Towards Region Discovery in Spatial Datasets

Wei Ding[1,⋆], Rachsuda Jiamthapthaksin[1], Rachana Parmar[1], Dan Jiang[1],
Tomasz F. Stepinski[2], and Christoph F. Eick[1]

[1] University of Houston, Houston TX 77204-3010, USA
{wding,rachsuda,rparmar,djiang,ceick}@uh.edu
[2] Lunar and Planetary Institute, Houston, TX 77058, USA
tstepinski@lpi.usra.edu

Abstract. This paper presents a novel region discovery framework geared towards finding scientifically interesting places in spatial datasets. We view region discovery as a clustering problem in which an externally given fitness function has to be maximized. The framework adapts four representative clustering algorithms, exemplifying prototype-based, grid-based, density-based, and agglomerative clustering algorithms, and then we systematically evaluated the four algorithms in a real-world case study. The task is to find feature-based hotspots where extreme densities of deep ice and shallow ice co-locate on Mars. The results reveal that the density-based algorithm outperforms other algorithms inasmuch as it discovers more regions with higher interestingness, the grid-based algorithm can provide acceptable solutions quickly, while the agglomerative clustering algorithm performs best to identify larger regions of arbitrary shape. Moreover, the results indicate that there are only a few regions on Mars where shallow and deep ground ice co-locate, suggesting that they have been deposited at different geological times.

Keywords: Region Discovery, Clustering, Hotspot Discovery, Spatial Data Mining.

1 Introduction

The goal of spatial data mining [1,2,3] is to automate the extraction of interesting and useful patterns that are not explicitly represented in spatial datasets. Of particular interests to scientists are the techniques capable of finding scientifically meaningful regions as they have many immediate applications in geoscience, medical science, and social science; e.g., detection of earthquake hotspots, disease zones, and criminal locations. An ultimate goal for region discovery is to provide search-engine-style capabilities to scientists in a highly automated fashion. Developing such a system faces the following challenges. First, the system must be able to find regions of arbitrary shape at different levels of resolution. Second, the system needs to provide suitable, plug-in measures of interestingness to instruct discovery algorithms what they should seek for. Third, the identified regions should be properly ranked by relevance.

[⋆] Also, Computer Science Department, University of Houston-Clear Lake.

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 88–99, 2008.

Fourth, the system must be able to accommodate discrepancies in various formats of spatial datasets. In particular, the discrepancy between continuous and discrete datasets poses a challenge, because existing data mining techniques are not designed to operate on a mixture of continuous and discrete datasets. Fifth, it is desirable for the framework to provide pruning and other sophisticated search strategies as the goal is to seek for interesting, highly ranked regions.

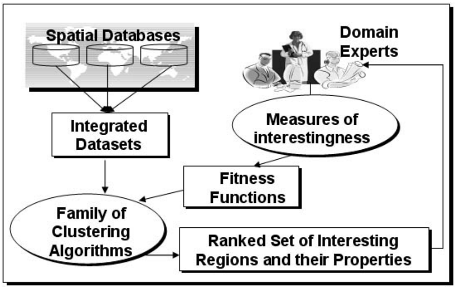

Spatial Databases
Domain Experts
Integrated Datasets
Measures of interestingness
Fitness Functions
Family of Clustering Algorithms
Ranked Set of Interesting Regions and their Properties

Fig. 1. Region discovery framework

This paper presents a novel region discovery framework (see Fig. 1) geared towards finding scientifically interesting places in spatial datasets. We view region discovery as a clustering problem in which an externally given fitness function has to be maximized. The framework adapts four representative clustering algorithms, exemplifying prototype-based, grid-based, density-based, and agglomerative clustering algorithms for the task of region discovery. The fitness function combines contributions of interestingness from constituent clusters and can be customized by domain experts. The framework allows for plug-in fitness functions to support a variety of region discovery applications correspondent to different domain interests.

Relevant Work. Many studies have been conducted in region discovery. These most relevant to our work are region-oriented clustering techniques and hotspot discovery. In our previous work, we have discussed a region discovery method that was restricted to one categorical attribute [4,5]. The integrated framework introduced in this paper is generalized to be applicable to both continuous and discrete datasets. The framework allows for various plug-in fitness functions and extends our work to the field of feature-based hotspot discovery (see Section 2). [1] introduces a “region oriented” clustering algorithm to select regions to satisfy certain condition such as density. This approach uses statistical information instead of a fitness function to evaluate a cluster.

