Spatial support and spatial confidence for spatial association rules

Patrick Laube[1], Mark de Berg[2], and Marc van Kreveld[3]

[1] Geomatics Department, The University of Melbourne, 3010 Parkville VIC, Australia, plaube@unimelb.edu.au
[2] Department of Mathematics and Computing Science, TU Eindhoven, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, mdberg@win.tue.nl
[3] Department of Computer Science, Utrecht University, P.O. Box 80.089, 3508 TB Utrecht The Netherlands, marc@cs.uu.nl

Abstract

In data mining, the quality of an association rule can be stated by its support and its confidence. This paper investigates support and confidence measures for spatial and spatio-temporal data mining. Using fixed thresholds to determine how many times a rule that uses proximity is satisfied seems too limited. It allows the traditional definitions of support and confidence, but does not allow to make the support stronger if the situation is "really close", as compared to "fairly close". We investigate how to define and compute proximity measures for several types of geographic objects—point, linear, areal—and we express whether or not objects are "close" as a score in the range $[0, 1]$. We then use the theory from so-called fuzzy association rules to determine the support and confidence of an association rule. The extension to spatio-temporal rules can be done along the same lines.

Keywords: *Spatial data mining, spatial association rule mining, fuzzy association rules, support, confidence.*

1 Introduction

Association rule mining (ARM) is one of the defining operations of data mining. The idea is best illustrated by the example of mining frequent item sets

in market-basket data (Agrawal *et al.* 1993). The task is finding sets of items that co-occur in user purchases more than a user-defined number of times (Gidofalvi and Pedersen 2005). A classical example of such an association rule is "If a transaction includes bread, then it includes butter." Of course, an association rule need not always be satisfied to be interesting: even though some transactions including bread do not include butter, it is still interesting to know that the rule holds for many transactions. This leads to the concepts of *support* (the number of transactions including bread and butter) and *confidence* (the fraction of all transactions including bread that also include butter) of association rules. Interesting rules are the ones where both the support and the confidence are high.

In a number of data-mining applications one has to deal with spatial and/or spatio-temporal data: crime hot-spot analysis, optimization of location-based services (LBS), public health and geomarketing applications (Gidofalvi and Pedersen 2005, Shekhar, Zhang, Huang and Vatsavai 2003) are important examples. The number of such applications is only growing because of the inexorable fusion of previously separate spatio-temporal data sources. When mining spatial data, spatial association rules are needed. A spatial association rule (sAR) is an association rule where at least one of the predicates is spatial (Shekhar and Huang 2001). Figure 1 illustrates such a spatial association rule. The rule captures a relation between *location* and *price* for items *houses*: "If a house is close to the river, then it is expensive."

Just as with conventional association rules, the quality of spatial association rules is determined by their support and confidence. When the rule uses proximity, one thus has to decide when objects are deemed "close". This is typically done by a Boolean distance buffer (Koperski and Han 1995, Miller and Han 2001): objects are close if their distance is less than some user-defined threshold d. In Figure 1, for example, the given threshold d_1 leads to a support of 6 and a confidence of 0.75. Note that a slightly smaller threshold d_2 would have given a support of 3 and a confidence of 1.0. Clearly, such strong sensitivity to a user-defined threshold is undesirable.

The thresholding problem not only arises for spatial relations. For example, the attribute "expensive" in the rule "If a transaction includes expensive wine, it also includes French cheese." is essentially non-binary. The same holds for attributes like "old", "tall", and so on. Hence, there has been some work on so-called *fuzzy association rules* (Kuok *et al.* 1998, Dubois *et al.* 2006). Here, instead of having a threshold that, for instance, defines whether wine

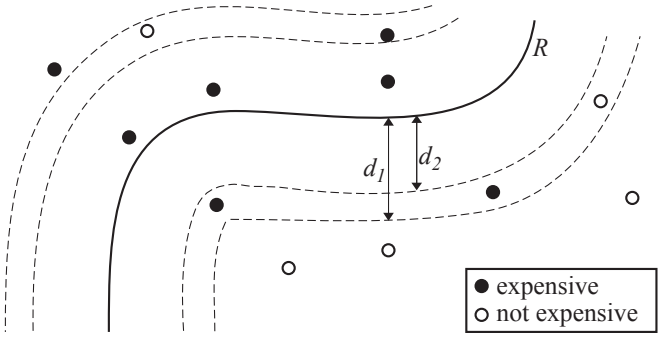

Figure 1: A set of expensive houses and a set of inexpensive houses, a river R, and two thresholds d_1 and d_2 to express "closeness". The rule "If a house is close to the river, then it is expensive" accounts for a support of 6 and a confidence of $6/8 = 0.75$ for d_1, and a support of 3 and a confidence of $3/3 = 1.0$ for d_2 respectively.

is expensive, the wine price is mapped to a score in the range $[0, 1]$. Then fuzzy-logic theory is used to measure support and confidence of an association rule—see the next section for details.

Given the variety and complexity of proximity relations between spatial entities, it should be clear that simple thresholding is often not the best approach for defining support and confidence in spatial association rules. Unfortunately, it seems that the much more appropriate concept of fuzzy association rules has been largely overlooked in spatial data mining: it is still standard practice to use thresholding when it comes to proximity relations. The main contribution of our paper is to investigate how fuzzy association rules can be applied in a spatial context. In particular, in this paper

- we explore the use of fuzzy association rules to spatial data mining,
- we investigate how to define suitable distance measures between various types of spatial objects—point, linear, areal—, we show how to compute these distance measures efficiently, and we discuss how to map these distance measures to scores in the range $[0, 1]$ for use in fuzzy spatial association rules,
- we extend our ideas to spatio-temporal association rules.

