

New Regional Co-location Pattern Mining Method Using Fuzzy Definition of Neighborhood

Mohammad Akbari¹, Farhad Samadzadegan²

¹ PhD Candidate of Dept. of Surveying & Geomatics Eng., University of Tehran,
Tehran, Iran
moakbari@ut.ac.ir

² Professor of Dept. of Surveying & Geomatics Eng., University of Tehran,
Tehran, Iran
samadz@ut.ac.ir

Abstract

Regional co-location patterns represent subsets of object types that are located together in space (i.e. region). Discovering regional spatial co-location patterns is an important problem with many application domains. There are different methods in this field but they encounter a big problem: finding a unique optimum neighborhood radius or finding an optimum k value for nearest neighbor features. Here, we developed a method that considers a neighborhood interval using fuzzy definition of neighborhood. It is easier to apply the proposed method for different applications. Also, this method mine regional patterns using a local tessellation (Voronoi Diagram) and finds patterns with a core feature. To test our method we used a synthetic data set and compared developed method with a naïve approach. The results show that the proposed method is more applicable and efficient.

Keywords: *co-location, pattern mining, fuzzy, neighborhood, regional.*

1. Introduction

Spatial data mining has been introduced for discovering interesting and previously unknown, but potentially useful patterns from large spatial databases [9], [16]. Spatial co-location patterns describe subsets of spatial features that are usually placed in close geographic proximity [8]. Spatial data mining has a wide range of applications in different fields, such as geographic information systems, geo-marketing, traffic control, database exploration, image processing, environmental studies etc. [10]. Spatial co-location pattern mining is one of the most important techniques of spatial data mining. It has recently been used for mining the spatial dependencies of objects in different applications [3], [14]. Extracting interesting and useful patterns from spatial data sets is more difficult than extracting the corresponding patterns from traditional numerical and categorical data due to the complexity of spatial data types, spatial relationships, spatial autocorrelation and time dependence of events [12]. Because of these local spatial relationships and the spatial

autocorrelation between objects, spatial co-location patterns have regional properties; therefore, methods that consider this condition in their process will yield more realistic results. Different algorithms have been proposed in spatial co-location pattern mining. In section 2, we will review the relevant works and identify shortcomings. In this research we extended existing methods to present a more capable co-location mining method:

- we use the Voronoi Diagram, to speed up the mining process;
- we find co-occurrence patterns with an emphasis on a so-called Pattern Core Element (PCE), to respond to some applications that require special attention to particular patterns;
- we extend our algorithm so that it considers a fuzzy neighborhood in mining process, to eliminate necessity of finding a unique optimum neighborhood radius in different applications.

The organization of the paper is as follows: the review of related research is given in Section 2. The proposed method is described in Section 3 and the results and discussion are described in Section 4. The conclusions and perspectives on future work are summarized in Section 5.

2. Literature review

Various researchers have focused on applying and extending methods for spatial co-location patterns mining for applications in different areas. Some work focused on global co-location patterns based on a fixed interest measure. Venkatesan et al. [11] used spatial statistics and data mining approaches to identify co-location patterns from spatial data sets. Huang et al. [5] developed join-based and Yoo and Shekhar [17], [18] proposed partial-join and join-less co-location algorithms using a fixed interest measure (i.e., spatial prevalence measure). The proposed system in [10] formalized the co-location

problem and showed the similarities and differences between the co-location rules problem and the classic association rules. Manikandan and Srinivasan [8] proposed a novel algorithm for co-location pattern mining which materializes spatial neighborhood relationships with no loss of co-location instances and reduces the computational cost with the aid of Prim's Algorithm.

In [7], to reduce the computation time of the database scanning, the authors used an R-tree index to mine the spatial co-location patterns. In [2] a novel and computationally efficient zonal co-location pattern mining algorithm was developed. This approach used an indexing structure (clQuad-tree) to store co-locations and their instances and handle dynamic parameters.

In summary, the above approaches have two main shortcomings. First, in some cases we have to determine co-location patterns of a desired parameter (or variable) with respect to other parameters e.g. co-location patterns of air pollution with respect to environmental parameters. Second, determination of an optimum unique neighborhood radius in co-location mining process for each application area isn't an easy task. Some studies exist that pursued partially the same goals, but have specific differences too. Xiong et al. [15] proposed a buffer-based model for mining co-location patterns over extended spatial objects. But this approach has some differences: first, they used a repetitive refinement and combinatorial search through costly overlay analysis to detect higher-level co-location patterns, while we first use a tessellation to define core element neighborhoods and index feature instances; thus all levels of co-location patterns are checked only once. Second, they use the Coverage Ratio, which is computationally expensive to determine suitable patterns, whereas we use the Participation Ratio and Index, which are computationally efficient.

