# Mining Spatial Data via Clustering

Vladimir Estivill-Castro
Intelligent Cooperative Information Systems Research Centre
Queensland University of Technology,
GPO Box 2434, Brisbane 4001, Australia.
vlad@icis.qut.edu.au

Alan T. Murray
Australian Housing and Urban Research Institute
Queensland University of Technology,
GPO Box 2434, Brisbane 4001, Australia.
a.murray@qut.edu.au

**Abstract**

*Contributions from researchers in Knowledge Discovery are producing essential tools in order to better understand the typically large amounts of spatial data in Geographical Information Systems. Clustering techniques are proving to be valuable in providing exploratory analysis functionality while supporting approaches for automated pattern discovery in spatially referenced data and for the identification of important spatial relationships. However, there is little recognition of the broader context for which many clustering approaches are related. The concern for efficient filtering of outliers has divested attention from the actual problem being solved, which has resulted in a lack of recognition for the variety of approaches that could be modeled and in the re-discovery of solution approaches that are problematic or inferior. We present an overview of non-hierarchical clustering approaches for spatial data analysis. Our panoramic view to clustering large spatio-referenced datasets suggests that more effective and efficient clustering methods are possible.*
KEYWORDS: *Clustering, Knowledge Discovery, Optimization, Local Search.*

## 1 Introduction

The emergence of Geographical Information Systems (GIS) with convenient and affordable methods for storing large amounts of spatial data has accelerated the rate at which sheer quantity of spatially referenced information is collected in electronic and magnetic media. It is argued that the GIS revolution and the increasing availability of GIS databases emphasizes the need for exploratory (discovery) rather than confirmatory methods of analysis [18]. Current research seeks techniques for artificial searchers that are able to hunt out localized patterns or database anomalies in geographically referenced data by reducing the need for direction ("where" to look or "what" to look for, or "when" to look). Spatial data mining [17] is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. The automatic knowledge discovery process in spatial databases aims at a) extracting interesting spatial patterns and features, b) capturing intrinsic relationships between spatial and non-spatial data, c) presenting data regularity concisely and at higher conceptual levels, and d) helping to reorganize spatial databases to accommodate data semantics and to achieve better performance.

Clustering is the task of identifying groups in a data set by some natural criteria of similarity [6].

For geo-referenced space the most obvious measure of similarity is Euclidean distance, although other derived geo-referenced distances are possible. Thus, similarity measurement between observations is relatively well-defined for geo-referenced data. This allows clustering (or cluster analysis) to be applied for knowledge extraction [2, 7, 8, 17, 24]. Although clustering reveals proximity associations, the partitioning of observations into clusters remains a delicate issue.

The idea behind the use of clustering of spatially referenced data is that it provides a means of generalization of the spatial component of the data associated with a GIS. This is complementary to the techniques for generalization used in data mining in relational databases [3]. Clustering hopefully identifies clouds and helps to filter outliers.

Given the importance of space, and thus of distance, we illustrate the problematic and concerning characteristics of the spatial data mining approaches recently developed in the Knowledge Discovery literature [17, 24]. Others [7] have criticized these approaches not on the fundamental issue of the problem formulation, but on their computational efficiency, in particular, on their space requirements. It will be shown here that this argument is misleading, because there are strong reasons for using medoid methods (and in fact, they can be implemented in linear space). Moreover, alternative density based approaches [7] require quadratic space for sorting all pair of distances that determine its density parameter.

A large family of clustering methods has emerged in the literature of statistics, machine learning, knowledge discovery and operations research, offering a spectrum of quality versus computational cost. Why has the field of Knowledge Discovery ignored some of the approaches suggested in other fields? Is the cluster quality dependent on solution methods or problem formulation? What are the trade-offs in cluster quality? We believe that the re-discovery of problematic clustering techniques and the apparent disagreement on their benefits and drawbacks for spatial data may be due to the lack of a panoramic view on non-hierarchical clustering methods. This paper serves as a basis for re-evaluation to take place so that tools for exploratory spatial data analysis in GIS are both efficient with respect to the usually sheer size of the data set, and capable of identifying meaningful clusters.

## 2 Formulations for clustering

An important component of clustering is the definition of the criterion function that measures the quality of a data partitioning. The problem is then finding the partition that optimizes the criterion function. The rationale for clustering is to identify and group observations in order to minimize within group difference [10, 19, 21]. This explicit interaction criterion (using two clusters) pictorially looks as illustrated in Figure 1.

