Mining significant crisp-fuzzy spatial association rules

Wenzhong Shi[a], Anshu Zhang[a]* and Geoffrey I. Webb[b]

[a] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, P.R. China; [b] Faculty of Information Technology, Monash University, Victoria 3800, Australia

Abstract: Spatial association rule mining (SARM) is an important data mining task for understanding implicit and sophisticated interactions in spatial data. The usefulness of SARM results, represented as sets of rules, depends on their reliability: the abundance of rules, control over the risk of spurious rules, and accuracy of rule interestingness measure (RIM) values. This study presents crisp-fuzzy SARM, a novel SARM method that can enhance the reliability of resultant rules. The method firstly prunes dubious rules using statistically sound tests and crisp supports for the patterns involved, and then evaluates RIMs of accepted rules using fuzzy supports. For the RIM evaluation stage, the study also proposes a Gaussian-curve-based fuzzy data discretization model for SARM with improved design for spatial semantics. The proposed techniques were evaluated by both synthetic and real-world data. The synthetic data was generated with predesigned rules and RIM values, thus the reliability of SARM results could be confidently and quantitatively evaluated. The proposed techniques showed high efficacy in enhancing the reliability of SARM results in all three aspects. The abundance of resultant rules was improved by 50% or more compared with using conventional fuzzy SARM. Minimal risk of spurious rules was guaranteed by statistically sound tests. The probability that the entire result contained any spurious rules was below 1%. The RIM values also avoided large positive errors committed by crisp SARM, which typically exceeded 50% for representative RIMs. The real-world case study on New York City points of interest reconfirms the improved reliability of crisp-fuzzy SARM results, and demonstrates that such improvement is critical for practical spatial data analytics and decision support.

Keywords: spatial association rules, fuzzy sets and logic, quality issues, statistical evaluation, spatial data mining

*: corresponding author. Email: anshu.zhang@connect.polyu.hk

To cite this article: Wenzhong Shi, Anshu Zhang & Geoffrey I. Webb (2018): Mining significant crisp-fuzzy spatial association rules, International Journal of Geographical Information Science
To link to this article: https://doi.org/10.1080/13658816.2018.1434525

1. Introduction

Spatial association rule mining (SARM) is an important topic in geographical information science (Ladner *et al*. 2003, Mennis and Liu 2005, Bogorny *et al*. 2008, Laube *et al*. 2008, Verhein and Chawla 2008, Baralis *et al*. 2012, Versichele *et al*. 2014, Faridi *et al*. 2017, Xu *et al*. 2016). SARM finds implicit patterns called *spatial association rules* that meet constraints on certain *rule interestingness measures* (RIMs) from spatial databases. With great power in revealing and prioritizing an enormous number of sophisticated data interactions, SARM has been of rising popularity in research and practical applications.

The usefulness of SARM results depends on its reliability, including *abundance* of authentic rules, the control over the *risk* of spurious rules, and *accuracy* of RIM values. Fuzzy SARM is a key to enhancing the reliability of resultant rules. Fuzzy data discretization divides numerical data attributes into linguistic concepts with fuzzy boundaries, and assigns membership functions to the concepts. This can relieve the inaccurate semantic representations and reduce unreliable rules result from hard divisions of gradual or vague concepts (Hüllermeier 2009, Farzanyar and Kangavari 2012). Studies have suggested that spatial data tends to be fuzzy in nature, and proposed fuzzy SARM for spatial relations such as proximity (Ladner *et al*. 2003, Laube *et al*. 2008). However, the ability of fuzzy SARM to generate more reliable results than ordinary (crisp) SARM, particularly in terms of RIM accuracy, has seldom been examined by quantitative studies. Also, existing data discretization models for fuzzy SARM generally lack systematic justifications for their mathematical rationales.

Furthermore, it is important to conduct statistical tests in SARM for controlling spurious rules and preventing them from accounting for a high percentage of resultant rules (Webb 2007, Zhang *et al*. 2016). Statistically sound evaluation (Webb 2007) is a particularly effective testing technique and can control the familywise error rate (FWER), the chance that entire result includes any spurious rules, upon a low user specified level, for example 5%. This technique has only been systematically applied to ordinary association rule mining, but not yet on fuzzy rule mining or fuzzy SARM.

This study presents a novel *crisp-fuzzy SARM* method that can effectively enhance the reliability of results in mining significant fuzzy spatial association rules. Significant rules refer to rules accepted by statistical tests. This method firstly performs statistically sound tests on crisp rules, and then evaluates RIMs of significant rules using fuzzy memberships. The method combines the abundance of true rules in ordinary SARM, high RIM accuracy in fuzzy SARM and low risk of spurious rules guaranteed by statistically sound tests. A new Gaussian-curve-based fuzzy data discretization model is also proposed for the RIM evaluation stage of crisp-fuzzy SARM. The model is designed for spatial concepts, especially nearness relations between spatial objects, and has systematic mathematical justifications.

The new method can markedly increase the number of resultant rules compared with conventional fuzzy SARM, avoid large positive errors in RIM values committed by ordinary SARM, and maintain minimal FWER because of statistically sound tests. A synthetic data experiment with predesigned rules and RIM values demonstrates the efficacy of the new method, and a real-world data experiment reconfirms its advantages and elaborates its practical value.

This article is organized as follows. Section 2 introduces relevant previous research. Section 3 presents crisp-fuzzy SARM and the new fuzzy data discretization model. Sections 4 and 5 present experimental evaluation of crisp-fuzzy SARM with synthetic and real-world data, respectively, evaluate the results and discuss their practical implications. Section 6 draws conclusions and suggests directions for future research.

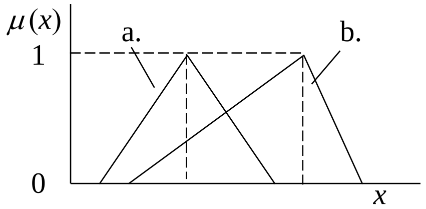
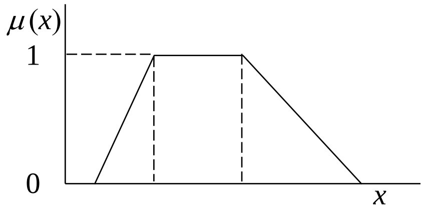
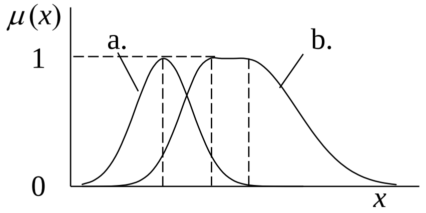
2. Backgrounds for proposed techniques

2.1. Fuzzy spatial association rules

Let *D* be a dataset, and each record $R \in D$ be a set of *items* in the form ‘attribute = value’. An *association rule* is a pattern $X \rightarrow Y$, where the *antecedent* $X = \{x_1 \ldots x_p\}$ and *consequent* $Y = \{y_1 \ldots y_q\}$ are itemsets consisting of items in *D*. $X \cup Y$ contains at most one item for each attribute. $X \rightarrow Y$ is a spatial association rule if it involves one or more spatial attribute(s). SARM includes two primary tasks: first, to compute necessary spatial attributes from geometries of spatial objects, either before or during the rule exploration. Second, to explore rules from data including the computed spatial attributes, using algorithms similar to those for general association rule mining, such as Apriori (Agrawal and Srikant 1994) type algorithms.

Before *D* is explored for rules, each numerical attribute *x* in *D* needs to be transformed into linguistic concepts via data discretization. A fuzzy data discretization model defines a *membership function,* μ_l : domain of *x* → [0,1], for each concept *l*. The *membership degree* of *x* in μ_l, $\mu_l(x)$, represents the degree to which *x* belongs to *l*. The *core* and *support* of μ_l is respectively $core(\mu_l) = \{x \in U | \mu_l(x) = 1\}$ and $supp(\mu_l) = \{x \in U | \mu_l(x) > 0\}$ (Bosc *et al*. 2007). Table 1 lists common types of previously proposed membership functions.

Table 1. Common forms of membership functions for fuzzy data discretization.

Form of $\mu(x)$	Graph	Reference(s)
Triangular		a. Bilaterally symmetry: Harrera and Martinez (2000), Chen *et al*. (2008), Carmona *et al*. (2010) b. Bilaterally asymmetry: Alhajj and Kaya (2008)
Trapezoidal		Ladner *et al*. (2003)
Gaussian-curve		a. Core for a single *x* value: Bordogna and Pasi (1993) b. Core for an *x* value range: Burda *et al*. (2014)

Conjunctions of multiple membership degrees are evaluated by t-norm, an associative, commutative and monotone function $\otimes$: $[0,\ 1]\times[0,\ 1]\rightarrow[0,\ 1]$, $\alpha\otimes 1=\alpha$ and $\alpha\otimes 0=0$ for each $\alpha\in[0,\ 1]$. SARM mostly adopts minimum t-norm: $\alpha\otimes_{\text{min}}\beta=\min(\alpha,\ \beta)$ and product t-norm: $\alpha\otimes_{\text{prod}}\beta=\alpha\beta$ (Laube *et al*. 2008). The *fuzzy support* of an itemset $V=\{'x_1=v_1'\ldots'x_m=v_m'\}$ is

$$supp(V)=\sum_{R\in D}\mu_{v_1}(r_1)\otimes\ldots\otimes\mu_{v_m}(r_m)\,. \tag{1}$$

Fuzzy SARM aims to find all spatial association rules that meet designated criteria, mostly about specified RIMs computed using fuzzy supports. Over 60 RIMs have been proposed for general association rules, most of which are applicable to SARM. Systematic reviews (Tew *et* *al*. 2013) revealed that many RIMs get very similar results when ranking the interestingness of rules. Some of the most common RIMs are:

- *support* (Agrawal *et al*. 1993): $supp\left(X \to Y\right) = supp\left(X \cup Y\right) = supp\left(\left\{x_1 \ldots x_p y_1 \ldots y_q\right\}\right)$;
- *confidence* (Agrawal *et al*. 1993): $conf\left(X \to Y\right) = supp\left(X \to Y\right) / supp\left(X\right)$;
- *improvement* (Bayardo *et al*. 2000): $imp(X \to Y) = conf(X \to Y) - \max_{Z \subset X}(conf(Z \to Y))$;
- *leverage* (Piatetsky-Shapiro 1991):
  $lev\left(X \to Y\right) = supp\left(X \to Y\right) - supp\left(X\right) supp\left(Y\right) / \left|D\right|$.