Hotspots are object clusters with respect to spatial coordinates. Detection of hotspots using variable resolution approach [6] was investigated in order to minimize the effects of spatial superposition. In [7] a region growing method for hotspot discovery was described, which selects seed points first and then grows clusters from these seed points by adding neighbor points as long as a density threshold condition is satisfied. Definition of hotspots was extended in [8] using circular zones for multiple variables.

Contributions. This paper presents a highly generic framework for region discovery in spatial datasets. We customize our discovery framework to accommodate raster, continuous, and categorical datasets. This involves finding a suitable object structure, suitable preprocessing techniques, a family of reward-based fitness functions for various measures of interestingness, and a collection of clustering algorithms. We systematically evaluate a wide range of representative clustering algorithms to determine when and which type of clustering techniques are more suitable for region discovery. We apply our framework to a real-world case study concerning ground ice on Mars and successfully find scientifically interesting places.

2 Methodology

Region Discovery Framework. Our region discovery method employs a reward-based evaluation scheme that evaluates the quality of the discovered regions. Given a set of regions $R = \{r_1, \ldots, r_k\}$ identified from a spatial dataset $O = \{o_1, \ldots, o_n\}$, the fitness of R is defined as the sum of the rewards obtained from each region r_j $(j = 1 \ldots k)$:

$$q(R) = \sum_{j=1}^{k} (i(r_j) \times size(r_j)^{\beta}) \tag{1}$$

where $i(r_j)$ is the interestingness measure of region r_j – a quantity based on domain interest to reflect the degree to which the region is "newsworthy". The framework seeks for a set of regions R such that the sum of rewards over all of its constituent regions is maximized. $size(r_j)^{\beta}$ $(\beta > 1)$ in $q(R)$ increases the value of the fitness nonlinearly with respect to the number of objects in O belonging to the region r_j. A region reward is proportional to its interestingness, but given two regions with the same value of interestingness, a larger region receives a higher reward to reflect a preference given to larger regions.

We employ clustering algorithms for region discovery. A region is a contiguous subspace that contains a set of spatial objects: for each pair of objects belonging to the same region, there always exists a path within this region that connects them. We search for regions $r_1, \ldots, r_k$ such that:

1. $r_i \cap r_j = \emptyset, i \neq j$. The regions are disjoint.
2. $R = \{r_1, \ldots, r_k\}$ maximizes $q(R)$.
3. $r_1 \cup \ldots \cup r_k \subseteq O$. The generated regions are not required to be exhaustive with respect to the global dataset O.
4. $r_1, \ldots, r_k$ are ranked based on their reward values. Regions that receive no reward are discarded as outliers.

Preprocessing. Preprocessing techniques are introduced to facilitate the application of the framework to heterogeneous datasets. Given a collection of raster, categorical, and continuous datasets with a common spatial extent, the raster datasets are represented as (<pixel>, <continuous variables>), the categorical dataset as (<point>, <category variables>)[1], and the continuous datasets as (<point>, <continuous variables>). Fig. 2 depicts our preprocessing procedure:

[1] To deal with multiple categorical datasets a single dataset can be constructed by taking the union of multiple categorical datasets.