Consider a set $S = \{s_1, s_2, \ldots, s_n\}$ of data items where each $s_i$ is a point in $d$-dimensional real space $\Re^d$. The grouping of these $n$ points into $k$ clusters can then be formulated as an optimization problem as follows. Let $d$ be a metric in $\Re^d$ and let $\mathcal{P}$ be the set of partitions of $S$ into $k$ groups (then, $P \in \mathcal{P}$ means $P = P_1|P_2|\ldots|P_k$ with $P_i \neq \phi$, for $i = 1, \ldots, k$, and $P_i \cap P_j = \phi$, for $i \neq j$, but $\bigcup_{i=1}^{k} P_i = S$). Minimization of the *Observation Interaction* ($\mathsf{OI}$) criterion is defined by the following equation:

$$\mathsf{OI}(P) = \min_{P \in \mathcal{P}} \sum_{i=1}^{k} \sum_{s_u, s_v \in P_i} w_u w_v d(s_u, s_v). \tag{1}$$

The weight $w_u$ may reflect attributes of the observation $s_u$. The distance $d(s_u, s_v)$ may be one of

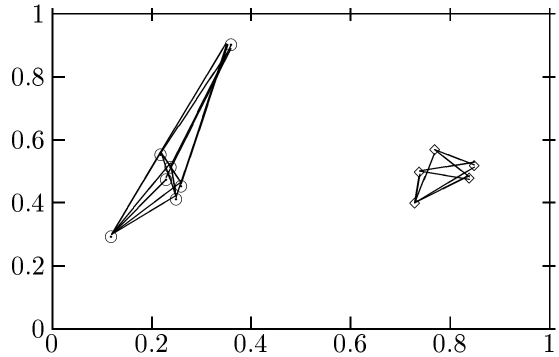

Figure 1: Group criterion.

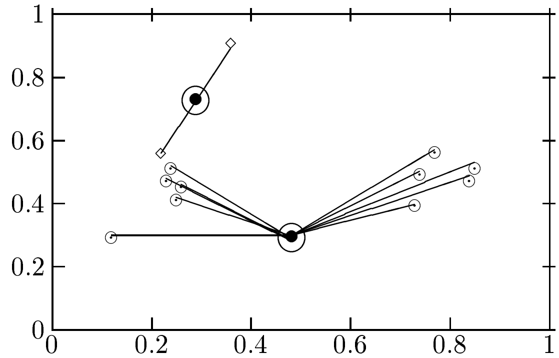

Figure 2: Center clusters.

Minkowski distances $d_m$ (i.e. $d_m(\vec{x},\vec{y}) = (\sum_{t=1}^{d} |x_t - y_t|^m)^{1/m}$ ). The case $m = 2$ corresponds to the Euclidean distance while the case $m = 1$ corresponds to the Manhattan distance. $\mathsf{OI}$ can be formulated as a non-linear optimization problem with $nk$ binary decision variables and $n + k$ linear constraints [16, 21] (a binary decision variable is either 1 or 0, while a linear constraint is a linear form of the decision variables).

A common approach to clustering is to identify a representative center for each cluster and to assess the quality of the clustering as the average distance between items and their representative center. Thus, the center points approach consists of finding $k$ centers $c_j$ (with $c_j \in \Re^d$, for $j = 1, \ldots, k$) such that the average distance to the closest center is minimized. More formally, we let $rep[s_i, C]$ be the closest point in $C = \{c_1, \ldots, c_k\}$ to $s_i$; that is, $\min_{j \in \{1,\ldots,k\}} d(s_i, c_j) = d(s_i, rep[s_i, C])$. The partition is defined by assigning each $s_i$ to its representative $rep[s_i, C]$. Those data items assigned to the same representative are in the same cluster; thus, the $k$ centers encode the partition of the data into $k$ groups. For $k = 2$, Figure 2 illustrates this criterion. The clustering problem translates to a combinatorial optimization problem where the goal is to find a set $C$ of representatives that optimizes the *Center Point* ($\mathsf{CP}$) criterion. Namely, we seek to minimize

$$\mathsf{CP}(C) = \frac{1}{n} \sum_{i=1}^{n} w_i d(s_i, rep[s_i, C]). \quad (2)$$

Equivalently, we seek to minimize $\sum_{i=1}^{n} w_i d(s_i, rep[s_i, C])$. This problem can be formulated as an

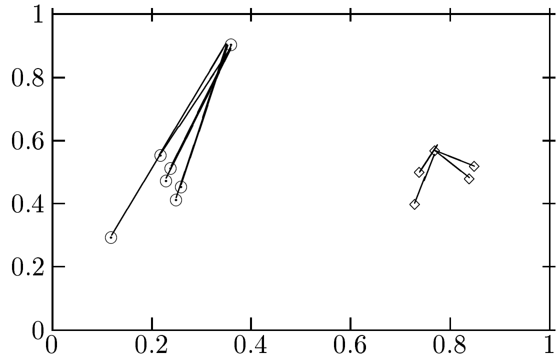


Figure 3: Medoid clusters.

optimization problem with $(2+n)k$ binary decision variables and $n$ linear constraints. However, the objective function is non-linear ($rep[s_i, C]$ is a function of $s_i$ and $C$).

The medoid approach is similar to the center approach, except that the set $C$ of representative centers is restricted to be a subset of $S$. This criterion, for $k=2$, is represented in Figure 3. The medoid approach has the advantage that it is robust with respect to outliers. The *Medoid Point* (MP) criterion is the minimization of

$$\mathsf{MP}(C) = \sum_{i=1}^{n} w_i d(s_i, rep[s_i, C]) \tag{3}$$

over all subsets $C$ of $S$ with $\|C\| = k$. MP is a linear-integer programming problem. Another relevant aspect of MP (and also of OI) is that, if desired, $d(s_u, s_v)$ can be computed in advance at the expense of $O(n^2)$ space. However, the drawback is that MP has $n^2+n$ binary decision variables and $n^2+n+1$ linear constraints.