Wan et al. [13] and Wan & Zhou [12] tried to neglect using a neighborhood radius in co-location mining, instead they presented a k-nearest feature technique to extract patterns. Although these methods don't need a neighborhood radius, but there is another problem to find an optimal k value. Also, they didn't consider a co-location pattern mining process based on a core element feature.

3. Proposed method

3.1 Basic Concepts

Definition 1: given a spatial framework, a **region** is a subset of the spatial framework.

Because of the spatial heterogeneity law of Geography "results of analysis vary from one place to another" [4], spatial objects have to be mined with a view

on local relationships. Therefore, by Definition 1, we characterize local spaces in order to mine co-location patterns. Supposing that D is a spatial framework, then $SD_i \div SD_i \subset D, (1 \leq i \leq n)$ will be a spatial region with spatial relationships being heterogeneous within it and $D = \{SD_1, SD_2, \dots, SD_n\}$ so that these spatial regions do not overlap.

Definition 2: Given a co-location pattern, a **Pattern Core Element (PCE)** has a feature type (point, line, or polygon) and has its instances in a spatial framework that serve as a basis to define and mine co-location patterns.

Consider an application such as car accident pattern mining. If we want to find patterns between car accidents and other related parameters such as distance to bars, population density, and time, then a car accident is a PCE and patterns will be found regarding it.

Definition 3: for a given PCE set, **Fuzzy Neighborhood** is defined using a couple of radiuses R_1 (lower bound) and R_2 (upper bound) so that form a Membership Function as Equation 1.

$$MF = \begin{cases} 1 & \text{(Compeletly in Neighborhood)} & \text{if } 0 < X \leq R_1 \\ 1 - \frac{X - R_1}{R_2 - R_1} & \text{(Partialy in Neighborhood)} & \text{if } R_1 < X \leq R_2 \\ 0 & \text{(Out of Neighborhood)} & \text{if } R_2 < X \end{cases} \quad (1)$$

Where,

MF: Membership Function that shows neighborhood relation value

R_1, R_2 : Lower and upper bound of neighborhood

X : Radial distance of a feature to PCE

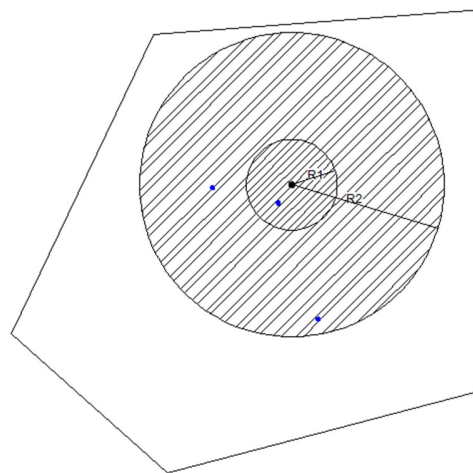


Figure 1. A region with PCE neighborhood illustration

Fig.1 shows a region, a PCE and neighborhood of it defined using R_1 and R_2 . As depicted in Fig. 1, a spatial feature (blue point) can have 3 positions in neighborhood of a PCE (black point). Then using fuzzy concepts we can define neighborhood relation of a spatial feature with PCE

using a membership function defined in Equation (1) and showed in Fig. 2.

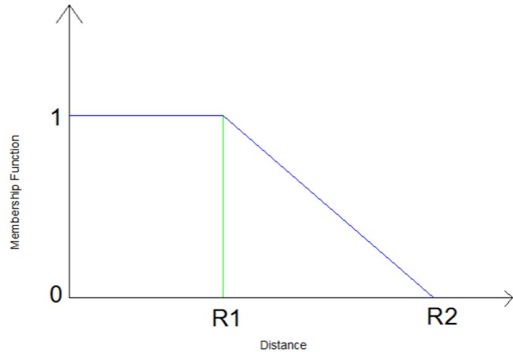


Figure 2. Membership function of PCE neighborhood

Definition 4: Given a region, a PCE, a set of spatial feature types, their instances, and a couple of neighborhood radiuses R_1 and R_2 , a **Regional co-location pattern** is a subset of feature types whose instances are in the neighborhood radiuses and have a fuzzy membership function value greater than 0.