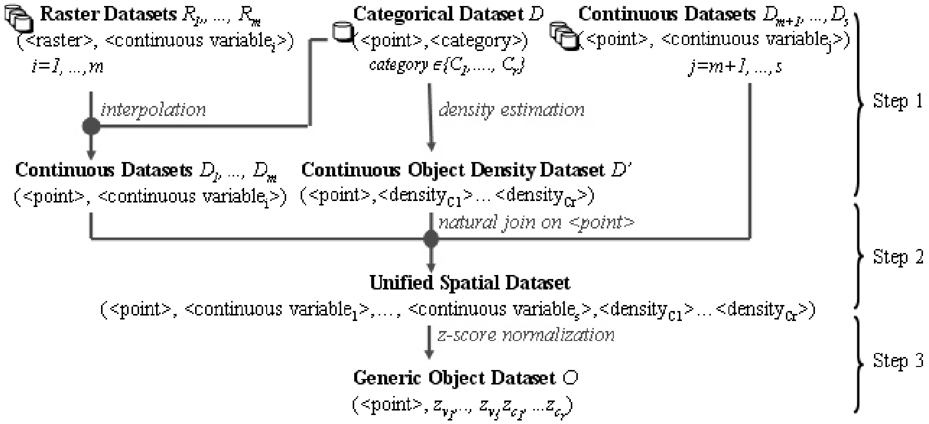

Raster Datasets R_1, ..., R_m
(<raster>, <continuous variable$_i$>)
$i=1, ..., m$
Categorical Dataset D
(<point>,<category>)
category $\in\{C_1, ..., C_r\}$
Continuous Datasets D_{m+1}, ..., D_s
(<point>, <continuous variable$_j$>)
$j=m+1, ..., s$
interpolation
density estimation
Step 1
Continuous Datasets D_1, ..., D_m
(<point>, <continuous variable$_i$>)
Continuous Object Density Dataset D'
(<point>,<density$_{C1}$>...<density$_{Cr}$>)
natural join on <point>
Step 2
Unified Spatial Dataset
(<point>, <continuous variable$_1$>,..., <continuous variable$_s$>,<density$_{C1}$>...<density$_{Cr}$>)
z-score normalization
Step 3
Generic Object Dataset O
(<point>, z_{v_1},.., z_{v_s}, z_{c_1}, ...z_{c_r})

Fig. 2. Preprocessing for heterogeneous spatial datasets

Step 1. Dataset Integration. Categorical datasets are converted into a continuous density dataset (<point>, <density variables>), where a density variable describes the density of a class for a given point. Classical density estimation techniques [9], such as Gaussian kernel functions, can be used for such transformation. Raster datasets are mapped into point datasets using interpolation functions that compute point values based on the raster values.

Step 2. Dataset Unification. A single unified spatial dataset is created by taking a natural join on the spatial attributes of each dataset. Notice that the datasets have to be made "join compatible" in Step 1. This can be accomplished by using the same set of points in each individual dataset.

Step 3. Dataset Normalization. Finally, continuous variables are normalized into z-scores to produce a generic dataset O=(<point>, <z-scores>), where z-score is the number of standard deviations that a given value is above or below the mean.

Measure of Interestingness. The fitness function $q(R)$ (Eqn. 1) allows a function of interestingness to be defined based on different domain interests. In our previous work, we have defined fitness functions to search risk zones of earthquakes [4] and volcanoes [5] with respect to a single categorical attribute. In this paper, we define *feature-based hotspots* as localized regions where continuous non-spatial features of objects attain *together* the values from the wings of their respective distributions. Hence our feature-based hotspots are places where multiple, potentially globally uncorrelated attributes happen to attain extreme values. We then introduce a new interestingness function i on the top of the generic dataset O: given set of continuous features $A = \{A_1, ..., A_q\}$ the interestingness of an object $o \in O$ is measured as follows:

$$i(A, o) = \prod_{j=1}^{q} z_{A_j}(o) \qquad (2)$$

where $z_{A_j}(o)$ is the z-score of the continuous feature A_j. Objects with $|i(A,o)| \gg 0$ are clustered in feature-based hotspots where the features in A happen to attain extreme values—measured as products of z-scores.

We then extend the definition of interestingness to regions: the interestingness of a region r is the absolute value of the average interestingness of the objects belonging to it:

$$i(A,r) = \begin{cases} (\frac{|\Sigma_{o \in r}\ i(A,o)|}{size(r)} - z_{\text{th}}) & \text{if } \frac{|\Sigma_{o \in r}\ i(A,o)|}{size(r)} > z_{\text{th}} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

In Eqn. 3 the interestingness threshold z_{th} is introduced to weed out regions with $i(r)$ close to 0, which prevents clustering solutions from containing only large clusters of low interestingness.

Clustering Algorithms. Our regional discovery framework relies on reward-based fitness functions. Consequently, clustering algorithms embedded in the framework, have to allow for plug-in fitness functions. However, the use of fitness function is quite uncommon in clustering, although a few exceptions exist, e.g., CHAMELEON [10]. Furthermore, region discovery is different from traditional clustering as it gears to find interesting places with respect to a given measure of interestingness. Consequently, existing clustering techniques need to be modified extensively for the task of region discovery. The proposed region discovery framework adapts a family of prototype-based, agglomerative, density-based, and grid-based clustering approaches. We give a brief survey of these algorithms in this section.

Prototype-based Clustering Algorithms. Prototype-based clustering algorithms first seek for a set of representatives; clusters are then created by assigning objects in the dataset to the closest representatives. We introduce a modification of the PAM algorithm [11] which we call SPAM (Supervised PAM). SPAM starts its search with a random set of k representatives, and then greedily replaces representatives with non-representatives as long as $q(R)$ improves. SPAM requires the number of clusters, k, as an input parameter. Fig. 3a illustrates the application of SPAM to a supervised clustering task in which purity of clusters with respect to the instances of two classes has to be maximized. SPAM correctly separates cluster A from cluster B because the fitness value would be decreased if the two clusters were merged, while the traditional PAM algorithm will merge the two clusters because they are in close proximity.