The clustering formulations OI, CP and MP have variants where $[d(s_u, s_v)]^2$ [22, 24] is used rather than using the metric $d(s_u, s_v)$ in Equations (1), (2) and (3). The CP problem with $[d(s_u, s_v)]^2$ has popularity because of its simple interpretation. For a given cluster, the center (or centroid) minimizes the total squared error (a differentiable function). Thus, CP has a closed form solution in special cases. Simple iterative methods like $k$-MEANS [13] offer a computationally efficient approach for heuristically solving CP based upon theoretical foundations [6]. Note that clustering in spatial data for knowledge discovery is rarely the search for a representative, and the use of continuous space may place the representative where it does not make much sense (for example, a school on a lake). Moreover, CP is known to work well when the clusters form essentially compact clouds that are rather well separated from each another [6]. CP with $[d(s_u, s_v)]^2$ may frequently split a large cluster in favor of a split with a slight reduction in squared error. The presence of outliers increases splits of clusters and decreases stability (shifting one outlier observation slightly has a large impact on the clusters).

The drawbacks in the use of $[d(s_u, s_v)]^2$ as a difference measure have been recognized for some time in facility location [16]. Although Kaufman and Rousseuw [12] advocate the use of $d(s_u, s_v)$ in their methods (i.e. PAM and CLARA) only a slight mention is given regarding the impacts due to outliers when $[d(s_u, s_v)]^2$ is used. The presence of outliers introduces large values of $d(s_u, s_v)$ relative to other values. The figures we have presented give a graphical illustration of why $d(s_u, s_v)$

should be used, rather than $[d(s_u, s_v)]^2$. Outliers, which by their nature should be regarded with less weight than other points, are given far much more weight than other points if the squared value is utilized.

It is notable that the recent literature in Knowledge Discovery for spatial data regard OI, CP and MP (and versions with $[d(s_u, s_v)]^2$) as almost equivalent to each other [17, 7, 24]. The concern has focused on computational resources, such as space, rather than solution concerns.

## 3 Solution methods

Obtaining optimal solutions for the three formulations of clustering presented for large data sets is unrealistic. They are NP-complete problems and cannot be expected to be solved optimally for the large number of observations (typically above several thousands) that Knowledge Discovery applications are pursuing. To find optimal solutions for problems of only about 50 observations (or perhaps 200 for MP) several approaches are possible [16]. For OI and CP the number of constraints and variables may be manageable using dynamic programming techniques, but the non-linearity of the objective function prohibits optimaly solving large instances of these problems, although linear transformations exists for OI and special cases of CP. MP is a linear criterion but with a quadratic number of decision variables and of constraints. Thus, heuristic approximations are essential. We will review heuristic approaches for OI, CP and MP among which, MP has received considerable attention in the literature [12, 15, 20]. In what follows, we assume that $n$ is large, the dimension $d$ is 2 or 3 and the number $k$ of clusters small, of the order of 10 or 50. Thus, $d$ and $k$ can be considered constant.

The OI formulation is, in a sense, the most attractive problem for Knowledge Discovery since the cohesion of clusters is maximized. However, no proposals for solving it for large instances appear in the literature. We will later discuss reason for this.

The CP formulation may be solved approximately using a variety of methods, many of them extremely efficient. $k$-MEANS [13] (or *Basic Isodata* [6]) is an iterative method that starts with a random set $C_0$ of $k$ centers and refines $C_t$ to $C_{t+1}$. Each refinement consists of using the partition defined by $C_t$ to compute a new set $C_{t+1}$. The new centers are the arithmetic means of the clusters of the previous partition. The method requires linear time (i.e. $\Theta(n)$ time) for a step advancing from $C_t$ to $C_{t+1}$, and thus, it is usually fast. Furthermore, it only requires $O(n)$ space. $k$-MEANS has foundations in the statistical analysis of fixture models [6]. From an optimization point of view, it usually converges to a local optima of poor quality and, from a pattern discovery point of view, it is sensitive to outliers or "noise".

Figure 4 presents three different clusterings obtained by $k$-MEANS (two of them are local optima). We can see that the few outliers have a significant effect on the result. Moreover, it is not hard to see that if the two outliers are slightly moved (or replaced by two randomly drawn observations) the partitions are likely to change. For this data set, a medoid approach results in the same clustering independent of the position of the outliers. The robustness of the medoid method is at the expense of additional computational resources, however. The fact that $k$-MEANS solutions are affected by outliers is concerning. On the other hand, for Data Mining, the generalization provided by $k$-MEANS could be qualitatively sufficient. For example, two of the partitions produced by $k$-MEANS in Figure 4 only differ in the classification of one outlier. Moreover, these solutions are qualitatively equivalent to the solution found by the medoid approach (in fact, one of them is the same). Thus, for deriving rules that are highly likely but not 100% exact, $k$-MEANS may be