RIMs exclusively for SARM, such as spatial support and spatial confidence (Laube *et al*. 2008), have also been proposed. Replacing all membership degrees in fuzzy SARM with binary memberships 0/1 reduces it to ordinary SARM.

Other criteria for pruning uninteresting rules can often be transformed into RIM criteria. For example, the non-redundant rule criterion (Zaki 2000) is entailed by 'improvement > 0'. The actionable rule criterion (Liu *et al*. 2001) accepts rules with positive improvements and higher confidences than $\varnothing \to y$ even after removing records that conform to any specializations of these rules.

2.2. Statistical tests on association rules

SARM data, sampled or population, are finite representations of associations between studied objects which can potentially repeat for infinite times in the real world. A rule might fulfil specified RIM constraints in data by chance rather than due to real-world associations of studied objects. If so, this rule will be spurious.

Statistical hypothesis tests have been developed to avoid spurious rules (Megiddo and Srikant 1998, Liu *et al*. 1999, Bay and Pazzani 2001, Zhang *et al*. 2004). The resultant *p* value of such a test is the probability that a rule $X \to Y$ has observed RIM value even if $X \to Y$ association is actually nonexistent, or the risk that $X \to Y$ is spurious. Only statistically *significant* rules with *p* values below the significance level *α*, say 0.05, are accepted. This study exemplifies such tests by the chi-square test for *productive* rules, a typical test for pruning redundant rules, as detailed in Appendix 1. The propose method also applies to other tests.

Normally, only a minority of the rules tested turn out to be interesting. In this case, when many rules are tested, the risk of accepting spurious rules may be very high, even each uninteresting rule has only a small probability below *α* to be accepted. Ordinary SARM studies show that most resultant rules can be spurious if tested at a 0.05 significance level (Webb 2007, Zhang *et al*. 2016).

Webb (2007) proposed the *statistically sound evaluation* which sets the significance level $\kappa = \alpha/s$, where *s* is the *search space* size of the rules, that is, the number of all potential rules that the data can constitute. Alternatively, a slightly higher and dynamic *κ* value can be determined by the Holm procedure (Holm 1979) to accept more rules. This technique can reduce the FWER to below 1% with $\alpha = 0.05$. Albeit highly effective, the technique is conservative and also rejects many authentic rules. In ordinary SARM, the technique usually reserves adequate number of rules for practical use. In fuzzy SARM, however, much fewer rules can be accepted due to the use of fuzzy memberships, as will be shown in 4.2.

3. Proposed techniques for mining significant fuzzy spatial association rules

3.1. Gaussian-curve-based fuzzy data discretization

Previous studies have proposed Gaussian-curve-based fuzzy data discretization models for individual concepts (Table 1). This study presents a more complete model for spatial attributes covering multi-concept relations, and justifies the mathematical characteristics of each model component for geographical studies.

The proposed model transforms a numerical spatial attribute *x* that is gradual or vague to ordinal concepts $l_1 \dots l_n$. For instance, *x* is distance and $\left(l_1,\ l_2,\ l_3\right) = \left(\text{near, medium, far}\right)$. The concepts are ordinal in that they cannot fully or partially imply one another, though their corresponding *x* value ranges can overlap. Therefore, $\left(l_1,\ l_2,\ l_3\right) = \left(\text{near, medium, medium to far}\right)$ is invalid. For concepts with well-defined boundaries, such as ‘above/below sea level’ for attribute ‘elevation’, the model is specialized to place crisp boundaries between the concepts.

The model has the following characteristics:

(1) For each membership function $\mu_{l_j}, 1 \le j \le n$, the sections with $0 < \mu_{l_j}\left(x\right) < 1$, named *transitions* after ‘transitions between concepts’, are Gaussian curves and can be symmetric or not. $core\left(\mu_{l_j}\right)$ is non-empty and arbitrary in size.

Gaussian curves have been widely used for characterizing degrees to which spatial attribute values belong to linguistic concepts, especially for proximity measures. Gaussian weighting function is most commonly used in fixed-kernel geographically weighted regression (Wu *et al*. 2014). The weight represents the impact factor due to nearness between spatial objects, and is equivalent to μ_{near} in fuzzy SARM. Robinson (2000) used Gaussian functions to learn fuzzy spatial relations via human-machine interaction. Worboys (2001) demonstrated with empirical study that the degrees people perceive two places as ‘near’ or ‘not near’, translating to μ_{near} and μ_{far} in fuzzy SARM, exhibit S-curve trend against Euclidean distances. The S-curve trend was visually interpreted and thus indistinguishable from a Gaussian-curve trend. The precise curve form is not that essential; the essence is that in contrast to triangular or trapezoidal membership functions, transition curves shall have larger slopes in the middle and smaller towards the endpoints. This reflects the fact that for *x* values in the middle of transitions, people have more uncertainty (modelled by curved slopes) in judging to which concept *x* should belong. Gaussian-curve transitions are also robust to uncertain and usually non-ideal value intervals of $core\left(\mu_{l_j}\right)$, since the curves smoothly connect to core endpoints (Bordogna *et al*. 1991).

Asymmetric transitions are critical for spatial concepts, as geographical data prevalently have heavy-tailed distributions. That is, the data includes a ‘head’ containing a minority of large-sized objects, and a ‘tail’ containing a majority of small-sized objects; ‘size’ may be population, connectivity, or other impact measures (Jiang 2013). Thus, the raw numerical data value intervals for low-impact to high-impact concepts should increase exponentially. For example, populations of small, medium and large towns are more likely 1:5:25 than 1:5:9. Then the left transition of the ‘medium town’ concept towards ‘small town’ shall be much narrower than the right one towards ‘large town’.

(2) For relations between $l_1 \ldots l_n$, each $supp\left(\mu_{l_{j-1}}\right)$ touches $core\left(\mu_{l_j}\right), 1 < j \leq n$, neither overlap nor disjoint.

This characteristic follows a common approach in past studies (Harrera and Martinez 2000, Alhajj and Kaya 2008, Carmona *et al*. 2010). Then $\mu_{l_j}(x)=1 \Leftrightarrow \mu_{i\neq j}(x)=0$, or each *x* value completely belongs to l_j if and only if it belongs to none of the other concepts. The proposed $\mu_{l_1}\ldots\mu_{l_n}$ are as below and illustrated in Figure 1:

$$\mu_{l_1}(x)=\begin{cases}1, & x\le c_{1\_\mathrm{R}}\\ \exp\left[-\left(x-c_{1\_\mathrm{R}}\right)^2\Big/\left(2\sigma_{1\_\mathrm{R}}^2\right)\right], & c_{1\_\mathrm{R}}<x<c_{2\_\mathrm{L}}\\ 0, & x\ge c_{2\_\mathrm{L}}\end{cases}$$

$$\mu_{l_j}(x)=\begin{cases}0, & x\le c_{(j-1)\_\mathrm{R}} \text{ or } x\ge c_{(j+1)\_\mathrm{L}}\\ \exp\left[-\left(x-c_{j\_\mathrm{L}}\right)^2\Big/\left(2\sigma_{j\_\mathrm{L}}^2\right)\right], & c_{(j-1)\_\mathrm{R}}<x<c_{j\_\mathrm{L}}\\ 1, & c_{j\_\mathrm{L}}\le x\le c_{j\_\mathrm{R}}\\ \exp\left[-\left(x-c_{j\_\mathrm{R}}\right)^2\Big/\left(2\sigma_{j\_\mathrm{R}}^2\right)\right], & c_{j\_\mathrm{R}}<x<c_{(j+1)\_\mathrm{L}}\end{cases}\quad,\ 1<j<n, \qquad (2)$$

$$\mu_{l_n}(x)=\begin{cases}0, & x\le c_{(n-1)\_\mathrm{R}}\\ \exp\left[-\left(x-c_{n\_\mathrm{L}}\right)^2\Big/\left(2\sigma_{n\_\mathrm{L}}^2\right)\right], & c_{(n-1)\_\mathrm{R}}<x<c_{n\_\mathrm{L}}\\ 1, & x\ge c_{n\_\mathrm{L}}\end{cases}$$

where $core\left(\mu_{l_j}\right)=\left[c_{j\_\mathrm{L}},\ c_{j\_\mathrm{R}}\right]$, and σ_{j_L} and σ_{j_R} are standard deviations of left and right transitions in μ_{l_j}.