Agglomerative Algorithms. Due to the fact that prototype-based algorithms construct clusters using nearest neighbor queries, the shape of clusters identified are limited to convex polygons (Voronoi cells). Interesting regions, and in particular, hotspots, may not be restricted to convex shapes. Agglomerative clustering algorithms are capable of yielding solutions with clusters of arbitrary shape by constructing unions of small convex polygons. We adapt the MOSAIC algorithm [5] that takes a set of small convex clusters as its input and greedily merges neighboring clusters as long as $q(R)$ improves. In our experiments

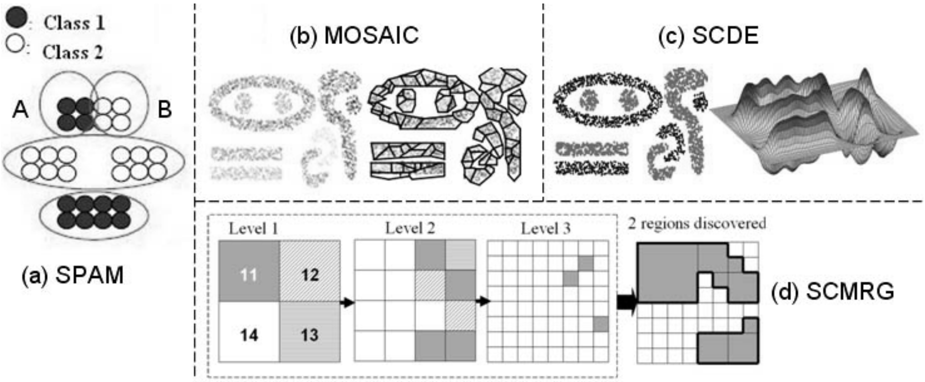


Fig. 3. Clustering algorithms

the inputs are generated by the SPAM algorithm. Gabriel graphs [12] are used to determine which clusters are neighbors. The number of clusters, k, is then implicitly determined by the clustering algorithm itself. Fig. 3b illustrates that MOSAIC identifies 9 clusters (4 of them are in non-convex shape) from the 95 small convex clusters generated by SPAM.

Density-Based Algorithms. Density-based algorithms construct clusters from an overall *density function*. We adapt the SCDE (Supervised Clustering Using Density Estimation) algorithm [13] to search feature-based hotspots. Each object o in O is assigned a value of $i(A,o)$ (see Eqn. 2). The influence function of object o, $f_{Gauss}(p,o)$, is defined as the product of $i(A,o)$ and a Gaussian kernel:

$$f_{Gauss}(p,o) = i(A,o) \times e^{-\frac{d(p,o)^2}{2\sigma^2}}. \tag{4}$$

The parameter σ determines how quickly the influence of o on p decreases as the distance between o and p increases. The density function, $\Psi(p)$ at point p is then computed as:

$$\Psi(p) = \sum_{o \in O} f_{Gauss}(p,o). \tag{5}$$

Unlike traditional density estimation techniques, which only consider the spatial distance between data points, our density estimation approach additionally considers the influence of the interestingness $i(A,o)$. SCDE uses a hill climbing approach to compute local maxima and local minima of the density function Ψ. These locales act as cluster attractors; clusters are formed by associating objects in O with the attractors. The number of clusters, k, is implicitly determined by the parameter σ. Fig. 3c illustrates an example in which SCDE identifies 9 regions that are associated with maxima (in red) and minima (in blue) of the depicted density function on the right.

Grid-based Algorithms. SCMRG (Supervised Clustering using Multi-Resolution Grids) [4] is a hierarchical, grid-based method that utilizes a divisive, top-down search. The spatial space of the dataset is partitioned into grid cells. Each grid cell at a higher level is partitioned further into smaller cells at the lower level, and this process continues as long as the sum of the rewards of the lower level cells $q(R)$ is not decreased. The regions returned by SCMRG are combination of grid cells obtained at different level of resolution. The number of clusters, k, is calculated by the algorithm itself. Fig. 3d illustrates that SCMRG drills down 3 levels and identifies 2 clusters (the rest of cells are discarded as outliers due to low interestingness).