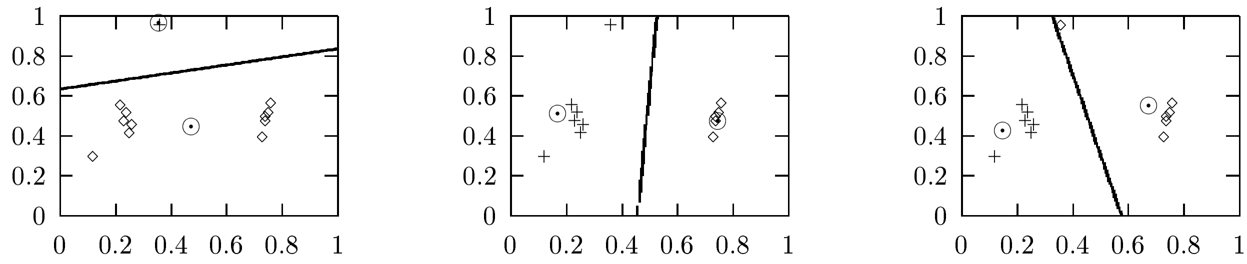

Figure 4: Three clusterings found by different executions of $k$-MEANS.

accurate enough; however, it has not been applied in data mining applications.

Other approaches to CP include Cooper's alternating heuristic [4], Simulated Annealing [22], Expectation Maximization [5] and Markov chain Monte-Carlo methods in Bayesian inferences [1] (in particular, Gibb's sampling). The solving of CP has not been used for cluster identification in GIS databases despite the computational efficiency of the methods for solving it (at least $k$-MEANS).

A variety of heuristic solution approaches exist for MP but require more computational effort than the approaches for solving CP. However, solving MP usually results in a clustering solution of higher quality.

The search space for MP is the set $X$ of all subsets $C \subseteq S$ with $\|C\| = k$ (thus, $X \subset 2^S$). The objective function $\mathsf{MP}(C) : X \to \Re$ assigns a quality clustering value to each subset $C \subseteq S$ as given by Equation (3). Recall that the elements of $C$ are taken as representatives and each $s_i$ belongs to its nearest representative. Heuristically searching $X$ for the smallest value $\mathsf{MP}(C)$ can be organized by providing $X$ with the structure of a graph whose set of nodes is the set $X$. Two nodes $C$ and $C'$ *are adjacent* if they differ in exactly one medoid (that is, $\|C \cap C'\| = k-1$). *Local Search* [11] in this graph is computationally feasible because computing $\mathsf{MP}(C')$ on an adjacent node $C'$ of $C$ requires $\Theta(n)$ time ($\Theta(n)$ time to find $rep[s_i, C']$ for $i \in \{1, \ldots, n\}$, and $\Theta(n)$ time to compute $\mathsf{MP}(C')$ as defined in Equation (3)). The medoid approach offers another advantage. The value $\mathsf{MP}(C')$ need not be computed, rather the gradient $\Delta(C, C') = \mathsf{MP}(C) - \mathsf{MP}(C')$ between $C$ and an adjacent node $C'$ can be used because it can be computed in $\Theta(n)$ time with a smaller constant factor under the $\Theta$.

Heuristics for MP have emerged in statistics, machine learning, knowledge discovery and operations research, and are essentially rediscoveries of each other because they are instances of *Local Search*. Such heuristics begin with a randomly selected solution $C_0$. Iteratively, the heuristic explores a subset $W$ of adjacent nodes of the current node $C_t$ and moves to the adjacent node $C_{t+1}$ that provides the best value of MP in $W$. The search halts when such a move is no longer possible. The move from $C_t$ to $C_{t+1}$ consists of interchanging one medoid $s_m \in C_t$ for one observation $s_u \in S - C_t$, so $C_{t+1} = S_t \cup \{s_u\} - \{s_m\}$. Thus, a local optima is reached when no interchange of a medoid for an observation results in an improved solution. Heuristics using this generic search schema are referred to as *interchange heuristics.*

**Global Hill climbing.-** GLOBAL HILL CLIMBING can be considered a discrete version of gradient-descent where the best local step towards the solution is taken. The window $W$ consists of all nodes adjacent to $C_t$, thus when the method halts a local optima is found. Note that the number of adjacent nodes to $C_t$ is $\Theta(n)$ and evaluating each requires $\Theta(n)$ time. GLOBAL HILL

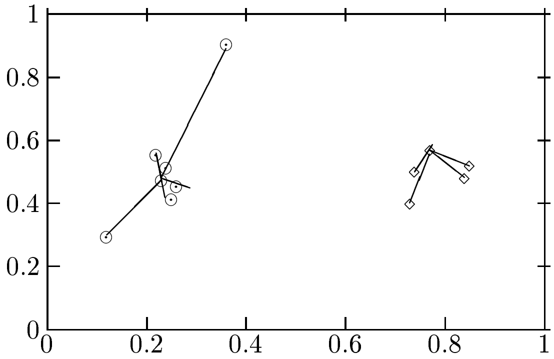


Figure 5: The MP criterion.

CLIMBING takes $\Theta(n^2)$ time to take a climbing step. For spatial clustering, this is the approach taken by PAM [12] and is embedded in derivations of PAM [17]. However, this is a rediscovery of the *global interchange* heuristic proposed by Goodchild and Noronha [9]. Solving MP using GLOBAL HILL CLIMBING has been shown to require more total computational effort than other interchange heuristics while finding solutions of equivalent quality [20].