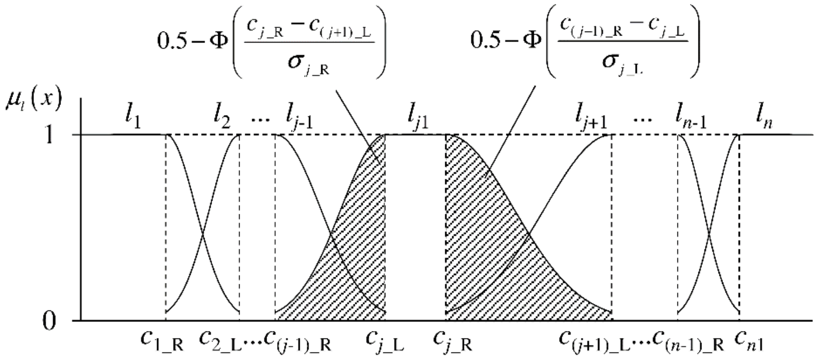

$0.5-\Phi\left(\frac{c_{j\_\mathrm{R}}-c_{(j+1)\_\mathrm{L}}}{\sigma_{j\_\mathrm{R}}}\right)$
$0.5-\Phi\left(\frac{c_{(j-1)\_\mathrm{R}}-c_{j\_\mathrm{L}}}{\sigma_{j\_\mathrm{L}}}\right)$
$\mu_l(x)$
1
0
l_1
l_2
...
$l_{j\text{-}1}$
l_{j1}
l_{j+1}
...
$l_{n\text{-}1}$
l_n
c_{1_R}
$c_{2\_\mathrm{L}}\cdots c_{(j\text{-}1)\_\mathrm{R}}$
c_{j_L}
c_{j_R}
$c_{(j+1)\_\mathrm{L}}\cdots c_{(n\text{-}1)\_\mathrm{R}}$
c_{n1}

Figure 1. Illustration of the proposed fuzzy data discretization model.

(3) Given no prior knowledge, there is an intuitively unbiased suggestion to set σ_{j_L} and σ_{j_R} to make the cumulative $\mu_{l_j}(x)$ is half the size of transition ranges, or takes half of full memberships of l_j in the transitions.

This makes $0.5-\Phi\left(\left(c_{(j-1)\_\mathrm{R}}-c_{j\_\mathrm{L}}\right)/\sigma_{j\_\mathrm{L}}\right)=0.5\times\left(c_{j\_\mathrm{L}}-c_{(j-1)\_\mathrm{R}}\right)$, where Φ is the cumulative standard normal distribution function. Resultantly, $\sigma_{j\_\mathrm{L}}=\left(c_{j\_\mathrm{L}}-c_{(j-1)\_\mathrm{R}}\right)/2.473$ and similarly $\sigma_{j\_\mathrm{R}}=\left(c_{(j+1)\_\mathrm{L}}-c_{j\_\mathrm{R}}\right)/2.473$. Following Bordogna and Pasi (1993), $\mu_{l_j}(x)$ values can be left out and equal to 0 for $x<c_{(j-1)\_\mathrm{R}}$ and $x>c_{(j+1)\_\mathrm{L}}$.

3.2. Crisp-fuzzy SARM for mining authentic and accurate rules

Effective control over spurious rules by statistically sound tests has been experimentally proven using predesigned associations between already categorized numerical attributes (Webb 2007, Zhang *et al*. 2016). This study seeks to use realistic synthetic data which contains associations of varying strength directly depending on raw numerical attribute values in a gradual manner, and some data disturbances (see 4.1). The statistically sound evaluation turns out to maintain very low FWER.

Meanwhile, artificial crisp representations of gradual or vague concepts are expected to distort, mostly exaggerate, RIM values. As exemplified in Table 2, supports of individual items like $supp(A)$ may not be exaggerated. More undesirably, positive associations between data items are overstated via t-norm operations. Thus, $supp(A\rightarrow B)$ can be significantly overestimated, and finally $imp(A\rightarrow B)$ and $lev(A\rightarrow B)$ become exaggerated. The exaggeration holds for both product and minimum t-norms, and worsens when rules contain more items and thus RIM evaluations include more t-norm operations.

Table 2. RIM value exaggerations due to crisp data discretization in a miniature database of four records. Numerical attributes *a* and *b* are discretized into ordinal concepts including l_A and l_B, respectively. Item *A* is $a = l_A$ and *B* is $b = l_B$.

Record#	Fuzzy				Crisp		
	$\mu_{l_A}(a)$	$\mu_{l_B}(b)$	$\mu_{l_A}(a)\otimes_{\text{prod}}$ $\mu_{l_B}(b)$	$\mu_{l_A}(a)\otimes_{\text{min}}$ $\mu_{l_B}(b)$	$\mu_{l_A}(a)$	$\mu_{l_B}(b)$	$\mu_{l_A}(a)\otimes$ $\mu_{l_B}(b)$
1	1	1	1	1	1	1	1
2	0.6	0.8	0.48	0.6	1	1	1
3	0.4	0.2	0.08	0.2	0	0	0
4	0	0	0	0	0	0	0
$supp(A)=\sum\mu_{l_A}(a)$			2	2			2
$supp(B)=\sum\mu_{l_B}(b)$			2	2			2
$supp(A\rightarrow B)=\sum\mu_{l_A}(a)\otimes\mu_{l_B}(b)$			1.56	1.8			2
$imp(A\rightarrow B)$ $=conf(A\rightarrow B)-conf(\varnothing\rightarrow B)$ $=supp(A\rightarrow B)/supp(A)-supp(B)/4$			0.28	0.4			0.5
$lev(A\rightarrow B)$ $=supp(A\rightarrow B)-supp(A)supp(B)/4$			0.56	0.8			1

Compared with support and confidence, RIMs evaluating how different the associations between items in rules are from independence among the items, such as improvement and leverage, appear to be exaggerated more severely. These RIMs are ‘margins’ of itemset supports, thus small overestimations in supports can be much amplified in them. In Table 2, the crisp rule has 28% exaggeration on $supp(A\rightarrow B)$ but 79% on $imp(A\rightarrow B)$ and $lev(A\rightarrow B)$ with respect to fuzzy product t-norm results. For experiments in Sections 4 and 5, improvement and leverage are typically exaggerated by at least 50%. Unfortunately, support and sometimes confidence mainly serve to control the number of rules considered, while other RIMs that are more severely exaggerated are usually of higher interest for users.

While fuzzy rules have more accurate RIM values, they are also more moderate and less significant in the statistically sound evaluation. This markedly reduces the rules accepted compared with ordinary SARM, as experimentally shown in Sections 4 and 5. To combine the abundance of resultant rules using crisp supports and higher accuracy of fuzzy RIM values, this study proposes *crisp-fuzzy SARM*:

- Firstly, perform statistically sound tests on rules, using crisp supports of itemsets involved;
- Then evaluate RIM values of significant rules accepted by the tests using fuzzy supports.

As will be elaborated in Section 4, statistically sound tests are still indispensable to controlling the FWER for fuzzy SARM. In this situation, crisp-fuzzy SARM can maximize the number of rules without sacrificing strong control over the risk of spurious rules. The statistical test stage shall use a crisp discretization model matching the fuzzy one for RIM evaluation:

$$\mu_{l_j}\left(x\right)=\begin{cases}1, & \left(c_{(j-1)\_\mathrm{R}}+c_{j\_\mathrm{L}}\right)/2\le x\le\left(c_{j\_\mathrm{R}}+c_{(j+1)\_\mathrm{L}}\right)/2\\0, & \text{otherwise}\end{cases}\quad. \tag{3}$$

Each *x* value in equation (3) has the same concept of maximum membership degrees as in the fuzzy model.

4. Experiment with synthetic data

4.1. Methods

The experiment data included three spatial point sets: objective, resource1 and resource2. Each point had two-dimensional coordinates (*x*, *y*). Each objective linked to a record containing nine attributes: near1, near2, other1, other2, noise1 to noise4, and outcome.

Attribute near1 and near2 were fuzzy and respectively represented nearness, or accessibility of objective to resource1 and resource2. They each had two values: Y (yes-near) and N (no-far). Following the proposed data discretization model in equation (2), $\mu_{\mathrm{Y}}\left(\mathrm{near1}\right)$ and $\mu_{\mathrm{N}}\left(\mathrm{near1}\right)$ were computed from the Euclidean distance between each objective and the nearest resource1, by taking $c_{1\_\mathrm{R}} = d_{0.01}$ and $c_{2\_\mathrm{L}} = d_{0.99}$, where $d_{0.01}$ and $d_{0.99}$ were the first and 99th percentiles of the distances for all objectives. Partial membership degrees were assigned to nearly full range of the original distance values to better differentiate accessibilities for different objectives. Cutting data at $d_{0.01}$ and $d_{0.99}$ was for lessening sensitivities of c_{1_R} and c_{2_L} to extreme data, following real-world SARM practice. All above equally applied for near2.

Attributes other1, other2 and noise1 to noise4 were categorical; other2 had 10 possible values (0~9) and the others had four (0~3). Values of these attributes were generated randomly, independently and equiprobably for every possible value.

The attribute 'outcome' was also fuzzy, with values Bad and Good. It was the only attribute with dependences on other attributes. The dependences were listed in Table 3: higher $\mu_{\mathrm{Y}}\left(\mathrm{near1}\right)$ improved the goodness of outcome unconditionally, while higher other2 values did so only when other1 = 0 or 2, and higher $\mu_{\mathrm{Y}}\left(\mathrm{near2}\right)$ did so only when other1 = 0 or 1. Such unconditional and conditional associations are both common in practical SARM. Attributes noise1 to noise4 had no association with outcome and were for examining the tolerance of the proposed techniques to irrelevant data. It was set that $\mu_{\text{Bad}}(\text{outcome}) = 1 - \mu_{\text{Good}}(\text{outcome})$.

