):58–61

Qian F, He Q, He J (2009a) Mining spread patterns of spatio-temporal co-occurrences over zones. In: Proceedings of the international conference on computational science and applications, pp 686–701. doi:10.1007/978-3-642-02457-3_57

Qian F, Yin L, He Q, He J (2009b) Mining spatio-temporal co-location patterns with weighted sliding window. IEEE Int Conf Intell Comput Intell Syst ICIS 2009:181–185. doi:10.1109/ICICISYS.2009.5358192

Rahimi Ghoroghi N (2012) Evaluation of geographical factors on Tehran air pollution and its relation with temperature inversion. First Conference of air and noise pollution management. Tehran, Iran. http://www.civilica.com/Paper-CANPM01-CANPM01_039.html. (In Persian)

Saadatabadi AR, Mohammadian L, Vazifeh A (2012) Controls on air pollution over a semi-enclosed basin, Tehran: a synoptic climatological approach. IJST 4:501–510

Safavi SY, Alijani B (2006) Evaluation of geographical parameters in Tehran air pollution. Geogr Res J 58:99–112 **(In Persian)**

Sakarde S, Choudhari M, Gode S (2014) Implementation of WSN based air pollution monitoring system using data mining technique. IJCSMC 3(6):790–795

Shad R, Mesgari MS, Shad A (2009) Predicting air pollution using fuzzy genetic linear membership kriging in GIS. Comput Environ Urban Syst 33(6):472–481

Shekhar S, Huang Y, Xiong H (2001) Discovering spatial co-location patterns: a summary of results. In: Proceeding of 7th international symposium on spatial and temporal databases (SSTD), Redondo Beach, CA, USA. doi:10.1007/3-540-47724-1_13

Sheng N, Tang UW (2013) Risk assessment of traffic-related air pollution in a world heritage city. Int J Environ Sci Technol 10(1):11–18

Silva LT, Pinho JL, Nurusman H (2014) Traffic air pollution monitoring based on an air–water pollutants deposition device. Int J Environ Sci Technol 11(8):2307–2318

Venkanna R, Nikhil GN, Rao TS, Sinha PR, Swamy YV (2014) Environmental monitoring of surface ozone and other trace gases over different time scales: chemistry, transport and modeling. Int J Environ Sci Technol 12(5):1749–1758

Vieira MR, Bakalov P, Tsotras VJ (2009) On-line discovery of flock patterns in spatio-temporal data. In: GIS '09 proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, pp 286–295. doi:10.1145/1653771.1653812

Wan Y, Zhou J (2008) KNFCOM-T: a k-nearest features-based co-location pattern mining algorithm for large spatial data sets by using T-trees. Int J Bus Intell Data Min 3(4):375–389. doi:10.150/IJBIDM.2008.022735

Wan Y, Zhou J, Bian F, (2008) CODEM: a novel spatial co-location and de-location patterns mining algorithm. In: Fifth international conference on fuzzy systems and knowledge discovery (FSKD'08), vol 2, pp 576–580

WHO (August 2011). Urban outdoor air pollution database. Geneva, Switzerland, Department of Public Health and Environment, World Health Organization; 2011. http://www.who.int/phe

Yoo JS, Bow M (2011). Mining top-k closed co-location patterns. In: Proceeding of IEEE international conference on spatial data mining and geographical knowledge services (ICSDM), Fuzhou, pp 100–105. doi:10.1109/ICSDM.2011.5969013