2 Background

2.1 Spatial data mining

Knowledge Discovery in Databases (KDD) is the overall process of discovering useful knowledge from data, and data mining is its most prominent step (Fayyad *et al.* 1996). Data mining is more specifically defined as "the application of specific algorithms for extracting patterns from data" (Fayyad *et al.* 1996, p.39). The need for spatial data mining and geographic knowledge discovery techniques has been widely acknowledged in both the GIScience (Miller and Han 2001, Roddick *et al.* 2001, Roddick and Lees 2001) and data-mining communities (Shekhar, Zhang, Huang and Vatsavai 2003). Spatial data mining is defined as the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial data sets (Shekhar, Zhang, Huang and Vatsavai 2003).

The challenges arising when mining spatial data include geographic measurement frameworks (formal and computational representations of geographic information requires the adoption of an implied topological and geometric measurement framework), spatial dependency (tendency of attributes at nearby locations in space to be related), spatial heterogeneity (an intrinsic degree of uniqueness at all geographic locations that makes global comparisons difficult), and the complexity of spatio-temporal objects and rules (whereas objects in non-spatial datasets normally are points in information space, spatial objects often have size, shape, and boundary properties) (Miller and Han 2001). Important output patterns for spatial data mining are spatial outliers (Shekhar, Lu and Zhang 2003, Lu *et al.* 2003, Ng 2001), spatial clusters (O'Sullivan and Unwin 2003, Sadahiro 2003), movement patterns (Laube *et al.* 2005, Gudmundsson *et al.* 2007), spatial co-location patterns (Shekhar and Huang 2001) and spatial association rules (Koperski and Han 1995).

In conclusion, spatial data mining is in many respects more complex than conventional data mining due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation (Shekhar, Zhang, Huang and Vatsavai 2003).

2.2 Spatial, spatiotemporal, and fuzzy association rules

Association rule mining has been an active research area ever since the seminal work by Agrawal *et al.* (1993). Less work has been seen extending association rules into the spatial domain. In an early work, Koperski and Han (1995) investigated *spatial association rules* (sAR) among a set of spatial and possibly some non-spatial predicates. They present optimisation techniques for association rules with a spatial antecedent and a non-spatial consequent ($is\_a(x, house) \wedge close\_to(x, river) \rightarrow is\_expensive(x)$) and with a non-spatial antecedent and a spatial consequent ($is\_a(x, gasStation) \rightarrow close\_to(x, highway)$). More recently, Gidofalvi and Pedersen (2005) describe an approach transforming the spatio-temporal rule-mining task to the traditional market-basket analysis task for the improvement of a location-based service application. Both approaches rely on support and confidence measures that are based on counts of pattern frequencies.

Association rules have also been used for the mining of object mobility patterns. Verhein and Chawla (2006) and Verhein and Chawla (2008) present a definition of *spatio-temporal association rules* (STARs) that specifically describe how objects move between regions over time, motivated by a scenario of mobile phone users moving in a cell-phone network. Although rather specific in their orientation toward mobility patterns between sets of cells, their inclusion of the spatial semantics of the cell sets into their support measure is relevant for our work. As they define support for their STARs in terms of transition counts from one cell to another, and since these cells can be very different in size, they suggest to include the size of the cells into their spatial support measure. Rules expressing transitions between small and restrictive cells are stronger than rules describing transitions between large and inclusive cells (Verhein and Chawla 2006).

In traditional market-basket analysis, when considering binary association rules $Ant \rightarrow Cons$, each transaction either completely satisfies Ant or it does not satisfy Ant at all, and the same is the case for $Cons$. This can also be applied to *quantitative association rules*, where attribute *intervals* are used (Dubois *et al.* 2006). An example for such a quantitative association rule is "If an employee is between 35 and 45 years of age, then his/her income is more than $100,000". Sometimes, however, one wishes to be less precise, and work with rules like "If an employee is middle-aged, then he/she has a high income." To this end, Kuok *et al.* (1998) introduced *fuzzy association rules*,

where crisp intervals are replaced by fuzzy intervals. In other words, strict membership functions are replaced by fuzzy membership functions giving *scores* in the range $[0, 1]$. Hence, the concepts of support and confidence must be redefined. Now suppose that for a database item x, the score functions $s_{Ant}(x)$ and $s_{Cons}(x)$ determine to what extent x satisfies the antecedent Ant and consequent $Cons$, respectively. Then the support and confidence of the rule $Ant \rightarrow Cons$ can be expressed using a so-called t-norm (Dubois *et al.* 2006). This is a function $\otimes : [0, 1] \times [0, 1] \rightarrow [0, 1]$ that is commutative, associative, monotone, and has 1 as its identity element (Hájek 1998). The support of $Ant \rightarrow Cons$ is now given by

$$\text{support} = \sum_x s_{Ant}(x) \otimes s_{Cons}(x) \tag{1}$$

and the confidence is given by

$$\text{confidence} = \frac{\sum_x s_{Ant}(x) \otimes s_{Cons}(x)}{\sum_x s_{Ant}(x)}. \tag{2}$$

A t-norm that is used often is the minimum t-norm, which takes the minimum of its two arguments; another possibility is the product t-norm, which multiplies its arguments (Dubois *et al.* 2006) Note that when $s_{Ant}(x)$ and $s_{Cons}(x)$ are either 0 or 1, then $s_{Ant}(x) \otimes s_{Cons}(x) = 1$ when $s_{Ant}(x) = s_{Cons}(x) = 1$ and $s_{Ant}(x) \otimes s_{Cons}(x) = 0$ otherwise. (This holds for both the minimum t-norm and for the product t-norm.) Hence, the definitions for support and confidence given above reduce to the standard definitions for the non-fuzzy case.