Table 3. Dependence of $\mu_{\text{Good}}(outcome)$ on other attributes.

other1	*other2*	Expectation of $\mu_{\text{Good}}(outcome)$ (Standard deviation = 0.15)
0	0, 1, 2, 3, 4	$\left(\mu_{\text{Y}}(\text{near1}) + \mu_{\text{Y}}(\text{near2})\right) / fac_1$[a] $- 0.35, -0.3, -0.25, -0.2, -0.15$
	5, 6, 7, 8, 9	$\left(\mu_{\text{Y}}(\text{near1}) + \mu_{\text{Y}}(\text{near2})\right) / fac_1 + 0.15, +0.2, +0.25, +0.3, +0.35$
1	Any	$\left(\mu_{\text{Y}}(\text{near1}) + \mu_{\text{Y}}(\text{near2})\right) / fac_1$
2	0, 1, 2, 3, 4	$\mu_{\text{Y}}(\text{near1}) / fac_2 - 0.35, -0.3, -0.25, -0.2, -0.15$
	5, 6, 7, 8, 9	$\mu_{\text{Y}}(\text{near1}) / fac_2 + 0.15, +0.2, +0.25, +0.3, +0.35$
3	Any	$\mu_{\text{Y}}(\text{near1}) / fac_2$

[a] *fac1*: mean value of all (near1 + near2); *fac2*: mean value of all near1; *fac1* and *fac2* are to adjust expectations of $\left(\mu_{\text{Y}}(\text{near1}) + \mu_{\text{Y}}(\text{near2})\right) / fac_1$ and $\mu_{\text{Y}}(\text{near1}) / fac_2$ to 0.5 and thus cancel out data variations due to capping $\mu_{\text{Good}}(outcome)$ values beyond [0, 1] to 0 and 1.

The predesigned data associations generated 118 productive rules with outcome values as the consequents:

- near1 = Y ∧ zero or one of other1 = 2 or 3 → outcome = Good (3 rules);
- near1 = N ∧ zero or one of other1 = 2 or 3 → outcome = Bad (3 rules);
- Zero or one of near1 = Y ∧ near2 = Y ∧ zero or one of other1 = 0 or 1 → outcome = Good (6 rules);
- Zero or one of near1 = N ∧ near2 = N ∧ zero or one of other1 = 0 or 1 → outcome = Bad (6 rules);

- Zero or one of near1 = Y $\wedge$ other1 = 2 $\wedge$ other2 = 5~9 → outcome = Good (10 rules);
- Zero or one of near1 = N $\wedge$ other1 = 2 $\wedge$ other2 = 0~4 → outcome = Bad (10 rules);
- Zero or more of near1 = Y, near2 = Y and other1 = 0 $\wedge$ other2 = 5~9 → outcome = Good (40 rules);
- Zero to three of near1 = N, near2 = N and other1 = 0 $\wedge$ other2 = 0~4 → outcome = Good (40 rules).

To evaluate the robustness of the proposed techniques, experiment groups (each called a *treatment*) were constructed using data in three sizes and various other settings, as summarized in Table 4. A sample dataset with the map and detailed description is provided as supplementary material. The *extraneous factors* simulated numerous affecting factors to real-world imperfect distance data, including but not limited to measurement errors. For instance, the distance between two places for all citizens can be longer than the recorded shortest path, if barrier-free paths between the places are long detours. Previous work has suggested that Gaussian-curve relations between concept memberships and raw numerical data are more usual (see 3.1). Still, some treatments also used ‘true’ and ‘perceived’ memberships with linear transitions, for comprehensively comparing the proposed data discretization model with triangular/trapezoidal ones.

Table 4. Various experiment settings in synthetic data experiment.

Item	Variations	Remarks
1. Data size (No. of objectives)	5000, 20,000, 80,000	
2. Spatial patterns of objectives and resources	Clustered, random, dispersed	Point sets must pass nearest neighbor index test (Mitchell 2005) for designated patterns with threshold p = 0.05
3. Extraneous factors	σ = 0, 10%, 20%	Added to data through multiplying raw near1/near2 by a random variable following normal distribution $N\left(1,\sigma^2\right)$
4. 'True' membership functions for near1/ near2 w.r.t. their raw values	Linear, Gaussian-curve	Linear: membership functions with linear transitions and endpoints coincided with the Gaussian one, e.g. $\mu_{\mathrm{Y}}\left(\text{near1}\right)=$ $\left(\text{near1}-d_{.99}\right)/\left(d_{.01}-d_{.99}\right)$ for $d_{.01}<$ near1$<d_{.99}$
5. Memberships used in statistical tests	Crisp, linear, Gaussian-curve	Matching crisp functions were defined following equation (3); linear and Gaussian-curve treatments used original membership values for outcome computed from Table 3
6. 'Perceived' memberships for evaluating RIMs	Crisp, linear, Gaussian-curve	For simulating how people perceive unknown 'true' relations

All these alternations multiplied into 486 unique treatments. Each treatment was applied to 10 independently generated datasets to produce stable average results, with each application called a *run*. In each run, rules were first extracted from data and tested against both unadjusted (with significance level α) and the statistically sound chi-square tests for productive rules at $\alpha=0.05$. The search space size of the test was s = 44,796 and the significance level was $\kappa=0.05/s=1.12\times10^{-6}$, computed according to Webb (2007).

4.2. Results

4.2.1. On true and spurious rules

Figure 2 illustrates the numbers of true and spurious rules and FWER against variations in data sizes, statistical soundness of tests and perceived memberships for fuzzy attributes. True and spurious rules were those accepted by statistical tests that were within and beyond predesigned productive rules, respectively. True rules only counted for 88 out of 118 predesigned rules containing near1 or near2, as the others were irrelevant to fuzzy nearness. Each plotted point was an aggregation for all spatial patterns of objectives and resources, true $\mu_{\text{Good}}(\text{outcome})$ and extraneous factors, as the plots had quite similar patterns when these settings varied.

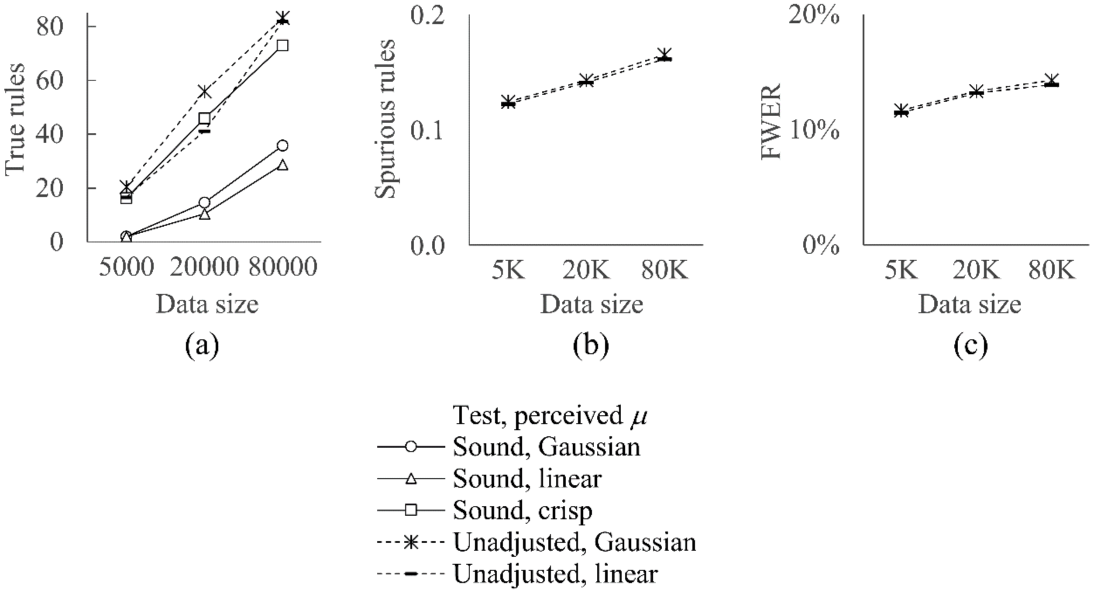


Figure 2. Abundance of true rules and avoidance of spurious rules for synthetic data experiment.

Statistically sound tests are proven indispensable for both crisp and fuzzy rules to strictly control spurious rules (Figure 2b, c). In crisp rule treatments, unadjusted tests resulted in dozens of spurious rules per run and 100% FWER, which were too large to be plotted. While this risk largely reduced in Gaussian-curve and linear fuzzy rule treatments, after unadjusted tests the FWER was still 10~15%, far above 5% as user specified by setting $\alpha = 0.05$. Although not shown in the figure, more than half of the runs produced rules with *p* values between 0.05 and 1. Thus, unadjusted tests with $\alpha = 0.1$ will produce over 50% FWER, while 10% was what the user expected. This is likely to be unacceptable in practice, as users cannot know the maximum risk of spurious rules under the significance level they set. Meanwhile, statistically sound tests did not accept any spurious rules, and thus their plots are absent from Figure 2b and c. Studies on ordinary SARM (Webb 2007, Zhang *et al*. 2016) reported 0.01~0.1% spurious rules and 0.1~1% FWER for statistically sound tests when $\alpha = 0.05$, which are still only several tenths of those for unadjusted tests on fuzzy rules, and much lower than the user specified 5% maximum FWER. In this study, this approach produced even fewer (actually zero) spurious rules, probably because the rule consequents were limited to outcome which was most related to other attributes.

On the condition that statistically sound tests were necessary, using crisp memberships exhibited great superiority in discovering more rules. Averaging results of all data sizes, statistically sound tests using crisp memberships discovered 2~4 times as many rules as using fuzzy ones, and comparable number of rules as unadjusted tests using fuzzy memberships (Figure 2a).

Overall, the results suggest that conducting statistically sound tests with crisp memberships in SARM is the best for finding abundant rules while maintaining strict control over spurious rules.

4.2.2. On RIM accuracy

Figure 3 shows the accuracy of RIM values against varying extraneous factors. Variations in other experiment settings made little difference in the changing trend of RIM accuracy. The plotted values are percentage errors of RIM values compared with their values computed using 'true' memberships of fuzzy attributes (see Table 4) and 0% extraneous factors.