Based on our problem statement that extracts patterns with PCEs, and also to consider spatial relationships of objects implicitly (in contrast to existing methods such as Yoo & Shekhar, [18], we define a spatial co-location pattern by Definition 4 and Equation 2.

$$RCP = \{f_i | f_i \in F \wedge MF(f_i) > 0\} \quad (2)$$

Where,

- f_i : A spatial feature and $1 \leq i \leq m$
- $MF(f_i)$: Fuzzy membership function value of f_i in PCE neighborhood
- F : Set of spatial features,
- RCP : Regional Co-location Pattern that is, a subset of problem spatial features which have close spatial relationships with a PCE.

Definition 5: Given a spatial framework and a set of regions (in our case, Voronoi regions), a PCE is the **centroid** of each region.

Based on Definition 1 and Definition 4, we will mine the regional co-location patterns in the subsets of the spatial framework, which we named Regions. Owing to the challenges of spatial space tessellation mentioned in [2], the proposed partitioning and indexing method for this research is the Voronoi diagram. Given a set of point sites and a distance measure in the plane, the Voronoi diagram partitions the plane into regions, one for each point site, containing all points of the plane that are closer to this site than to any other [6]. Based on these properties of Voronoi

diagrams, we formed Voronoi regions for each PCE to satisfy Definition 1.

3.2 Problem Development

The fundamental idea of regional co-location pattern mining is to reduce search space, increase efficiency, achieve results based on spatial concepts, and index space while preserving relevant co-location patterns. As [2] mentioned, known spatial indexing structures are not so applicable for co-location pattern mining. Here, we propose a spatial index structure to efficiently mine regional co-location patterns. A Voronoi tessellation is used as a Voronoi index that organizes the access to spatial data. By obviating the need to examine objects which are outside the area of interest, the Voronoi index can enhance performance.

In contrast to the existing models for co-location pattern mining, in this model a new definition of neighborhood for PCEs has been developed that this fuzzy definition of neighborhood eliminates necessity of finding a unique optimal neighborhood radius. Also, by applying interval distances to co-location mining process it can provide more accurate results.

Algorithm 1 gives the pseudo code of the proposed method. In this algorithm (cf. pseudo-code in Algorithm 1), line 1 initializes the parameters; line 2 creates the Voronoi regions based on PCEs and spatially indexes the features to the Voronoi cells; lines 3 through 8 represent the core of the algorithm and are explained in detail below; and line 9 returns the results.

Algorithm 1: Regional Co-location Pattern Mining

Inputs:

- F**: a set of distinct spatial feature types
- FI**: a set of feature type instances
- R_1** : spatial neighborhood radius (Lower bound)
- R_2** : spatial neighborhood radius (Upper bound)
- PCE**: a set of pattern core elements
- θ_s** : a spatial prevalence threshold

Output:

- Spatial Co-location patterns whose spatial prevalence indices are greater than θ_s .

Variables:

- k**: co-location size
- C_k** : set of candidate size k co-locations
- I_k** : set of instances of size k co-locations

SCP_k: set of spatially prevalent size k co-locations

Algorithm

```

1: Initialization, k=1, SCPk = Ck= F
2: gen_Voronoi (PCE, F, FI)
3: while (not empty Ck)
4:   Ck+1= gen_co-location_candidates (PCE, SCPk)
5:   Ik+1 = gen_co-location_instances (Ck+1, Ik, R2)
6:   MFk+1 = calc_membership_function (Ik+1, R1, R2)
7:   SCPk+1 = mine_spatial_prev_co-location (Ck+1, MFk+1, θs)
8: k=k+1
9: end while
10: return {SCP2 , ..., SCPk+1 }