2.3 Proximity, nearness, and fuzzy neighbourhoods

An sAR may include spatial predicates such as $close\_to$, $adjacent\_to$, or $inside$. Most sARM approaches use thresholds such as "within a distance of 80km" when modeling proximity (Koperski and Han 1995). Revolving around Tobler's first law of Geography, claiming that "close things are more likely to be related than distant things" (Tobler 1970), the field of Geographical Information Science developed a wide range more sophisticated models to express and measure proximity relations. See Miller and Wentz (2003) for an introductory text on geographic relationships. Worboys (1996) discusses distance and proximity relationships between entities in geographic spaces,

putting an emphasis on representations that go beyond metric spaces, allowing, for instance, asymmetry in distance relations (e.g. travel time, uphill vs. downhill). In a later piece of work, Worboys (2001) explores the vagueness of the spatial relation "near". He specifically suggests the extension of sets with broad boundaries as nearness neighbourhoods to fuzzy neighbourhoods with continuous measures of nearness between 0 and 1 . Worboys's work focusses on how people perceive and reason with vague concepts such as nearness, so it is different from our approach in Section 4 where we try to quantify proximity automatically.

Most work on the vague spatial relations *proximity* and *nearness* agree on the context dependency of such concepts. However, if context knowledge can be provided by expert users, concepts such as fuzzy neighbourhoods and the like can be powerful tools for modeling proximity and nearness in a sARM context.

3 Quality measures for spatial association rules

In many non-spatial applications the numerical values of the attributes are already given and it is rather straightforward to map these values to scores in the range $[0, 1]$. In the spatial domain, however, this is not the case. In this paper we discuss the application of concepts from fuzzy-association-rule theory to spatial association rules, in particular to rules involving proximity.

3.1 Scoring proximity

In section 2.2 we have seen that in order to apply the theory of fuzzy ARM, we need to define score functions in the range [0,1] for the antecedent and for the consequent of the association rule. The task of defining score functions for proximity can be split into two subtasks. Consider as an example the condition "close to the river". Then the first task is to determine how to quantify distance to the river, and the second task is to convert this distance to a score in the range $[0, 1]$. Suppose for the moment that we have a suitable distance function *dist*. Traditionally, a user would define a threshold parameter d and one would say that an object o is close to a river R if $dist(o, R) \leq d$. Instead, we use two parameters, d_c and d_f, with $d_c < d_f$, and define

$$score(o) = \begin{cases} 1 & \text{if } dist(o, R) \leq d_c \\ \frac{d_f - dist(o,R)}{d_f - d_c} & \text{if } d_c < dist(o, R) \leq d_f \\ 0 & \text{if } dist(o, R) > d_f \end{cases} \quad (3)$$

Typically, the value of d_c would be somewhat smaller than the strict threshold d would be, while d_f would be somewhat larger. Figure 2 illustrates this definition.

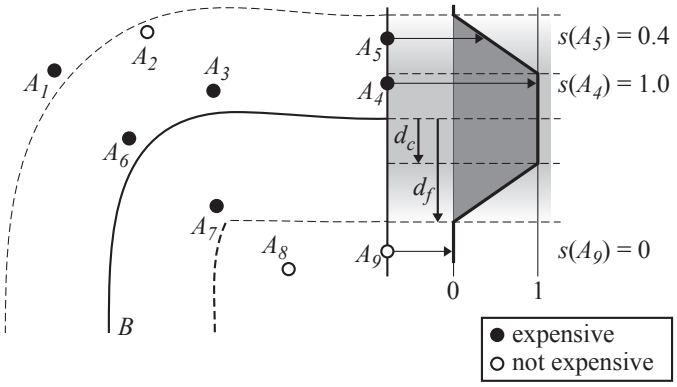

Figure 2: A linear score function allows the computation of spatial support for the antecedent "A is close to B". The score function rises from 0 at distance d_f to 1 at distance d_c. Example point A_9 provides support 0, A_5 accounts for support 0.4, and A_4 for 1.0. Note that both A_4 and A_3 provide the same support 1.0 even if A_3 is closer to B than A_4.

The best values for the parameters d_c and d_f depend on the application, and can be determined by the user. Note however, that even though we now require two parameters rather than one, the score function is much less sensitive to the exact value of the parameters than was the case for simple thresholding (see Figure 1). Indeed, the score function is continuous in the parameters d_c and d_f, so changing their values slightly will result in only a slight change in the score function.

Above we have used a piecewise linear score function. When so desired, one could also use a smooth Gaussian-like score function. Furthermore, the score

function can easily be adapted to capture other proximity concepts such as "far". All that we need is a suitable distance function *dist*. In the Section 4 we investigate how one can define distance in various different settings.