**Randomized Hill climbing.-** Since $n$ is typically large for spatial databases, an alternative approach is not to visit every $C'$ adjacent node to $C_t$, but only a uniform sample of constant size [12, 17]. This has the advantage that it reduces the time it takes for moving from $C_t$ to a better $C_{t+1}$. If $m$ is the size of the sample $W$, then the step time is $\Theta(nm)$ time. However, the guarantee of a local optimum is lost. Sampling approaches have been shown to be inferior to GLOBAL HILL CLIMBING and other more sophisticated heuristics for rendering a high quality solution [14].

**Local Hill climbing.-** LOCAL HILL CLIMBING [15, 20] takes an approach that is between RANDOMIZED HILL CLIMBING and GLOBAL HILL CLIMBING. It remains deterministic and will find a local optimum just as GLOBAL HILL CLIMBING does. This method explores the adjacent nodes of $C_t$, but as soon as one with a better value for MP is found (an $C'$ with $\Delta(R_t, R') < 0$), the searcher declares this the new current solution $C_{t+1}$. LOCAL HILL CLIMBINGincorporates a search strategy to explore only a fraction of the adjacent nodes of the current solution $C_t$. This reduces the time to take an improvement to $O(n)$ time in most cases, and is an improvement of a factor of $n$ over GLOBAL HILL CLIMBING. To achieve this, the observations are enumerated as $s_i$, for $i = 1, \ldots, n$. Each $s_i$ that is not a medoid, is evaluated for a possible exchange with a medoid. This is $O(kn) = O(n)$ time per observation $s_i$ when $k$ is small and $n$ is large. The next round of exploration after an improvement is started with $s_{i+1}$ (with $s_1$ following $s_n$). A full round on the observations with no improvement gives the local optima and the search halts.

Results of LOCAL HILL CLIMBING applied to facility location problems have shown remarkable merits for this heuristic [15, 20]. However, to the best of our knowledge, this approach has not been investigated for clustering large sets of spatial data. Although LOCAL HILL CLIMBING moves to a better node $C_{t+1}$ faster than GLOBAL HILL CLIMBING, the relative and absolute improvement could potentially be smaller. Thus, in general search problems it may take several steps to achieve the improvement GLOBAL HILL CLIMBING makes in one step. Note that the complexity of the step $C_t$ to $C_{t+1}$ for LOCAL HILL CLIMBING is adaptive to the shape of the objective function. It requires $O(\delta_t n)$ time, where $s_{(i+\delta_t) mod\ n}$ is the next observation where an improvement occurs after the improvement that occurred at $s_i$. Note that the size of the window $W$ is $\delta_t$ and is dynamic. In the best case it is as small as 1, while at the local optima it consists of all adjacent nodes.

Because LOCAL HILL CLIMBING exploits the sharpness of the objective function, it is much more efficient than GLOBAL HILL CLIMBING. Our experimental results discussed later indicate that clustering spatial data is such an objective function with no cliffs. Figure 5 illustrates an adjacent node to Figure 3 that offers an immediate improvement. The solution in Figure 3 has a medoid on an outlier but has many adjacent nodes that are a significant improvement of MP. Each data point in the left cloud defines an exchange with the medoid which improves MP. So, at least for outliers, there are no cliffs and LOCAL HILL CLIMBING profits from this. Moreover, it will rarely place a representative on an outlier after a few steps.

**Distance restricted hill climbing.-** Given that MP is a problem within a geographical context, the window $W$ may be selected as a subset of the neighbors of $C_t$ using some geographical requirement. This approach attempts to reduce the candidate neighbors evaluated by requiring interchange candidates to be within a specific distance of the medoid they may replace. (recall that computing $d(s_m, s_u)$ requires constant time while evaluating a potential interchange requires $O(n)$ time). This approach is common in the statistical clustering literature and for spatial data has been recently rediscovered [12, 24]. After our description of LOCAL HILL CLIMBING it should be clear that this approach is inferior. Besides the problem of learning or discovering the cut-off value, this method cannot adapt its window to the region that it is exploring. The most obvious consequence is that it cannot guarantee local optimality. Moreover, this approach has been shown to result in poor quality solutions [23].

**Backtraking heuristics.-** The methods discussed thus far adopt a monotone descent for finding a local optima. Usually, it is impossible to reach the optimum by a sequence of solutions $\{R_t\}_{t \in N}$ whose function values are monotone. SIMULATED ANNEALING has been successfully used in clustering [22] as well as in location-planning models [15] (whose integer programming formulation is more generic than the corresponding formulation for spatial clustering). SIMULATED ANNEALING derives its name from the process of annealing solids where high energy is initially applied and the process cooled down. In our current setting, SIMULATED ANNEALING resembles a combination of RANDOMIZED HILL CLIMBING and LOCAL HILL CLIMBING. From the current solution $C_t$, the neighbors are randomly explored. When a neighbor $C'$ has an improved value $\Delta(R_t, R') < 0$, SIMULATED ANNEALING adopts $C'$ as $C_{t+1}$ and moves ahead as a LOCAL HILL CLIMBING. When the neighbor $C'$ offers a worse value of the objective function a random choice may accept it as $C_{t+1}$ (as a function of the parameters, namely the temperature $T$). Another analogical search is GENETIC ALGORITHMS which simulate the evolution process and have been used for clustering spatial data [8], but with very small data sets and with limited elliptical shape for the clouds forming the clusters. Our initial experimentation with SIMULATED ANNEALING and GENETIC ALGORITHMS has shown that their computational requirements are not competitive in clustering.