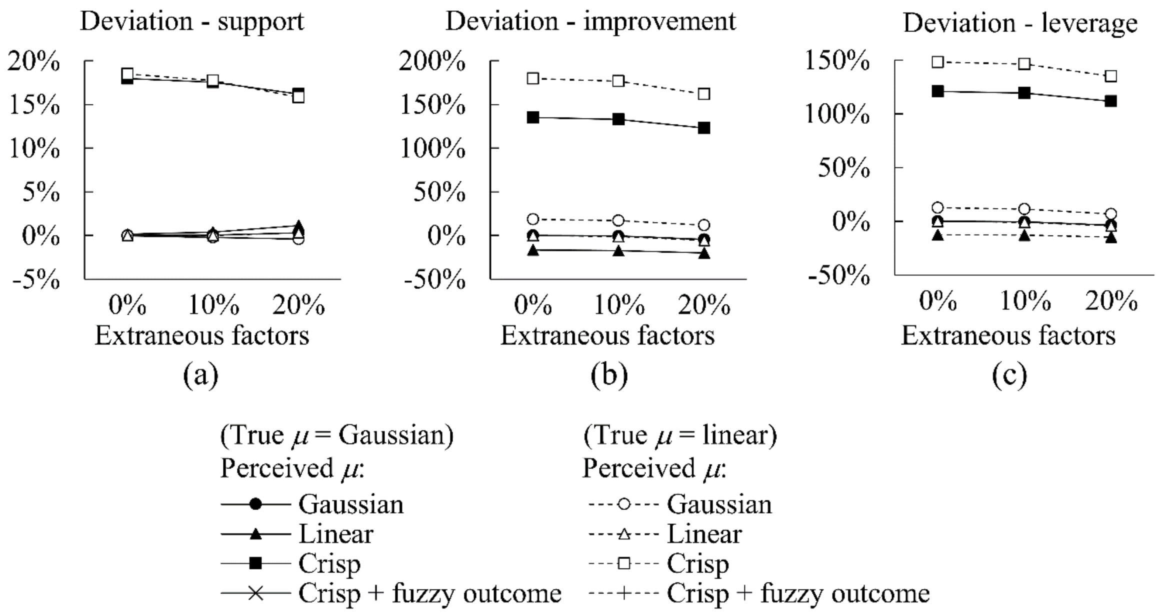


Figure 3. Rule Interestingness Measure (RIM) accuracies for the synthetic data experiment.

Crisp rules committed 15~20% positive error on supports and over 100% errors on improvements and leverages (Figure 3). This confirms the inference in 3.2 about large exaggerations on RIMs by the crisp membership, especially on RIMs other than support which are usually more useful for decision support. As extraneous factors grew, RIM values generally decreased, except for minor increases in fuzzy supports (Figure 3). This conforms to the expectation that extraneous factors blur data associations and weaken rules.

Gaussian-curve and linear perceived memberships caused positive and negative RIM errors, respectively, when true memberships were the opposite. Yet such errors were only 10~20% of those in counterpart crisp treatments. Thus, in RIM evaluation, by replacing crisp membership for gradual or vague concepts with fuzzy ones, there can be a major improvement in

RIM accuracy, even when true memberships might not be accurately defined based on existing expert knowledge.

When the true membership is unknown, the proposed Gaussian-curve model is still recommended for the RIM evaluation stage. First, as suggested in 3.1, Gaussian-curve memberships are widely regarded as a better illustration of linguistic concepts. Second, Gaussian-curve memberships produced larger RIM values than linear ones (Figure 3), since linear membership degrees change more constantly (with constant slopes) across transitions between concepts. Reductions in RIM values, due to practically inevitable extraneous factors, may partially offset positive errors in RIM values caused by the Gaussian-curve model while enlarging negative errors committed by the linear model.

5. Experiment with real-world data

5.1. Data and methods

This case study investigated the association between the popularity of points of interest (POIs) and their location factors in location-based social networks (LBSNs). Findings about such associations can be very useful for POI popularity prediction, business site selection, and POI recommendation.

The experiment data included the Foursquare POI and user check-in data from April 2012 to September 2013 in New York City (NYC) (Yang *et al*. 2016)[1], and NYC road network (OpenStreetMap contributors 2017) and subway entrances (Metropolitan Transportation Authority 2017) in ESRI Shapefiles. Each POI record contained geographical coordinates and the type of a POI, and each check-in record included the POI-ID, user-ID, and timestamp.

[1] Data is available at: https://drive.google.com/file/d/0BwrgZ-IdrTotZ0U0ZER2ejI3VVk/view

POIs of six types at the highest level of Foursquare official POI type taxonomy were taken as the *objectives* to be investigated. The number of check-ins (#checkins) at each objective was counted from the data and used to indicate its popularity. Location factors of each objective were distances from it to nearest roads, subway entrances and 14 types of *neighbors*. The neighbors were POIs whose types were also defined based on Foursquare POI types. The objective dataset was then constructed by adding #checkins and location factor values to objective POI records. Attributes in the objective data are listed in Table 5, and a part of data is mapped in Figure 4.

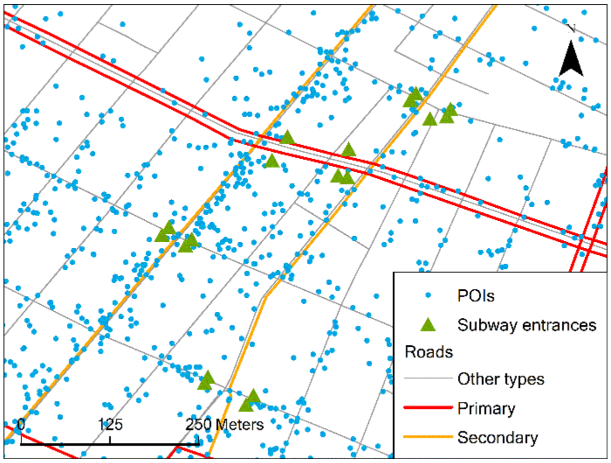


Figure 4. Map of the study area.

Table 5. Attributes of objective dataset.

		Attribute name	Description
General	1	ID	
	2	type	Objective type, including: 1. Arts & entertainment (3117 POIs) 2. Food (15,496 POIs) 3. Nightlife spot (3520 POIs) 4. Outdoors & recreation (4568 POIs) 5. Shop & service (10,680 POIs) 6. Travel & transport (4120 POIs)
	3	#checkins	No. of check-ins
Distance to transport facilities	1	d_road_primary	Distance to nearest trunk/primary road
	2	d_road_secondary	Distance to nearest trunk/primary/secondary road
	3	d_subway_1, d_subway_5	Distances to 1st and 5th nearest subway entrances
Distance to neighbor POIs	1	d_arts_1, d_arts_5	(Distances to 1st and 5th nearest POIs of following type; the same below) arts & entertainment, excluding POIs in neighbor type 3
	2	d_college_1, d_college_5	College & university
	3	d_stadium_park_1, d_stadium_park_5	Stadium, theme park and zoo
	4	d_food_1, d_food_5	Food
	5	d_nightlife_1, d_nightlife5	Nightlife spot
	6	d_outdoors_1, d_outdoors_5	Outdoors & recreation
	7	d_ reside_1, d_reside_5	Residence
	8	d_shop_1, d_shop_5	Shop & service
	9	d_mall_1, d_mall_5	Mall/department store
	10	d_gov_1, d_gov_5	Government building
	11	d_medical_1, d_medical_5	Medical center
	12	d_office_1, d_office_5	Office
	13	d_school_1, d_school_5	School
	14	d_spiritual_1, d_spiritual_5	Spiritual center

On the objective data of each type, ordinary, crisp-fuzzy and conventional fuzzy SARM were conducted. For each objective type and SARM approach, different schemes were assessed for building data discretization models, and the one yielding the largest number of significant rules was finally used. Each attribute was discretized into two or three concepts, that is, 'near, far' or 'near, mid, far' for distances, and 'low, high' or 'low, mid, high' for #checkins. Concept boundaries for the crisp discretization were determined using equisize and head/tail breaks classification (Jiang 2013). For simplicity, the scheme was the same for distances to roads, and the same for other distance attributes. For the matching fuzzy discretization model, the membership functions were determined to maximize the proportion of transitions in each concept as in the synthetic data experiment.

For each objective, rules like 'distance attribute values→ #checkins = high' were explored and examined using the statistically sound chi-square test for productive rules at α=0.05. The rules might contain one item in the antecedent for objective types 1, 4 and 5, and up to two items for other types. The test employed Holm Procedure, during which only attributes in rules already accepted or being evaluated were taken into the search space.

The experiment results of crisp-fuzzy SARM were also compared with those from regression and *M* index computation (Marcon and Puechy 2010), two common analyses for evaluating locations of business POIs. The results of these are given in section 5.3.

5.2. Results

The experiment results reconfirmed the advantages of crisp-fuzzy SARM in obtaining abundant rules and accurate RIM values. Thanks to the statistical test based on crisp data discretization, ordinary and crisp-fuzzy SARM discovered nearly 50% more rules than fuzzy SARM (Table 6). Also, crisp-fuzzy SARM avoided the large exaggerations of RIM values in ordinary SARM caused by crisp data discretization for gradual concepts, which should be the main reason for the around 50% average discrepancies in rule leverages of ordinary and crisp-fuzzy SARM results (Table 6).

Table 6. Real data experiment result on number of significant rules and RIM accuracy.