3.2 Spatial support and spatial confidence

First we discuss how to define support and confidence in the simplest case, namely where both the antecedent and the consequent consist of a single predicate. Recall from Section 2.2 that in order to define support and confidence, we need to choose a suitable t-norm. Although the minimum t-norm is often used, we believe that in spatial association rules the product t-norm is more appropriate. As an example, consider the rule "If a house is close to the sea, then it is expensive." Suppose we have a score $s_{Ant}(H)$ in the range $[0, 1]$ that captures to what extent a house H is close to the sea, how to obtain such a score is discussed in the next section. We also map the price of H to a score $s_{Cons}(H)$ in the range $[0, 1]$ to capture to what extent a house is expensive. For example, houses that cost \$1,000,000 or more get a score of 1, houses that cost less than \$500,000 get a score of 0, and in between we interpolate the score. Using the product t-norm in the definition of support then gives us the spatial support

$$\text{spatial support} = \sum_H s_{Ant}(H) \times s_{Cons}(H), \tag{4}$$

where the sum is over all houses H. Similarly, spatial confidence is defined as

$$\text{spatial confidence} = \frac{\sum_H s_{Ant}(H) \times s_{Cons}(H)}{\sum_H s_{Ant}(H)}. \tag{5}$$

We mentioned that in a spatial context the product t-norm is usually the most appropriate. Indeed, in the example above, houses that are rather close to the sea (say, score 0.7) and very expensive (score 0.9) should give more support than houses that are rather close (score 0.7) to the sea and somewhat expensive (score 0.7), as is given with the product t-norm ($0.7 * 0.9 = 0.63 > 0.7 * 0.7 = 0.49$).

Note that the antecedent and/or the consequent of a sAR can be composed of several predicates. In the next subsections we discuss two examples of this.

3.2.1 A score for two antecedents combined by AND

Consider the sAR "If a house is close to the sea and close to a big city, then it costs at least $800,000.", and suppose we have a score for "close to the sea" and a score for "close to a big city". From these two scores we must then compute an overall score for the antecedent. For this we need another t-norm. We believe that also here the product t-norm is the most appropriate. Indeed, if an expensive house is very close to the sea (score 1) and somewhat close to a big city (say, score 0.5) then it should support the rule somewhat, while if an expensive house is somewhat close to the sea (score 0.5) and somewhat close to a big city (score 0.5), then it should support the rule less. We get this behavior by multiplying the scores: in the first case we then have a score for the antecedent of 0.5, and in the second case we have a score of 0.25. Note that in standard fuzzy logic, the AND-operator gives the minimum (instead of the product) of the two scores. The semantics of the product appears more suitable in our case than the semantics of the minimum, as we argued briefly.

3.2.2 A score for two antecedents combined by OR

Consider the sAR: "If a house is close to an airport or close to a highway, then it has good sound insulation." If a well-insulated house is already very close to a highway, then proximity to an airport is irrelevant for the support of the rule, which simply should be 1, no matter how close or far any airport is. But if the house is somewhat close to a highway and somewhat close to an airport (both with score 0.5), then it supports the rule more than when only one of these antecedents was present. A t-conorm—this is similar to a t-norm, except that it has 0 as identity element—that nicely captures this behavior is the Einstein sum, defined as $\frac{s_1+s_2}{1+s_1 s_2}$ for two scores s_1 and s_2.

4 Proximity measures

In this section we discuss how to quantify proximity or, in other words, how to define distance measures for various types of geographic objects. Since spatial association rules have most use in two-dimensional space, we will limit ourselves to this case; the ideas, however, easily extend to three-dimensional objects.

In the object view, one generally distinguishes three types of geographic

objects in two-dimensional space: zero-dimensional objects (points), one-dimensional objects (linear objects, typically polylines), and two-dimensional objects (areas, typically polygons). When measuring the distance between such objects A and B, we thus have a number of different cases: point-to-point, point-to-polyline, and so on. (Note that, depending on the application, the point-to-polyline case need not be the same as the polyline-to-point case.) The next three subsections discuss the point-to-point, point-to-polyline, and point-to-polygon cases, respectively; the last subsection then comments on some other cases.

4.1 The point-to-point case

We begin with the case of point-to-point proximity. As an example, consider the rule "If a street-crime incident is close to an ATM, then it is a pickpocket case." Here both the location of the incident and the location of the ATM are points. Note, however, that the rule speaks of *an* ATM. Thus we are not measuring the distance to a specific ATM, but to any ATM. In more abstract terms, the point-to-point situation is often as follows. We are given a point A and a set S_B of points, and we are interested in $\mathit{dist}(A, S_B)$, the distance from A to the set S_B. In our example, A would be a street-crime incident and S_B would consist of all ATM locations. In this example, a natural interpretation of the rule is to consider the distance from the incident to the nearest ATM. This corresponds to setting

$$\mathit{dist}(A, S_B) = \min_{B \in S_B} d(A, B),$$

where $d(A, B)$ denotes the distance between the points A and B. The distance $d(A, B)$ can be the Euclidean distance, the travel time on the road, or any other distance measure between individual points, but in any case $\mathit{dist}(A, S_B)$ would be defined by the nearest point $B \in S_B$—see Figure 3.

If S_A is a set of n points, S_B is a set of m points, and $d(A, B)$ denotes the Euclidean distance, then we can compute $\mathit{dist}(A, S_B)$ for each $A \in S_A$ in $O(m \log m + n \log m)$ time in total. To this end we compute the Voronoi diagram of S_B, preprocess it for efficient planar point location, and then perform a query with each point $A \in S_A$ to find its closest point in S_B. The first two steps take $O(m \log m)$ time, and each of the point-location queries takes $O(\log m)$ time (de Berg *et al.* 2008). The network version—here the points from S_A and S_B lie on a network with E edges, and distances are

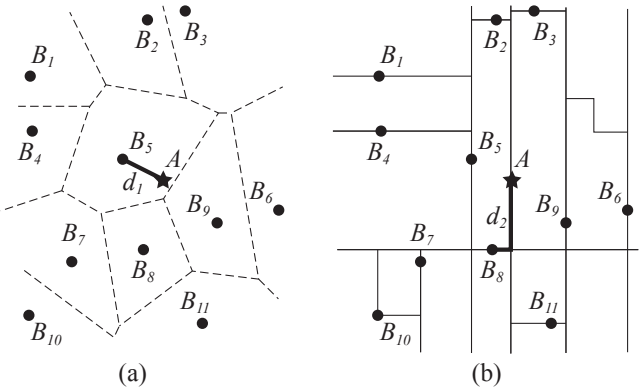

Figure 3: Score function for the point-to-point case. (a) The Voronoi diagram is used to define the nearest ATM location B_5 to a crime spot A. (b) If travel time is assessed on a street network a different ATM is closest, B_8. The Euclidean distance d_1 and the Manhattan distance d_2 are used when computing distance $d(A, B)$.

measured along the network—can be solved in $O((n + m) \log(n + m) + E)$ (Erwig 2000).