TABU SEARCH [20] can be considered an organized search that may be stopped when time resources reach a predefined limit or when a local optimum cannot be exited. TABU SEARCH will supply the best solution found so far. Thus, it is like repeating the search heuristic but not starting from a random point. TABU SEARCH typically works with LOCAL HILL CLIMBING since little space is required to save information about a decision taken on an earlier visit to a node, which facilitates backtracking. For large data sets, initial experimentation with TABU SEARCH shows that it is computationally inefficient.

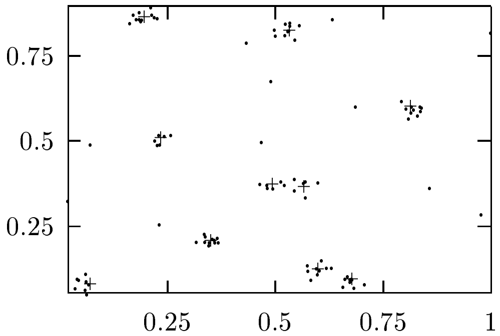


Figure 6: A test set.

**Other heuristics.-** The methods presented so far are considered local search methods because they progress towards convergence based solely on a neighborhood of a current solution. A very common approach to introduce some global search capability to these methods is simply to restart the method several times. The initial start point in the search space is reselected randomly, and the final solution is the best obtained among the individual local searches. For data-mining of spatial data, this is the approach taken in CLARANS [17], with 2 searches by RANDOMIZED HILL CLIMBING.

Another approach is to use a sample of the data set. This reduces the problem size from $n$ to the sample size, and if the patterns of interest are frequent, they should show up in the sample. A final approach is hybrids which combine two or more of the previous heuristics, using the solution of one as the starting point for the next, usually the second one being a more sophisticated but more computationally intensive approach. The idea is that the fast heuristic will quickly locate regions of interest in the search space to be exploited by the more robust but slower searcher. This can also be done across problems. For example, we could use the relatively efficient approaches for MP to obtain a starting point for approximately solving OI using a local search heuristic.

Perhaps now is the time to see why it is computationally more expensive to solve OI, even with local search. It is not hard to see that Equation (1) has $\Theta(n^2)$ terms in the sum (refer to Figure 1 (a)). Thus, computing $\mathsf{OI}(C)$ for a partition $C$ requires $\Theta(n^2)$ time. A naive implementation of a local search heuristic would require $\Theta(n^2)$ time per neighbor explored in $W$ for the current solution (for example, GLOBAL HILL CLIMBING would require $\Theta(n^3)$ per step to improve the current solution). With respect to CP and MP this is a factor of $O(n)$ extra per step towards the solution, which is infeasible. A less naive approach would be to structure the search space as a graph with nodes being the partitions of $P$. In this graph, two partitions are *adjacent* if all their parts are equal except for two, where one observation $s_u$ has been removed from one part but added to another part. This allows one to compute the gradient between two neighbors in $O(n)$ time, at the expense of enlarging the number of neighbors to the current solution from $n-k$ to $kn$.

# 4 Computational requirements

We have implemented the heuristics for solving OI, CP and MP in C so that most of the code is common except for the loop that moves from $C_t$ to $C_{t+1}$ (but the subroutines to compute $\Delta(C_t, C')$

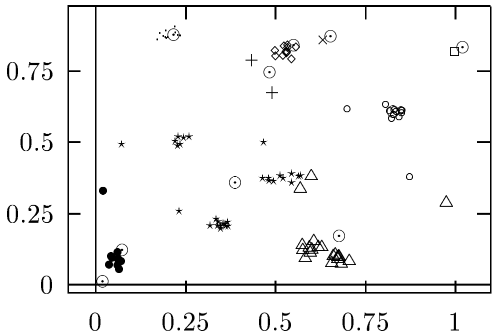


Figure 7: Clustering with $k$-Means.

are shared). This allows us to compare the efficiency of the methods with the profiler *gprof* and by CPU time. For this, we have generated test data by selecting $k$ points $\vec{c}_1, \ldots \vec{c}_k$ randomly and uniformly in $[0,1] \times [0,1]$. These points are virtual centers that are not part of the data set. The smallest separation $D = \min_{i \neq j} d(\vec{c}_i, \vec{c}_j)$ is used to determine a common virtual radius $r = D/2$. With this information, $(1-N)n$ data points are chosen by first selecting randomly a virtual center $\vec{c}_i$ and then selecting polar coordinates $\gamma, \psi$ ($\gamma$ is uniform in $[0,r]$ and $\psi$ is uniform in $[0, 2\pi]$). The point in the data set is $\vec{c}_i + (r \sin\gamma, r\cos\gamma)$. Note that at least two virtual clouds touch. The data set is shuffled with $Nn$ randomly and uniformly selected points where $N \in [0,1]$ is a percentage of the amount of noise. Figure 6 shows a test set generated with $n = 100$, $k = 10$ and $N = 10$. The virtual centers are shown with $+$.