	Objective type						Average
	1	2	3	4	5	6	
Max. No. of items in rule antecedent	1	2	2	1	1	2	
(a) No. of significant rules							
Ordinary/crisp-fuzzy SARM	22	68	71	15	27	13	36 (44% larger)
Fuzzy SARM	11	60	56	5	17	1	25
(b) Average leverage of significant rules							
Ordinary SARM	0.0159	0.0154	0.0319	0.0105	0.0103	0.0106	0.0157 (50% larger)
Crisp-fuzzy SARM	0.0125	0.0109	0.0220	0.0075	0.0060	0.0032	0.0104

Table 7 lists the top-10 rules in terms of leverage discovered by crisp-fuzzy SARM for each objective type. We select leverage for ranking the rules and the importance of location factors, as leverage measures the total gain of popularity by all objectives of a certain type compared with if the popularity is irrelevant to the location factors. Discoveries from the top-10 and other rules are summarized below. Note that many of these discoveries are made only by crisp-fuzzy SARM: ordinary SARM resulted in very different ranks of rules (Table 7), due to heterogeneous exaggerations of leverages among the rules, while many rules to be analyzed were missing in the conventional fuzzy SARM result.

Table 7. Top-10 rules for each objective type in terms of leverage, and some other rules for food venues.

Rank		Rule antecedent	Leverage
Fuzzy	Ordinary	(all with consequent '#checkins=high')	(fuzzy)
(a) Arts & entertainment		Scheme: Chkin3equisize_Rd3_Nb2[a]	
1	5	d_arts_1 = near	0.0176
2	3	d_food_5 = near	0.0163
3	8	d_food_1 = near	0.0159
4	1	d_nightlife_5 = near	0.0158
5	2	d_arts_5 = near	0.0158
6	6	d_office_5 = near	0.0150
7	10	d_outdoors_1 = near	0.0135
8	9	d_shop_5 = near	0.0133
9	4	d_outdoors_5 = near	0.0131
10	12	d_office_1 = near	0.0130
(b) Food		Scheme: Chkin3equisize_Rd3_Nb2	
1	1	d_nightlife_5 = near	0.0204
2	9	d_nightlife_1 = near	0.0181
3	2	d_food_5 = near	0.0180
4	5	d_office_1 = near	0.0179
5	4	d_arts_5 = near	0.0170
6	20	d_arts_1 = near	0.0154
7	3	d_shop_5 = near	0.0150
8	22	d_food_1 = near	0.0146
9	21	d_office_1 = near	0.0144
10	8	d_spiritual_5 = near	0.0138
20*	31	d_office_5 =mid ∧ d_road_primary = mid	0.0126
22*	28	d_nightlife_5 = near ∧ d_road_primary = mid	0.0125
30*	39	d_ food_5 = near ∧ d_road_primary = mid	0.0116
(c) Nightlife spot		Scheme: Chkin3equisize_Rd2_Nb2	
1	1	d_nightlife_5 = near	0.0345
2	27	d_nightlife_1 = near	0.0300
3	13	d_food_5 = near	0.0294
4	20	d_nightlife_5 = near ∧ d_road_primary = near	0.0294
5	18	d_arts_1 = near	0.0292
6	6	d_arts_5 = near	0.0282
7	16	d_nightlife_1 = near ∧ d_arts_1 = near	0.0280
8	37	d_food_1 = near	0.0275
9	29	d_arts_1 = near ∧ d_road_primary = near	0.0275
10	31	d_food_5 = near ∧ d_road_primary = near	0.0265

(d) Outdoors & recreation			Scheme: Chkin2headtail_Rd3_Nb2
1	6	d_shop_1 = near	0.0085
2	5	d_food_1 = near	0.0084
3	3	d_nightlife_5 = near	0.0084
4	1	d_shop_5 = near	0.0083
5	12	d_road_secondary = near	0.0082
6	2	d_food_5 = near	0.0081
7	4	d_gov1 = near	0.0078
8	11	d_office_1 = near	0.0077
9	9	d_arts_5 = near	0.0074
10	7	d_outdoors_5 = near	0.0073
(e) Shop & service			Scheme: Chkin2equisize_Rd3_Nb2
1	4	d_office_1 = near	0.0082
2	7	d_nightlife_1 = near	0.0076
3	8	d_shop_5 = near	0.0075
4	5	d_food_5 = near	0.0074
5	1	d_college_1 = near	0.0074
6	2	d_medical_5 = near	0.0072
7	18	d_shop_1 = near	0.0071
8	6	d_nightlife_5 = near	0.0070
9	11	d_medical_1 = near	0.0070
10	10	d_college_5 = near	0.0069
(f) Travel & transport			Scheme: Chkin2headtail_Rd3_Nb2
1	10	d_office_1 = near	0.0075
2	1	d_nightlife_5 = near	0.0074
3	6	d_nightlife_1 = near	0.0068
4	3	d_nightlife_5 = near $\wedge$ d_road_secondary = mid	0.0067
5	8	d_food_5 = near	0.0066
6	11	d_spiritual_1 = near	0.0065
7	12	d_office_5 = near	0.0062
8	9	d_office_1 = near $\wedge$ d_road_secondary = mid	0.0062
9	5	d_food_5 = near $\wedge$ d_road_secondary = mid	0.0059
10	4	d_spiritual_1 = near $\wedge$ d_road_secondary = mid	0.0058

[a] Equisize classification with 3 concepts for #checkins, 3 concepts for distance to roads (Rd), and 2 concepts for neighbors (Nb). Ditto for other objective types.

(1) Nearness to POIs of the same type

We call a location factor *favorable* if it is associated with high popularity of the objectives. For most objective types, nearness to neighbors of the same type is highly favorable (Table 7a, rule 1, 5; 8b, rule 3, 8; 8c, rule 1, 2; 8e, rule 3, 7). Apparently, these objectives benefit from an agglomeration effect, that is, consumers tend to prefer spatially agglomerated business sites of a certain type over dispersed sites (De Beule *et al*. 2015). This effect was originally reported for retail stores, and studies have seldom compared the degrees of such effect on shops and other business sectors. Judged by ranks of corresponding rules, this study shows that during the study period, the agglomeration effect in NYC was the strongest for nightlife spots, followed by arts & entertainment spots, shops and food venues. Thus, new nightlife spots are most strongly recommended to be located around established nightlife hotspots for higher potentials to become popular.

(2) Nearness to neighbors of other types

Among the 14 neighbor types, nearness to food venues, shops, nightlife spots, arts & entertainment spots and offices are generally the most favorable for the popularity of objectives (Table 7). Somewhat unexpectedly, proximity to nightlife spots is often a top factor, except for arts & entertainment and outdoor spots. It is known that people tend to visit most outdoor venues and many arts & entertainment spots (such as museums and zoos) in the daytime. Compared with extremely common food venues and shops, clustering of nightlife spots seems a better indicator of prosper business and entertainment areas attracting large visitor flows. It is recommended to carefully consider nearby nightlife spots when selecting business sites in NYC and similar cities.

Key factors for POI popularity also vary with objective types (Table 7). Interested parties may refer to rules for these factors with importance ranks for each objective type, and see whether the nearest neighbor (suggested by d_1) or the richness of surrounding neighbors (suggested by d_5) is more meaningful for indicating popular objectives. Rules with two factors in the antecedent suggest that objectives at locations with both factors have statistically significantly higher potentials to be popular than those at locations with either factors. It can be worthwhile for business owners to pay extra costs (such as higher rents) to acquire potential business sites with both these factors.

(3) Accessibility to transport facilities

Distances from business sites to transport facilities direct influence their visibility and accessibility by users, which are top factors for business site popularity (Roig-Tierno *et al*. 2013). Yet the resultant rules suggest only moderate influence of transport facilities, as distances to major roads and subways appeared infrequently in top-ranked rules. One reason may be that transportation in NYC is generally quite convenient: 80% of the POIs are within 500m to primary-or-above roads and 200m to secondary-or-above ones. This discovery suggests that in modern metropolises with convenient transportation, the relative importance of transportation convenience and neighboring POIs to the success of businesses may need to be reevaluated.

High popularity of objectives was mostly associated with nearness to major roads as expected, yet it was associated with medium d_road_primary for food venues and d_road_secondary for travel & transport venues (Table 7b, f). Indeed, popular travel & transport venues are sometimes away from major roads to meet people’s demand or avoid noises. For food venues, however, the result deviates from common belief and the literature. The interval between transition midpoints of $\mu_{\text{mid}}\left(\text{d\_road\_primary}\right)$ was 80~351m, which might not suggest poor accessibility, but should impact the venues' visibility compared with those along primary roads. It is worth further investigation that if such impaired visibility is largely offset by online word-of-mouth in metropolises with intensive use of LBSNs.

5.3 Comparison with alternative analytics

(1) Regression

For each objective type, linear regression was performed on #checkins with each location factor as the independent variable. All variable values were transformed to base-10 logarithms since their distributions were skewed to large values. Nearly all regression coefficients were negative (Table 8), showing that nearness to corresponding neighbors is favorable for POI popularity. For travel & transport venues, however, the coefficient of d_road_secondary was positive and significant (p=0.026), suggesting that remote POIs are more popular. This should be a misunderstanding, as crisp-fuzzy SARM reveals that d_road_secondary = mid is favorable for the popularity of these POIs (Table 7f). As the factors have small R^2 (Table 8) and weak correlations to POI popularity, it is hard to avoid this misunderstanding by observing the data. A quadratic regression against #checkins with d_road_secondary as the independent variable resulted in a regression function with negative quadratic coefficient and a vertex corresponding to d_road_secondary=77m. This function agreed to the crisp-fuzzy SARM result where the interval between transition midpoints of $\mu_{\text{mid}}\left(\text{d\_road\_secondary}\right)$ was 14~164m. However, without the reconfirmation by crisp-fuzzy SARM, it could be difficult for users to trust the quadratic regression result, due to the opposite and statistically significant linear regression result.

Table 8. Summary of regression results.