Sometimes not only the closest point of S_B matters. For example, consider the rule: "If a hotel is close to castles, then it is used mostly by tourists." In this case, not just the nearest castle matters, but also other nearby castles and their proximity: the more castles in the vicinity, the more the antecedent of the rule applies. One possibility to handle this is as follows. First, for two points A and B—a hotel and a castle—we define $w(A, B)$ to be a value in the range $[0, 1]$ that expresses to closeness of A and B. We do this in such a way that $w(A, B) = 0$ if A and B are really close to each other, $w(A, B) = 1$ if A and B are really far from each other, and $w(A, B)$ increases linearly in between. (Note the similarity to the way we mapped distance to score in the previous section.) Then we define

$$\mathit{dist}(A, S_B) = \sum_{B \in S_B} w(A, B).$$

Observe that $\mathit{dist}(A, S_B)$ can range from 0 (if all castles in S_B are really close to A) to $|S_B|$ (if all castles are really far from A); this distance[1] will be

[1]Note that adding another point to S_B that is far away from A will increase $\mathit{dist}(A, S_B)$. If

mapped to a score in the range $[0, 1]$, as usual. Computing $\mathit{dist}(A, S_B)$ for each $A \in S_A$ now takes $O(nm)$ time in total, since the distance to all other points in S_B can play a role. However, one would expect that only few points B are relevant—that is, have $w(A, B) > 0$—for a given A. Hence, one can speed up the computation in practice by using spatial index structures such as R-trees to quickly find for each $A \in S_A$ the relevant points $B \in S_B$.

4.2 The point-to-polyline case

Next we discuss how to measure the distance between a point and a single polyline. As in the point-to-point case, we may want to extend this definition to the case of multiple polylines. This can be done by considering the closest polyline (as in the street-crime/ATM example) or in some more involved manner (as in the hotel/castles example). For the point-to-polyline case, however, these considerations already play a role when considering the distance to a single polyline. Consider for example the rule "If a house is close to the river, then the occupants own a boat." Here what matters could be how quickly the occupants can reach the river. For a point A and polyline B—the house and the river—we would then consider the minimum distance from A to any point on B:

$$\mathit{dist}(A, B) = \min_{p \in B} d(A, p),$$

where $d(A, p)$ denotes (for instance) Euclidean distance. But now consider a house and its proximity to a highway, for the purpose of studying problems due to noise: "If a house is close to the highway, ...". Clearly, a house that is within 500m from the highway over a stretch of 1.2km of that highway, suffers more noise pollution than a house that is within 500m over a stretch of 0.7km—see Figure 4. Hence, minimum distance does not seem an appropriate distance measure in this example. One possible solution is to proceed similarly to the hotel/castles example. Thus we first define for points $p \in B$ a function $w(A, p)$ that is 0 (resp. 1) if A and p are really close to (resp. far from) each other, and that increases linearly in between, and we define

this is undesirable, one may reverse the definition of $w(A, B)$ so that $w(A, B) = 1$ for points that are close to (instead of far from) A. This also means one should change the mapping from distance to score, as larger "distance" now implies more castles that are closer, which should lead to a higher score.

$$\mathit{dist}(A, B) = \int_0^1 w(A, p(x))dx$$

Now $\mathit{dist}(A, B)$ can vary between 0 (if the entire highway is very close to A) to $\mathit{length}(B)$ if the entire highway is far from A.

Let S_A be a set of n points whose proximity to a polyline B consisting of m line segments is needed. For a score based on the minimum distance only, similar to the point-to-point case, we use Voronoi diagrams of line segments to compute $\mathit{dist}(A, B)$ for all $A \in S_A$ in $O(m \log m + n \log m)$ time in total. If we use the definition using the integral, then we cannot use Voronoi diagrams to reduce the running time and we will need $O(nm)$ time in the worst case. As before, index structures like R-trees can be used to speed up the efficiency in practice.

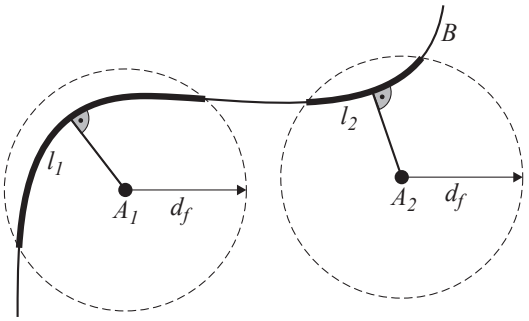

Figure 4: The point-to-polyline case for "if a house is close to the highway". Even though in terms of absolute distance A_1 and A_2 are equally close to B, when assessing the proximity with respect to the stretch l_x covered by a disc of radius d_f, then A_1 is closer to B than A_2.