Table 1 presents the average CPU time in seconds of 10 executions (and the corresponding 95% confidence interval) for some of the heuristic solution methods discussed here. Local Hill climbing proves to be a far more efficient method than approaches presented recently in the Data Mining literature. $k$-Means CPU time to convergence is $O(c_K n)$ where $c_K$ is the number of improving steps, and $c_K$ does not depend on $n$. Similarly, the total time to converge to a local optima for Global Hill climbing (and thus PAM) is $O(c_G n^2)$ time where $c_G$ are the number of improving steps. Randomized Hill climbing (and thus CLARANS) also requires $O(c_R n^2)$ time where $c_R$ are the number of its improving steps ($c_R$ is slightly smaller than $c_G$). The total time for convergence to a local optima of Local Hill climbing is $O(c_L n)$ time where $c_L$ are the number of its improving steps, but $c_L$ is not a constant but $c_L = \sum_{t=1} \delta_t$. Experimentally, $c_L \approx 4n$ for Local Hill climbing to converged. In this time Global Hill climbing has only taken 4 improving steps towards an equivalent local optima.

It should be noted that $k$-Means is faster than Local Hill climbing, but Local Hill climbing solutions appear to be of better quality. Figure 7 shows the clustering using $k$-Means, where centers are indicated with $\odot$. Figure 8 illustrates that Local Hill climbing discovers 9 of the 10 clusters, and only places one medoid on an outlier. However, although the data was generated with $k = 10$ as explained before, two virtual centers happen to be so close that a statistical test favors the use of 9 groups rather than 10. Thus, Local Hill climbing finds a high quality solution that is superior to $k$-Means, which merges three far apart clusters into one. As one would expect, because Local Hill climbing finds a local optima, the quality of the solution by Local Hill climbing is not improved by PAM (Global Hill climbing) nor by CLARANS (Randomized

| $n$ | $k$-Means | | Local Hill climbing | | CLARANS | | PAM | |
|---|---|---|---|---|---|---|---|---|
| 100 | 0.08 | ± 0.02 | 10.1 | ± 2.0 | 44.2 | ± 4.2 | 61 | ± 8.1 |
| 200 | 0.17 | ± 0.04 | 41.2 | ± 5.3 | 204.3 | ± 32.4 | 350.3 | ± 60.9 |
| 500 | 0.50 | ± 0.21 | 338.2 | ± 10.3 | 1435.3 | ± 123.8 | 2318.7 | ± 240.2 |
| 1000 | 1.01 | ± 0.30 | 939.2 | ± 42.3 | - | - | - | - |

Table 1: Solution times in CPU seconds.

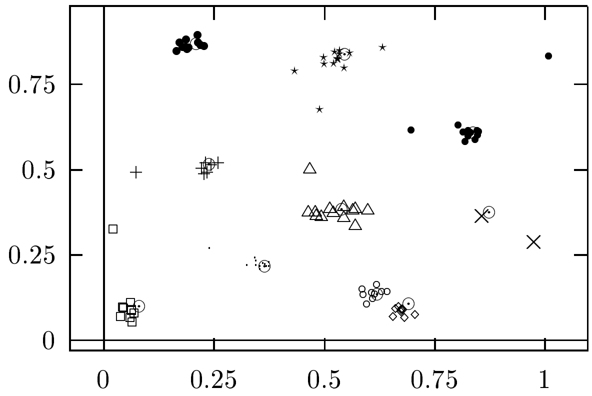


Figure 8: Solution via Local Hill climbing.

Hill climbing).

# 5 Conclusion

We have presented a panoramic view of non-hierarchical clustering problems and solution methods. We have noted that the use of clustering is central for pattern discovery in spatial data but the methods that have emerged in the field of Data Mining fail to recognize and benefit from previous research. This has allowed us to propose Local Hill climbing as an immediate improvement for clustering spatial data using the medoid approach. Also, we have underlined why the use of $[d(s_u, s_v)]^2]$ should be avoided, most notably when data contains outliers. This is in contrast to doing clustering in several phases, each phase filtering some outliers [24]. We have clarified the virtues of the medoid approach for solving the MP problem; namely

- it is approximately solvable by local search in $O(n)$ space (and not $O(n^2)$ space [7]),
- it offers a potetial approach for solving OI as an initial method in a hybrid local search,
- it is robust with respect to outliers.