	Linear regression: $\log_{10}$(#checkins+1) = $\log_{10}(\beta$(d*) +1)+c [a]	
	Average β	Average R^2
1. Arts & entertainment	-0.0737	0.0056
2. Food	-0.0736	0.0067
3. Nightlife spot	-0.1366	0.0161
4. Outdoors & recreation	-0.0451	0.0021
5. Shop & service	-0.0356	0.0024
6. Travel & transport	-0.0316	0.0016

[a] d* are distance attributes (location factors) in Table 5. #checkins and d* values were shifted by 1 to avoid taking logarithms on zero values.

Further, if multiple location factors are regressed against POI popularity, it could be difficult to find the redundant factors that cannot significantly improve the model fitness, and thus do not deserve extra costs paid by business owners. A factor is obviously non-redundant if it markedly improves the adjusted R^2. Yet in this and many other studies, individual factors account for only small R^2, and can slightly improve the adjusted R^2 at best. To our knowledge, no method could evaluate whether such improvement of adjusted R^2 is statistically significant and thus likely to hold in future. For SARM, redundant factors can be identified by statistically sound tests for productive rules, as shown in 5.2.

(2) *M* Index

The *M* index between business sectors (such as store types) A and B is

$$M_{\mathrm{AB}} = \left(\overline{n_{\mathrm{B}}(\mathrm{A})/N(\mathrm{A})}\right)\Big/\left(n_{\mathrm{B}}/N\right), \qquad (4)$$

where $\left(\overline{n_{\mathrm{B}}(\mathrm{A})/N(\mathrm{A})}\right)$ is the average proportion of B stores among all stores around (within a distance r to) each A store, and $\left(n_{\mathrm{B}}/N\right)$ is the proportion of B stores in the study area. The

Jensen's Quality (Jensen 2006) of location *p* for opening an A store is such computed that each B store around *p* increases the quality if $M_{\mathrm{AB}} > 1$, and reduces the quality otherwise. Later, this quality measure was expanded to broader POI types (Karamshuk *et al*. 2013, Lian *et al*. 2017).

Table 9 shows the *M* values between POI types in this study computed with *r*=200m following Karamshuk *et al*. (2013). The result at *r*=100m computed following Jensen (2006) showed the same patterns. Even most neighbor types were found favorable for the popularity of objectives by crisp-fuzzy SARM and regression, many neighbors had *M*<1 for all objectives, showing that they geographically 'repel' those objectives. Indeed, most of these neighbors, like colleges and schools (types 2 and 14), tend to locate in different urban functional areas from business POIs. According to the Jensen's Quality, larger numbers of such neighbors reduce the popularity of nearby POIs, which is somehow unreasonable. Hence, the *M* index is not so suitable as SARM for investigating the location factors in this study.

Table 9. *M* values of objectives at *r*=200m.

Objective type	Neighbor type													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Arts & entertainment	2.67	0.67	1.37	0.89	1.10	0.67	0.58	0.86	0.61	0.69	0.56	1.64	0.36	0.51
2. Food	1.06	0.69	0.44	1.11	1.13	0.55	0.62	1.00	0.71	0.66	0.75	1.63	0.38	0.54
3. Nightlife spot	1.22	0.61	0.46	1.07	1.62	0.55	0.68	0.93	0.57	0.60	0.64	1.48	0.38	0.50
4. Outdoors & recreation	1.26	0.81	0.88	0.93	0.98	1.02	0.69	0.95	0.82	0.99	0.77	1.81	0.43	0.54
5. Shop & service	1.05	0.71	0.39	0.97	0.98	0.56	0.56	1.23	1.14	0.65	0.78	1.78	0.34	0.49
6. Travel & transport	1.30	0.73	0.68	1.04	1.07	0.62	0.64	0.95	0.97	0.84	0.76	1.80	0.41	0.61

The above comparisons show that crisp-fuzzy SARM can better reveal interactions between location factors and POI popularity in several aspects than alternative analytics. While crisp-fuzzy SARM cannot directly predict POI popularity, it can generate improved understandings of data interactions that are essential for proper use of predictive analytics and accurate predictions.

6. Conclusions

This article presents crisp-fuzzy SARM, a new method for improving the reliability of SARM results. This method integrates statistically sound tests on crisp rules and evaluation of RIMs based on fuzzy supports. In the RIM evaluation, the method also adopts an original Gaussian-curve-based fuzzy data discretization model with enhanced spatial semantics and mathematical justifications compared with existing models. This method can achieve all-round improvement in the reliability of SARM results, by combining abundance of rules in ordinary SARM, high RIM accuracy of fuzzy SARM and low risk of spurious rules attained by statistically sound tests.

The proposed method can markedly increase the resultant rules, typically by 50% or more compared with conventional fuzzy SARM. The method also greatly reduces the positive errors in RIM values caused by crisp data discretization for gradual or vague concepts. Such errors are usually at least 50% for representative RIMs in ordinary SARM. The risk that the entire SARM result contains any spurious rules is below 1%. A case study on POI popularity in LBSNs demonstrates the practical value of crisp-fuzzy SARM for data analysis and decision support.

Albeit improved by the new method, SARM results have further reliability concerns. First, past studies have justified Gaussian-curve-based data discretization models mostly for point-to-point distances. Future studies need to evaluate Gaussian-curve-based models, or find better models for other proximity measures such as point-to-polyline and polygon-to-polygon proximity (Laube *et al*. 2008), as well as directional and topological relations.

Second, apart from mathematical forms of membership functions, SARM results also closely rely on specific data discretization schemes, including the number of concepts and the membership function for each concept. However, expert knowledge for identifying proper discretization schemes is often inadequate, which poses a substantial risk of producing poor

SARM results. A promising solution to this problem lies in evolutionary algorithms (EAs) for association rule mining (Alhajj and Kaya 2008, Chen *et al*. 2008, Carmona *et al*. 2010). EAs can find membership functions that achieve near-optimum for one or multiple objectives, based on data distribution alone, thereby constructing data discretization schemes that are appropriate for specific user demands. The objectives can be RIMs or other constraints. Much research is needed to evaluate the efficacy of and develop new algorithms for EA-based crisp-fuzzy SARM with statistically sound tests.

Acknowledgements

We would like to thank the editor Dr. Shawn Laffan and the anonymous reviewers for their insightful comments and substantial help on improving this article. We also thank Prof. Jose Torero Cullen, Prof. Asif Usmani and Dr. Xu Yang for suggestions on the case study, and Miss Zhou Xiaolin for English editing. This study is funded by the National Natural Science Foundation of China (41331175), Ministry of Science and Technology (2017YFB0503604), and Hong Kong Polytechnic University (1-ZE24, 4-ZZFZ).

Appendix 1. Chi-square test for productive association rules

An association rule $X \to Y$ is productive if $imp(X \to Y) > 0$. Unproductive rules contain redundant items in X that cannot improve $conf\,(X \to Y)$, or cannot strengthen the association in $X \to Y$, thus mostly should be removed from SARM results. Referring to the definition of $imp(X \to Y)$ in 2.1, the productivity of $X \to Y$ can be tested by:

$$\forall Z \subseteq X,\ \Pr\left(Y \mid X\right) > \Pr\left(Y \mid X \setminus Z\right). \qquad (5)$$

This study follows accepted practice (Webb 2007) to conduct a more computationally economic test, the result of which is quite similar to that of testing (5), on

$$\forall m = 1 \ldots p,\ \Pr\left(Y \mid X\right) > \Pr\left(Y \mid X \setminus \left\{x_m\right\}\right) \qquad (6)$$

for $X = \left\{x_1 \ldots x_p\right\}$. The corresponding null hypothesis is $\exists x \in X,\ \Pr\left(Y \mid X\right) \le \Pr\left(Y \mid X \setminus \left\{x_m\right\}\right)$, suggesting that $conf\,(X \to Y) > conf\,(X \setminus \left\{x_m\right\} \to Y)$ purely by chance.

(6) can be evaluated the chi-square test commonly used for testing likewise conditions:

$$\chi^2 = \frac{\left(ad - bc\right)\left(a + b + c + d\right)}{\left(a + b\right)\left(c + d\right)\left(a + c\right)\left(b + d\right)}, \qquad (7)$$

where

$$\begin{aligned} a &= supp\left(X \bigcup Y\right) \\ b &= supp\left(X \bigcup \neg Y\right) \\ c &= supp\left(\left(X \setminus \left\{x_m\right\}\right) \bigcup \neg \left\{x_m\right\} \bigcup Y\right) \\ d &= supp\left(\left(X \setminus \left\{x_m\right\}\right) \bigcup \neg \left\{x_m\right\} \bigcup \neg Y\right) \end{aligned}, \qquad (8)$$

"\" denotes set difference, and "¬" means that the record does not contain x_m or all items in Y. For each $x_m \in X$, a possibility p_m is looked up from the χ^2 table for the computed χ^2 value with one degree of freedom. The final p value of the test, equal to $\max\left(p_m\right)$, is the probability that $X \to Y$ has observed value of $imp(X \to Y)$ in data if the null hypothesis is true, and is equivalent to the risk that $X \to Y$ is spurious. $X \to Y$ is accepted by the test as reflecting a real-world association if p is below the significance level, such as 0.05, otherwise it is rejected and removed from resultant rules.

Fisher exact test (Agresti 1992) is a popular alternative for evaluating (6), yet it only handles integral supports and thus is inapplicable to fuzzy SARM. Chi-square is less accurate than the Fisher exact test for crisp rules in small datasets, due to the approximation of integral supports by continuous χ^2 distribution (McDonald 2014). Yet this no longer holds for fuzzy supports which are continuous in nature. This study applies the chi-square test to both crisp and fuzzy rules, for fair comparisons between results of corresponding experiment groups.