4.3 The point-to-polygon case

Also in the point-to-polygon proximity case, there are several ways to quantify proximity. Depending on the application, we can use the shortest distance from the point to the polygon, the total length of the boundary of the polygon within a certain distance from the point, or the total area of the polygon

within a certain distance from the point—see Figure 5. The shortest distance could be appropriate when scoring access to a water in an agricultural property evaluation. Here the only fact that matters might be the shortest distance, as this is directly linked to development costs. When evaluating potential sites for a new beach-side hotel, the actual length of the beach stretch within some walking distance might be more important than the area of the water body itself. Finally, areas could be important in a bird ecology study evaluating potential nesting sites. In Figure 5, A_1 and A_2 could be nesting sites for birds that need access to the forest B for food supply. Even though A_1 is in absolute distance further away from forest B, it has much more forest within distance d_f than A_2 and, hence, may be the better nesting site.

Note that the definition of the distance function $dist(A, B)$ for a point A and polygon B in the latter case (the nesting example) is very similar to the definition in the highway example in the previous subsection. Hence, we can define the distance in a similar way, using a (double) integral over the polygon B.

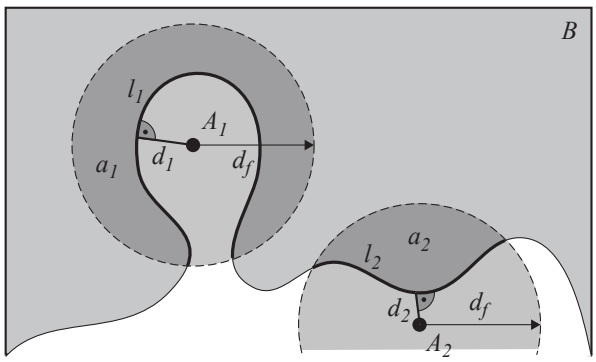

Figure 5: The point-to-polygon case. The proximity of point A with respect to area B is assessed by several measures derived from a circular template around A. In the most simple case minimal distances d_1 and d_2 are used. Alternatively, the length of the boundary cut out by the circular template might define the proximity, with l_1 outperforming l_2. Finally, if the covered area matters, proximity depends on the areas a_1 and a_2.

4.4 Other cases

There are several more cases, namely those where A is of a linear or areal type. Examples are: "If a forest is close to highways, then it contains only few deer", and: "If an urban development area is close to existing urban areas, then the houses will have small gardens".

First consider the case where we want to quantify the distance from a polygon to a set of points: "If a lake is close to dump sites, then its water is polluted." For this rule, we want the polygon-to-point proximity to take all nearby dump sites and their distances into account. Thus we should define distance similar to some examples given before. Note that this way we already defined a score in the range [0,1] to capture proximity, so the machinery of section 3.1 to convert distance to score is not necessary. The same is true for the examples discussed next.

Now let's consider polygon-to-polygon proximity. Observe first that the polygon types can be such that overlap is not possible ("If a lake is close to forests, ..") or such that overlap is possible ("If a forest is close to municipalities with many elderly people, ..."). These cases give rise to different score functions. If overlap is not possible, then the simplest definition of the score is based on the minimum distance and is like the point-to-point case, see Figure 6.

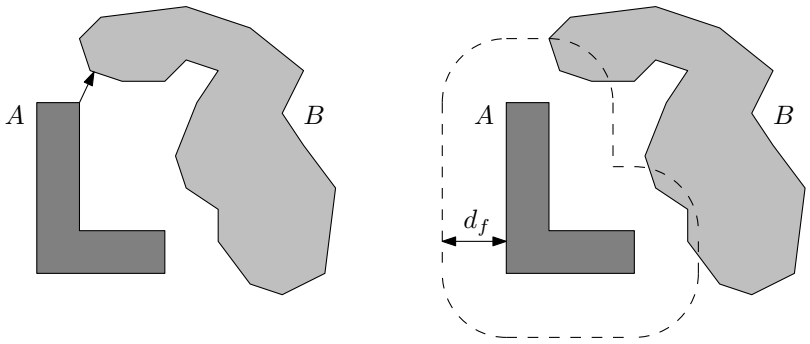

Figure 6: Two simple ways to define polygon-polygon proximity.

Another simple definition is based on the area of B inside a buffer of width d_f around A, see Figure 6. For example, if R denotes the buffer region, then one could say that if more than, say, 30% of R is covered by B, then the score

is 1. If 0% of R is covered, then the score is 0. In between these percentages, we can let the score depend linearly on the percentage covered, similar to some examples given before.

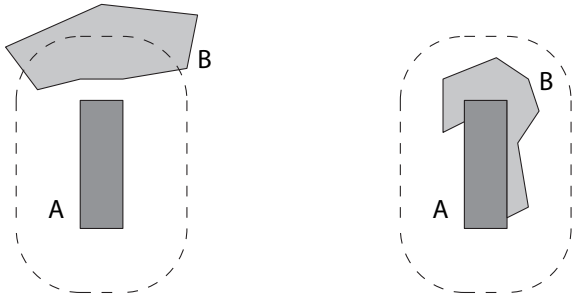

Figure 7: Discriminating two cases where 15% of the area of the buffer of A is covered by B.

Note that such a definition may sometimes be too limited. It does not discriminate on the distance from A, and also not on which parts of A are actually close to B. Consider Figure 7. In both cases, about 15% of the area of the buffer of A is covered by B, but in the right case, A appears closer to B, because for every point in A, the polygon B is closer than in the left case.

We can refine our definition by using a weighted buffer and taking integrals, similar to what we did before. For example, define the weight $w(p)$ of a point p to be 0 if the distance from p to A is at least d_f, and let its weight be $(d_f - \mathit{dist}(p, A))/d_f$ otherwise. We can now define the score depending on

$$\iint_B w(p)\ dxdy.$$

We can divide this value by the integral of the weight in the whole buffer region of A, and if this is more than, say, 0.3, let the score be 1. The score can decrease linearly in the outcome of the division.