```

This algorithm has several functions that can be explained as follows. In line 4, the function `gen_co-location_candidates()` generates size-k+1 candidate co-location patterns C_{k+1} based on all size-k prevalent patterns using an apriori-based method [1] and PCEs as pattern core elements. In line 5, the function `gen_co-location_instances()` works similarly to [5] by joining neighbor instances of size-k spatial co-location patterns, generating the instances of candidate C_{k+1} . In line 6, the function `calc_membership_function()` calculates neighborhood value for each co-location instances based on Definition 3. In line 7, the function `mine_spatial_prev_co-location()` evaluates the candidates to find those patterns whose spatial prevalence criteria are greater than a threshold (θ_s). The spatial prevalence criterion of patterns in this research is the participation index such as in [5], but as mentioned, since we want to handle a fuzzy neighborhood for PCEs then it is necessary to extend the existing criteria. Therefore, we developed a new participation ratio according to the following Equation 3.





$$\Pr(C, f_i) = \frac{\sum MF(f_i)}{N(f_i)} \quad (3)$$

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.

4. Results and discussion

We evaluated the proposed method with synthetic data generated using methodologies used to evaluate algorithms for mining association rules [1]. In this experiment, we used data to test the impact of regional space partitioning and fuzzy definition of neighborhood on

co-location pattern mining. The data includes four feature types A, B, C, and D as shown in Table 1.

Table 1. Feature types in dataset				
Feature Type				
Feature Label	A (PCE)	B	C	D

The data distribution in spatial framework is presented in Fig 3. We mined co-location patterns of these data by three different methods based on PCEs. First, a naïve approach used for mining patterns and as shown in Fig 4, neighborhood regions were created with neighborhood radius $R=1.5$ km. The co-location mining results are presented in Table 2. This process led to the co-location of {A, C} with a threshold level of 0.55.

Second, as shown in Fig. 5, after Voronoi tessellation, a fuzzy neighborhood based on Definition 3 created with lower and upper bounds $R_1=1$ km and $R_2=2$ km respectively. Then co-location mining process was done regarding Definition 4 and the results are presented in Table 3. This process also presented a co-location of {A, C} with a threshold level of 0.55.