References

Agrawal, R., Imielinski, T., and Swami, A., 1993. Mining associations between sets of items in massive databases. IP. Buneman and S. Jajodia, eds. *1993 ACM-SIGMOD International conference on management of data*. May 25-28 1993 Washington, DC, USA. New York: ACM, 207–216.

Agrawal, R., and Srikant, R., 1994. Fast algorithms for mining association rules. In: J.B. Bocca, M. Jarke, and C. Zaniolo, eds. *Proceedings of the 20th International Conference on Very Large Data Bases*, 12-15 September 1994 Santiago de Chile, Chile. San Francisco: Morgan Kaufmann Publishers, 487–499.

Agresti, A., 1992. A survey of exact inference for contingency tables. *Statistical Science*, 7 (1), 131–153. doi:10.1214/ss/1177011454

Alhajj, R. and Kaya, M., 2008. Multi-objective genetic algorithms based automated clustering for fuzzy association rules mining. *Journal of Intelligent Information Systems*, 31, 243–264. doi:10.1007/s10844-007-0044-1

Baralis, E., et al., 2012. Generalized association rule mining with constraints. *Information Sciences*, 194, 68–84. doi:10.1016/j.ins.2011.05.016

Bay, S.D. and Pazzani, M.J., 2001. Detecting group differences: mining contrast sets. *Data Mining and Knowledge Discovery,* 5 (3), 213–246. doi:10.1023/A:1011429418057

Bayardo Jr., R.J., Agrawal, R., and Gunopulos, D., 2000. Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, 4 (2/3), 217–240. doi:10.1023/A:1009895914772

Bogorny, V., Kuijpers, B., and Alvares, L.O., 2008. Reducing uninteresting spatial association rules in geographic databases using background knowledge: a summary of results. *International Journal of Geographical Information Science,* 22 (4), 361–386. doi:10.1080/13658810701412991

Bordogna, G., Carrara, P., and Pasi, G., 1991. Query term weights as constraints in fuzzy information retrieval. *Information Processing & Management*, 27 (1), 15–26. doi:10.1016/0306-4573(91)90028-K

Bordogna, G. and Pasi, G., 1993. A fuzzy linguistic approach generalizing Boolean information retrieval: a model and its evaluation. *Journal of the American Society for Information Science*, 44 (2), 70–82. doi:10.1002/(ISSN)1097-4571

Bosc, P., et al., 2007. Adjusting the core and/or the support of a fuzzy set - A new approach to fuzzy modifiers. In: *IEEE International Fuzzy Systems Conference 2007*, 23–26 July 2007 London, 1–6.

Burda, M., Pavliska, V., and Valasek, R., 2014. Parallel mining of fuzzy association rules on dense data sets. In: *2014 IEEE International Conference on Fuzzy Systems*, 6–11 July 2014 Beijing.

Carmona, C.J., et al., 2010. NMEEF-SD: non-dominated multiobjective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. *IEEE Transactions on Fuzzy Systems*, 18 (5), 958–970. doi:10.1109/TFUZZ.2010.2060200

Chen, C., et al., 2008. A multi-objective genetic-fuzzy mining algorithm. In: *2008 IEEE International Conference on Granular Computing*, 26–28 August 2008 Hangzhou, 115–120.

De Beule, M., Van den Poel, D., and Van de Weghe, N., 2015. Assessing the principles of spatial competition between stores within a retail network. *Applied Geography*, 62, 125–135. doi:10.1016/j.apgeog.2015.04.015

Faridi, M., Verma, S., and Mukherjee, S., 2017. A novel algorithm of weighted fuzzy spatial association rule mining (WFSARM) for wasteland reclamation, *Journal of Information and Optimization Sciences*, online first. doi:10.1080/02522667.2017.1372920.

Farzanyar, Z. and Kangavari, M., 2012. Efficient mining of fuzzy association rules from the preprocessed dataset. *Computing and Informatics*, 31, 331–347.

Herrera, F. and Martinez, L., 2000. A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Transactions on Fuzzy Systems*, 8 (6), 746–752. doi:10.1109/91.890332

Holm, S., 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.

Hüllermeier, E., 2009. Fuzzy methods in data mining. In: J. Wang, ed. *Encyclopedia of data warehousing and mining*. 2nd ed. Hershey, PA: IGI Global, 907–912.

Jensen, P., 2006. Network-based predictions of retail store commercial categories and optimal locations. *Physical Review E,* 74 (3), 035101. doi:10.1103/PhysRevE.74.035101

Jiang, B., 2013. Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *The Professional Geographer,* 65 (3), 482–494. doi:10.1080/00330124.2012.700499

Karamshuk, D., et al., 2013. Geo-spotting: mining online location-based services for optimal retail store placement. In: *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2013 Chicago, 793–801.

Ladner, R., Petry, F.E., and Cobb, M.A., 2003. Fuzzy set approaches to spatial data mining of association rules. *Transactions in GIS,* 7 (1), 123–138. doi:10.1111/tgis.2003.7.issue-1

Laube, P., Berg, M., and Kreveld, M., 2008. Spatial support and spatial confidence for spatial association rules. In: *The 13th international symposium on spatial data handling: headway in spatial data mining*. Springer, 575–593.

Lian, J., et al., 2017. Restaurant survival analysis with heterogeneous information. In: *The 26th international conference on World Wide Web companion,* 3–7 April 2017 Perth, Australia, 993–1002.

Liu, B., Hsu, W., and Ma, Y., 1999. Pruning and summarizing the discovered associations. A. Ruas and C. Gold, eds. *The 13th international symposium on spatial data handling: headway in spatial data mining,* 23-25 July 2008 Montpellier, France. Berlin: Springer-Verlag, 575–593..

Liu, B., Hsu, W., and Ma, Y. 2001. Identifying non-actionable association rules. In: D. Hsu, et al., eds. *The 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 26– 29 August 2001 San Francisco. New York: ACM, 329–334 doi:10.1145/502512

Marcon, E. and Puechy, F., 2010. Measures of the geographic concentration of industries: improving distance-based methods. *Journal of Economic Geography,* 10, 745–762. doi:10.1093/jeg/lbp056

McDonald, J.H., 2014. *Handbook of biological statistics*. 3rd ed. Baltimore: Sparky House Publishing.

Megiddo, N. and Srikant, R., 1998. Discovering predictive association rules. In: *The fourth international conference on knowledge discovery and data mining*. Menlo Park: AAAI, 27–78.

Mennis, J. and Liu, J., 2005. Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change. *Transactions in GIS*, 9 (1), 5–17. doi:10.1111/tgis.2005.9.

Metropolitan Transportation Authority, 2017. *Subway entrances*. Available from: https://data.cityofnewyork.us/Transportation/Subway-Entrances/drex-xx56 [Accessed 17 August 2017].

Mitchell, A., 2005. *The ESRI guide to GIS analysis, volume 2: spatial measurements and statistics*. Redlands: ESRI Press.

OpenStreetMap contributors, 2017. Planet dump [Data file from 21 September 2017]. Available from: https://planet.openstreetmap.org.

Piatetsky-Shapiro, G., 1991. Discovery, analysis, and presentation of strong rules. In: G. Piatetsky-Shapiro and J. Frawley, eds. Knowledge discovery in databases. Menlo Park: AAAI/MIT Press, 229–248.

Robinson, V.B., 2000. Individual and multipersonal fuzzy spatial relations acquired using human-machine nteraction. *Fuzzy Sets and Systems*, 113, 133–145. doi:10.1016/S0165-0114(99)00017-2

Roig-Tierno, N., et al., 2013. The retail site location decision process using GIS and the analytical hierarchy process. *Applied Geography*, 40, 191–198. doi:10.1016/j.apgeog.2013.03.005

Tew, C., et al., 2014. Behavior-based clustering and analysis of interestingness measures for association rule mining. *Data Mining and Knowledge Discover*y, 28 (4), 1004–1045.

Verhein, F. and Chawla, S., 2008. Mining spatio-temporal patterns in object mobility databases. *Data Mining and Knowledge Discovery*, 16, 5–38. doi:10.1007/s10618-007-0079-5

Versichele, M., et al., 2014. Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: a case study of Ghent, Belgium. *Tourism Management*, 44, 67–81. doi:10.1016/j.tourman.2014.02.009

Webb, G.I., 2007. Discovering significant patterns. *Machine Learning*, 68, 1–33. doi:10.1007/s10994-007-5006-x

Worboys, M.F., 2001. Nearness relations in environmental space. *International Journal of Geographical Information Science*, 15 (7), 633–651. doi:10.1080/13658810110061162

Wu, B., Li, R., and Huang, B., 2014. A geographically and temporally weighted autoregressive model with application to housing prices. *International Journal of Geographical Information Science*, 28 (5), 1186–1204. doi:10.1080/13658816.2013.878463

Xu, Y., et al., 2016. Another tale of two cities: understanding human activity space using actively tracked cellphone location data. *Annals of the American Association of Geographers*, 106 (2), 489–502.

Yang, D., Zhang, D., and Qu, B., 2016. Participatory cultural mapping based on collective behavior data in location based social networks. *ACM Transaction on Intelligent Systems and Technology*, 7, 3.

Zaki, M.J., 2000. Generating non-redundant association rules. In: R. Ramakrishnan, et al., eds. The 6th ACM SIGKDD international conference on knowledge discovery and data mining, 20–23 August 2000 Boston, MA, USA. New York: ACM, 34–43 doi:10.1145/347090

Zhang, A., Shi, W., and Webb, G.I., 2016. Mining significant association rules from uncertain data. *Data Mining and Knowledge Discovery*, 30 (4), 928–963. doi:10.1007/s10618-015-0446-6

Zhang, H., Padmanabhan, B., and Tuzhilin, A., 2004. On the discovery of significant statistical quantitative rules. In: *The tenth international conference on knowledge discovery and data mining*. New York: ACM, 374–383.