5 Spatio-temporal support and confidence

This section investigates the extension of our fuzzy spatial association rules into the temporal domain. We start our consideration of spatio-temporal association rules by extending the basic point-to-point case with a temporal

antecedent T: "If A is close to B, around time T, then $Cons$". This rule has two antecedents, namely one capturing spatial proximity, "A is close to B", and one assessing temporal proximity, "A is present around time T". An example could be "If mobile phone users are close to the city centre at noon, then they are businessmen". Note that combining scores for the two antecedents with the product t-norm appears appropriate.

If we perceive the time as a one-dimensional space, we can model temporal proximity similar to spatial proximity. Time stamps of events and episodes may be located on a time axis, allowing the assessment of various temporal relationships. Just as with spatial proximity, also temporal relations may be extended beyond discrete classes. Predicates such as "at noon", "in the morning", or "at night" may be assessed using score functions allowing for a refined concept of temporal proximity. Unlike spatial relations, temporal relations are in general directed. A relation such as "after sunset" is directed and this direction has to be included when modeling temporal proximity.

Figure 8 illustrates examples for a score function for temporal proximity for a linear and a cyclical time axis. The first example refers to the "baby boomer" generation. In order to qualify as a baby boomer, a person has to be born clearly after e_1, the end of World War II. Whereas births up until 1965 generally qualify for the baby-boomer generation (score 1), there might be a transition period that can be modeled with a directed score function. By contrast, when referring to the decade "the Seventies" one might want to also include funny hair cuts and bell-bottoms found in 1969 and 1981. Hence, the use of an extended bi-directed temporal proximity score function that leaps beyond the crisp decade e_2 might be appropriate.

Similar concepts can be applied for temporal proximity on a cyclic time scale. Whereas the event e_3 "sunset" is crisply defined, there may be leeway for the temporal predicate "after sunset". Finally, for many spatio-temporal association rule mining applications refering to the episode e_4 "in the morning", an extended temporal neighborhood function may be more adequate than a crisp interval ranging from 6am to 12pm.

6 Discussion

Our exploratoy study on quality measures for spatial association rules illustrates that it is imperative to include spatial semantics when adopting as-

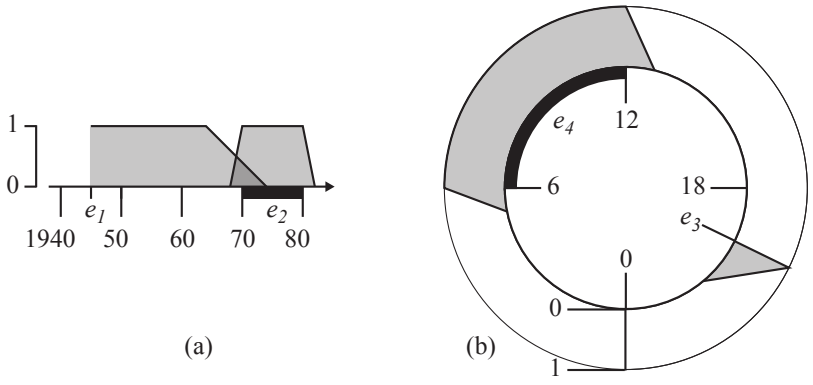

Figure 8: Temporal proximity can be modeled with uni-directional (events e_1 and e_3) or bi-directional (episodes e_2 and e_4) score functions on both linear (a) and cyclic (b) time axes.

sociation rules for assessing the complex relationships between spatial objects. This finding complies with similar conclusions from work on spatio-temporal ARs in a mobility context (Verhein and Chawla 2006, Verhein and Chawla 2008). As we have argued and illustrated when defining spatial support and spatial confidence, the spatial context of specific sAR applications makes the selection of some t-norms for score combinations more suitable over others. Similarly, the semantics of the variability of interrelations amongst spatial objects leads to a wide range of sensible proximity measures.

Allowing score functions for both antecedents and consequents in various conjunctions, requires the specification of rules on how to combine separate scores. The presented t-norms and t-conorms offer a suitable framework for such guidelines.

If one wants to use fuzzy association rules in a spatial context, one has to quantify proximity. We have shown how this can be done for a number of different cases, involving proximity between features of various types (point, linear, areal). Often spatial association-rule mining is only a first step in analyzing spatial data. This step gives a number of association rules, which can then be subjected to further investigation using e.g. statistical methods. Note that these statistical methods would also need to quantify proximity. Hence, our work on proximity measures not only has applications in spatial

association-rule mining, but it can also be useful in statistical methods for spatial data analysis (or, in fact, any type of quantitative analysis involving spatial data).

Being one-dimensional in nature, the temporal dimension *per se* adds little complexity to our proximity discussion. However, taken in combination with almost arbitrarily complex score functions for spatial proximity, the suggested temporal proximity score functions sum up to a powerful tool when assessing spatio-temporal association rules.

7 Conclusions

This paper contributes to the theory of spatial data mining by refining quality measures for spatial association rules. We present a conceptual framework for *spatial support* and *spatial confidence* applying concepts from the theory of fuzzy association-rule mining. The major contribution of this paper is twofold: First, we introduce fuzzy association rule quality measures into the spatial domain. Second, we explore various possibilities for defining and efficiently computing suitable proximity measures amongst objects in space-time. With a series of illustrative examples we have shown that developing spatial and spatio-temporal quality measures for association rules presents a set of interesting proximity problems, ranging from simple point-to-point constellations to rather complex polygon-to-polygon scenarios.