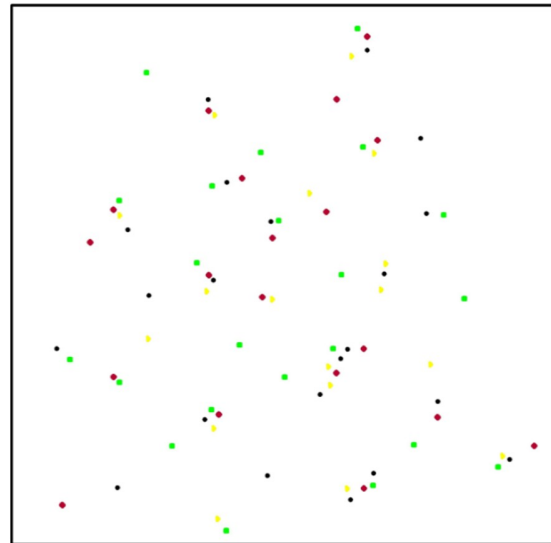


Figure 3. data distribution in spatial framework

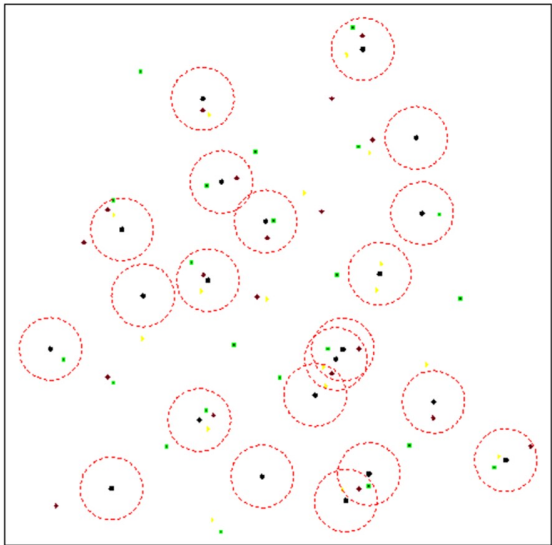


Figure 4. Core element neighborhood regions with $R=1.5$ km

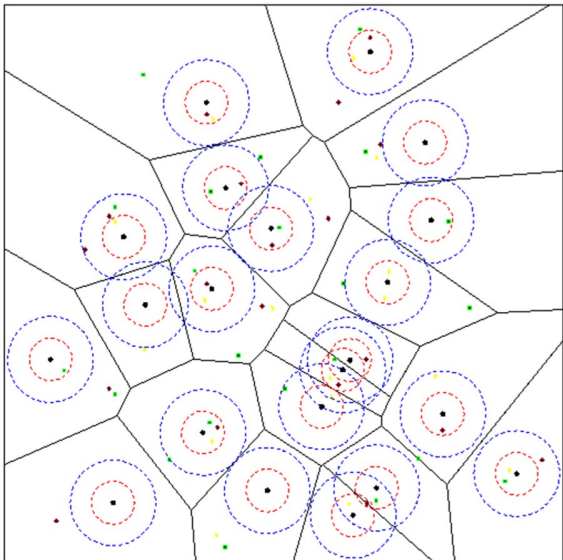


Figure 5. Local tessellation and pattern core element's neighborhood regions with $R1=1$ km & $R2=2$ km













Table 2. Results of co-location mining with a naïve approach						
Co-location						
Pr(Participation ratio)	0.57	0.50	0.48	0.59	0.67	0.63
PI (Participation Index)	0.50		0.48		0.63	
Prevalent(0.55)	No		No		Yes	

Table 3. Results of co-location mining of proposed method						
Co-location						
Pr(Participation ratio)	0.57	0.47	0.52	0.65	0.61	0.62
PI (Participation Index)	0.47		0.52		0.61	
Prevalent(0.55)	No		No		Yes	

Evaluation of results show that different implementations of co-location mining cause to almost similar patterns but there are 2 key differences in our proposed method against the other method: First, as shown in Fig. 5, we consider a fuzzy neighborhood using a lower and upper bound that makes this method more applicable for different application areas even you don't have a deep knowledge about that field. Second, as shown in Fig. 6, the required time for mining patterns in the 2 evaluated method is quite different. When you use a local tessellation such as Voronoi Diagram, then the required time for mining process will reduce considerably.

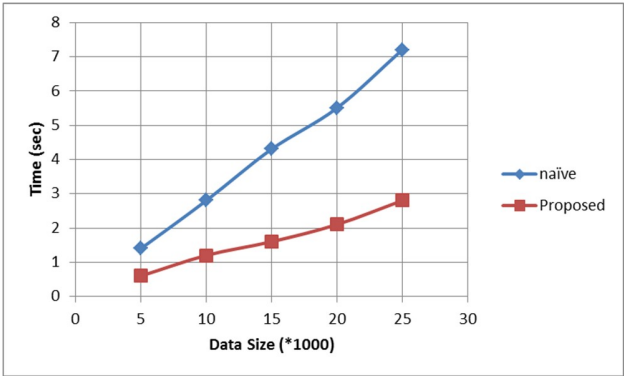


Figure 6. Execution performance for evaluated methods

5. Conclusion and future works

Based on the existing techniques for spatial co-location mining, we developed a new method for regional co-location pattern mining. Our proposed method has several important extensions. First, the pattern search is localized by indexing the data to a Voronoi tessellation prior to the co-location mining process, thus reducing the computational cost. Second, another property of our proposed method is that it considers a fuzzy neighborhood instead of defining a robust neighborhood. In later case you need a deep knowledge of application domain to determine a unique optimum neighborhood distance, but in our method you define neighborhood with a lower and upper bound that is more reliable and applicable. To test our proposed method, we implemented it with the C# programming language and applied it by a synthetic dataset. The results of our experiments suggest that our method have a better performance than a naïve approach and can facilitate applying co-location mining process for different application domains.

In future studies, we would like to apply and test the proposed method with different real datasets. We also intend to extend our model for all feature types (point, line and polygon). Furthermore, future research may want to

consider time as an independent dimension in co-location mining.