We should point out that most of the literature in facility location and operations research regards a local search method as superior when, for equivalent computational effort, the solutions obtained are of higher quality in terms of the optimization criteria. Perhaps for Knowledge Discovery this is not the best criteria for evaluating a clustering technique. For pattern spotting, an optimal solution for $\mathsf{MP}(C)$ or $\mathsf{CP}(C)$ is not required, rather what is required is that the partition defined by using $C$ as representatives results in a Voronoi Diagram whose components closely reflect the spatial patterns in data. Figures 3 and Figure 5 give two different sets of medoid solutions such that $\mathsf{MP}$ is much worse in one case than the other. However, the clusterings are equivalent. For spatial pattern spotting one may be willing to sacrifice quality in the final value of $\mathsf{MP}(C)$, $\mathsf{CP}(C)$ or even $\mathsf{OI}(P)$ if in fact the partition is close enough to make the pattern known. Figure 7 contrasts this perspective in that the quality of the partition produced by $k$-MEANS is poor with just a few outliers despite the fact that the patterns are well concentrated clouds. On the other hand, the time requirements for local search methods grows significantly with $n$, and for data mining applications, we should expect $n$ to be large. The extreme is PAM (that is, GLOBAL HILL CLIMBING). We can run $k$-MEANS almost an extra time for each additional point since $Time_{Global}(n) \approx nTime_{kmeans}$ or

$$\begin{aligned} Time_{Global}(n+1) &\approx (n+1)Time_{kmeans} \\ &= Time_{Global}(n) + Time_{kmeans} \end{aligned}$$

Since there is such a large gap between the computational requirements of $k$-MEANS for approximately solving $\mathsf{CP}$ and local search methods for approximately solving $\mathsf{MP}$, it seems natural to seek a balance. Perhaps methods like Expectation-Maximization or Gibb's sampling offer an alternative for such balance. How would this compare with a new hybrid approach, or to our approach suggested here for solving $\mathsf{OI}$? We expect to address such questions in future research.

## References


[1] J. Besang and P.J. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society B*, 55(1):25–77, 1993. part of a series of papers on Gibbs sampling.

[2] T. Brinkhoff and H.P. Kriegel. The impact of global clustering on spatial database systems. In J. Bocca, M. Jarke, and C. Zaniolo, editors, *Proceedings of the 20th Conference on Very Large Data Bases (VLDB)*, pages 168–179, San Francisco, CA, 1994. Santiago, Chile, Morgan Kauffman Publishers.

[3] Y. Cai, N. Cercone, and J. Han. Attribute-oriented induction in relational databases. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery in Databases*, pages 213–228, Menlo Park, CA. USA, 1991. AAAI Press.

[4] L. Cooper. Heuristic methods for location-allocation problems. *SIAM Review*, 6:37–53, 1964.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likehood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[6] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, US, 1973.

[7] M. Ester, H.P. Kriegel, S. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, and U. Fayyad, editors, *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231, Menlo Park, CA, 1996. AAAI, AAAI Press.

[8] A. Ghozeil and D.B. Fogel. Discovering patterns in spatial data using evolutionary programming. In J.R. Koza, editor, *Genetic Programming: Proceedings of the First Annual Conference*, pages 521–527, Cambridge, MA, 1996. MIT Press.

[9] M Goodchild and V. Noronha. Location-allocation for small computers. Monograph 8, University of Iowa, 1983.

[10] R. Jensen. A dynamic programming algorithm for cluster analysis. *Operations Research*, 12:1034–1057, 1969.

[11] D.S. Johnson, C.H. Papadimitrou, and M. Yanakakis. How easy is local search? *Journal of Computer System Sciences*, 37:79–100, 1988.

[12] L. Kaufman and P.J. Rousseuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons, NY, US, 1990.

[13] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L. Le Cam and J. Neyman, editors, *5th Berkley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967. Volume 1.

[14] A.T. Murray and R.L. Church. Heuristic solution approaches to operational forest planning problems. *OR Spektrum*, 17:193–203, 1995.

[15] A.T. Murray and R.L. Church. Applying simulated annealing to location-planning models. *Journal of Heuristics*, 2:31–53, 1996.

[16] A.T. Murray and V. Estivill-Castro. Cluster discovery techniques for exploratory spatial data analysis. Submitted manuscript, 1997.

[17] R.T. Ng and J. Han. Efficient and effective clustring methods for spatial data mining. In J. Bocca, M. Jarke, and C. Zaniolo, editors, *Proceedings of the 20th Conference on Very Large Data Bases (VLDB)*, pages 144–155, San Francisco, CA, 1994. Santiago, Chile, Morgan Kauffman Publishers.

[18] S. Openshaw. Two exploratory space-time-attribute pattern analysers relevant to GIS. In S. Fotheringham and P. Rogerson, editors, *Spatial Analysis and GIS*, pages 83–104, London, UK, 1994. Taylor and Francis.

[19] M. Rao. Cluster analysis and mathematical programming. *Journal of the American Statistical Association*, 66:622–626, 1971.

[20] D. Rolland, E. Schilling and J. Current. An efficient tabu search procedure for the $p$-median problem. *European Journal of Operations Research*, 96:329–342, 1996.

[21] K. Rosing and C. ReVelle. Optimal clustering. *Environment and Planning A*, 18:1463–1476, 1986.

[22] S.Z. Selim and K. Alsultan. A simulated annealing algorithm for the clustering problem. *Pattern Recognition*, 24(10):1003–1008, 1991.

[23] P. Sorensen. Analysis and design of heuristics for the $p$-median location-allocation problem. Master's thesis, Department of Geography, University of California, Santa Barbara, 1994.

[24] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH:an efficient data clustering method for very large databases. *SIGMOD Record*, 25(2):103–114, June 1996. Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data.