Acknowledgments

This research was initiated during the GADGET Workshop on Geometric Algorithms and Spatial Data Mining, funded by the Netherlands Organisation for Scientific Research (NWO) under BRICKS/FOCUS grant number 642.065.503. Patrick Laube's work is funded by the Australian Research Council (ARC), Discovery grant DP0662906. Mark de Berg was supported by the Netherlands' Organisation for Scientific Research (NWO) under project no. 639.023.301. Finally, the authors wish to thank three anonymous reviewers for their valuable comments.

References

Agrawal, R., Imieliski, T. and Swami, A. (1993). Mining association rules between sets of items in large databases.

de Berg, M., Cheong, O., van Kreveld, M. and Overmars, M. (2008). *Computational Geometry: Algorithms and Applications*, 3nd edn, Springer-Verlag, Berlin.

Dubois, D., Hüllermeier, E. and Prade, H. (2006). A systematic approach to the assessment of fuzzy association rules, *Data Min. Knowl. Discov.* **13**(2): 167–192.

Erwig, M. (2000). The graph Voronoi diagram with applications, *Networks* **36**(3): 156–163.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *AI Magazine* **17**(3): 37–54.

Gidofalvi, G. and Pedersen, T. B. (2005). Spatio-temporal rule mining: Issues and techniques, *Data Warehousing and Knowledge Discovery, Proceedings*, Vol. 3589 of *Lecture Notes in Computer Science*, Springer-Verlag, Berlin, pp. 275–284.

Gudmundsson, J., van Kreveld, M. and Speckmann, B. (2007). Efficient detection of patterns in 2D trajectories of moving points, *GeoInformatica* **11**(2): 195–215.

Hájek, P. (1998). *Metamathematics of Fuzzy Logic*, Kluwer.

Koperski, K. and Han, J. (1995). *Discovery of Spatial Association Rules in Geographic Information Databases*, Proceedings of the 4th International Symposium on Advances in Spatial Databases, Springer-Verlag.

Kuok, C., Fu, A.-C. and Wong, M. (1998). Mining fuzzy association rules in databases, *SIGMOD Record* **27**: 41–46.

Laube, P., Imfeld, S. and Weibel, R. (2005). Discovering relative motion patterns in groups of moving point objects, *International Journal of Geographical Information Science* **19**(6): 639–668.

Lu, C.-T., Chen, D. and Kou, Y. (2003). Detecting spatial outliers with multiple attributes, *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence 2003 (ICTAI'04)*, pp. 122–128.

Miller, H. J. and Han, J. (2001). Geographic data mining and knowledge discovery: An overview, *in* H. J. Miller and J. Han (eds), *Geographic data mining and knowledge discovery*, Taylor and Francis, London, UK, pp. 3–32.

Miller, H. J. and Wentz, E. A. (2003). Representation and spatial analysis in geographic information systems, *Annals of the Association of American Geographers* **93**(3): 574–594.

Ng, R. T. (2001). Detecting outliers from large datasets, *in* H. J. Miller and J. Han (eds), *Geographic data mining and knowledge discovery*, Taylor and Francis, London, UK, pp. 218–235.

O'Sullivan, D. and Unwin, D. J. (2003). *Geographic Information Analysis*, John Wiley and Sons, Hoboken, NJ.

Roddick, J. F., Hornsby, K. and Spiliopoulou, M. (2001). An updated bibliography of temporal, spatial, and spatio-temporal data mining research, *in* J. F. Roddick and K. Hornsby (eds), *Temporal, spatial and spatio-temporal data mining, TSDM 2000*, Vol. 2007 of *Lecture Notes in Artificial Intelligence*, Springer, Berlin Heidelberg, DE, pp. 147–163.

Roddick, J. F. and Lees, B. G. (2001). Paradigms for spatial and spatio-temporal data mining, *in* H. J. Miller and J. Han (eds), *Geographic data mining and knowledge discovery*, Taylor and Francis, London, UK, pp. 33–49.

Sadahiro, Y. (2003). Cluster detection in uncertain point distributions: a comparison of four methods, *Computers, Environment and Urban Systems* **27**(1): 33–52.

Shekhar, S. and Huang, Y. (2001). Discovering spatial co-location patterns: A summary of results, *Advances in Spatial and Temporal Databases, Proceedings*, Vol. 2121 of *Lecture Notes in Computer Science*, Springer-Verlag, Berlin, pp. 236–256.

Shekhar, S., Lu, C. T. and Zhang, P. S. (2003). A unified approach to detecting spatial outliers, *Geoinformatica* **7**(2): 139–166.

Shekhar, S., Zhang, P., Huang, Y. and Vatsavai, R. R. (2003). Trends in spatial data mining, *in* H. Kargupta, A. Joshi, K. Sivakumar and Y. Yesha (eds), *Data Mining: Next Generation Challenges and Future Directions*, MIT/AAAI Press.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region, *Economic Geography* **46**(2): 234–240.

Verhein, F. and Chawla, S. (2006). Mining spatio-temporal association rules, sources, sinks, stationary regions and thoroughfares in object mobility databases, *Database Systems for Advanced Applications*, pp. 187–201.

Verhein, F. and Chawla, S. (2008). Mining spatio-temporal patterns in object mobility databases, *Data Mining and Knowledge Discovery* **16**(1): 5–38.

Worboys, M. F. (1996). Metrics and topologies for geographic space, *in* M. J. Kraak and M. Molenaar (eds), *Advances in Geographic Information Systems Research II: Proceedings of the International Symposium on Spatial Data Handling, Delft*, Taylor & Francis, London, UK, pp. 365–376.

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