References

- [1] Agarwal, R., & Srikant, R., "Fast algorithms for Mining Association Rules", In Proceeding of 20th International Conference on Very Large Data Bases (VLDB), 1994, pp. 487-499.
- [2] Celik, M., Kang, J. M., & Shekhar, S., "Zonal Co-location Pattern Discovery with Dynamic Parameters", In Proceeding of Seventh IEEE International Conference on Data Mining, 2007, pp. 433-438, Omaha, NE, DOI: 10.1109/ICDM.2007.102.
- [3] Ding, W., Jiamthapthaksin, R., Parmar, R., Jiang, D., Stepinski, T. F., & Eick, C. F., "Towards Region Discovery in Spatial Datasets", In Proceeding of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2008, pp. 88-99, Osaka, Japan, DOI: 10.1007/978-3-540-68125-0_10.
- [4] Goodchild, M. F., "The fundamental laws of GIScience", Invited talk at University Consortium for Geographic Information Science, University of California, Santa Barbara, 2003.
- [5] Huang, Y., Shekhar, S., & Xiong, H., "Discovering co-location patterns from spatial datasets: A general approach" IEEE Trans. on Knowl. and Data Eng., 16(12), 2004, pp. 1472-1485.
- [6] Icking, C., Klein, R., Kollner, P., & Ma, L., "Java Applets for the Dynamic Visualization of Voronoi Diagrams", Comput. Sci. in Perspect., Lect. Notes in Comput. Sci., 2598, 2003, pp. 191-205, DOI: 10.1007/3-540-36477-3_14.
- [7] Manikandan, G., & Srinivasan, S., "Mining of Spatial Co-location Pattern Implementation by FP Growth", Indian J. of Comput. Sci. and Eng. (IJCSE), 3(2), 2012, pp. 344-348, ISSN: 0976-5166.
- [8] Manikandan, G., & Srinivasan, S., "Mining Spatially Co-Located Objects from Vehicle Moving Data", Eur. J. of Sci. Res., 68(3), 2012, pp. 352-366, ISSN: 1450-216X.
- [9] Miller, H. J., & Han, J., Geographic Data Mining and Knowledge Discovery, 2nd edition, London: CRC Press, published, 486pp, 2009.
- [10] Priya, G, Jaisankar, N., & Venkatesan, M., "Mining Co-location Patterns from Spatial Data using Rulebased Approach" Int. J. of Glob. Res. in Comput. Sci., 2(7), 2011, pp. 58-61.
- [11] Venkatesan, M. S., Arunkumar, Th., & Prabhavathy, P., "Discovering Co-location Patterns from Spatial Domain using a Delaunay Approach", Procedia Engineering, International conference on modeling optimization and computing, 38, 2012, pp. 2832-2845.
- [12] Wan, Y., & Zhou, J., "KNFCOM-T: a k-nearest features-based co-location pattern mining algorithm for large spatial data sets by using T-trees", Int. J. of Bus. Intell. and Data Min., 3(4), 2008, pp. 375-389, DOI: 10.1501/IJBIDM.2008.022735.
- [13] Wan, Y., Zhou, J., & Bian, F., "CODEM: a novel spatial co-location and de-location patterns mining algorithm" In Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on (Vol. 2, pp. 576-580). IEEE.
- [14] Xiao, X., Xie, X., Luo, Q., & Ma, W., "Density based co-location pattern discovery", In Proceeding of ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM-GIS), 2008, Irvine, CA, USA, DOI: 10.1145/1463434.1463471.
- [15] Xiong, H., Shekhar, S., Huang, Y., Kumar, V., Ma, X., & Yoo, J. S., "A framework for discovering co-location patterns in data sets with extended spatial objects" In Proceeding of the 2004 SIAM international conference on data mining (SDM'04), Lake Buena Vista, FL, 2004, pp. 78-89.
- [16] Yoo, J. S., & Bow, M., "Mining Top-k Closed Co-location Patterns", In Proceeding of IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM), Fuzhou, 2011, pp. 100-105, DOI: 10.1109/ICSDM.2011.5969013.
- [17] Yoo, J. S., & Shekhar, S., "A Partial Join Approach for Mining Co-location Patterns", In Proceeding of ACM SIGSPATIAL International conference on Advances in Geographic Information Systems (ACM-GIS), 2005, pp. 241-249, doi:10.1145/1032222.1032258.
- [18] Yoo, J. S., & Shekhar, S., "A Join-less Approach for mining Spatial Co-location Patterns", IEEE Trans. on Knowl. and Data Eng., 18(10), 2006, pp. 1323-1337, DOI: 10.1109/ICDM.2005.8.

Mohammad Akbari is PhD. Candidate of Dept. of Surveying & Geomatics Engineering, Collage of Engineering, University of Tehran since 2009. His field for PhD and Master is GIS. He graduated for MSc. in 2009 from university of Tehran. His interest fields are data mining and air pollution modeling.

Farhad Samadzadegan is Professor of Dept. of Surveying & Geomatics Engineering, Collage of Engineering, University of Tehran. He is the head of a research and development group.