3 A Real-World Case Study: Ground Ice on Mars

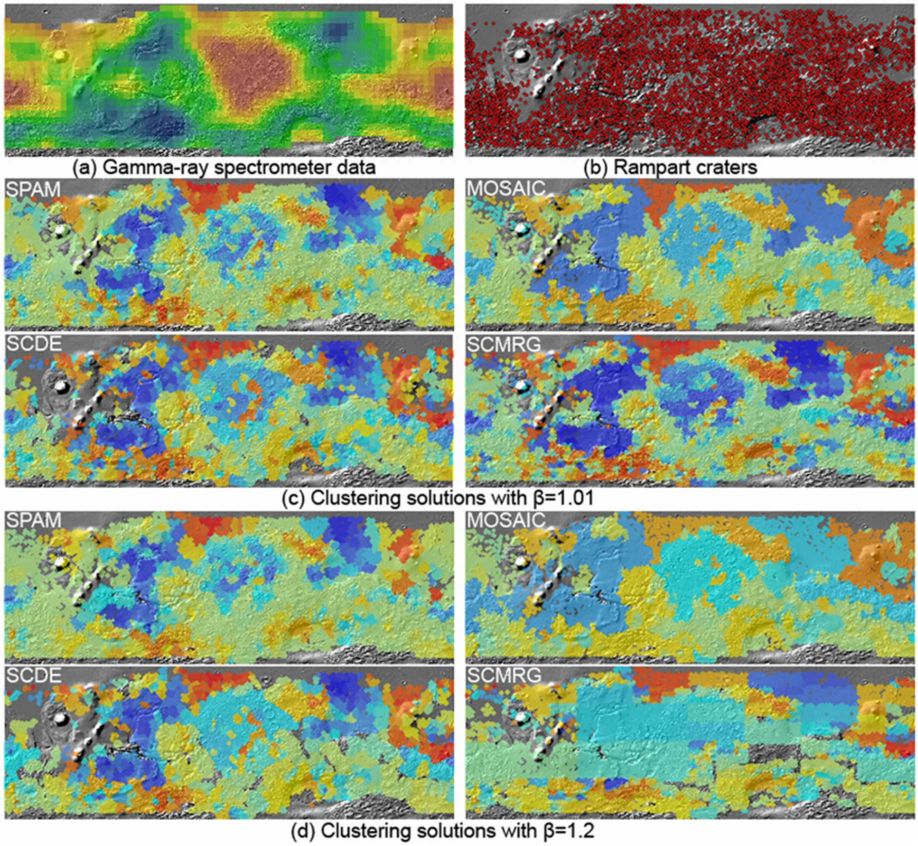
Dataset Description and Preprocessing. We systematically evaluate our region discovery framework on spatial distribution of ground ice on Mars. Mars is at the center of the solar system exploration efforts. Finding scientifically interesting places where shallow and deep ice abundances coincide provides important insight into the history of water on Mars. *Shallow ice* located in the shallow subsurface of Mars, within an upper 1 meter, is obtained remotely from orbit by the gamma-ray spectrometer [14] (see Fig. 4a, shallow ice in $5^o \times 5^o$ resolution). A spatial distribution of *deep ice*, up to the depth of a few kilometers, can be inferred from spatial distribution of rampart craters [15] (see Fig. 4b, distribution of 7559 rampart craters restricted to the spatial extent defined by the shallow ice raster). Rampart craters, which constitute about 20% of all the 35927 craters on Mars, are surrounded by ejecta that have patterns like splashes and are thought to form in locations once rich in subsurface ice. Locally-defined relative abundance of rampart craters can be considered a proxy for the abundance of deep ice.

Using the preprocessing procedure outlined in Section 2 we construct a generic dataset (<longitude, latitude>, z_{di}, z_{si}) where <longitude, latitude> is the coordinate of each rampart crater, z_{di} denotes the z-score of deep ice and z_{si} denotes the z-score of shallow ice. The values of these two features at location p are computed using a $5^o \times 5^o$ moving window wrapped around p. The shallow ice feature is an average of shallow-ice abundances as measured at locations of objects within the window, and the deep-ice feature is a ratio of rampart to all the craters located within the window.

Region Discovery Results. SPAM, MOSAIC, SCDE, and SCMRG clustering algorithms are used to find feature-based hotspots where extreme values of deep ice and shallow ice co-locate on Mars. The algorithms have been developed in our open source project *Cougar*[2] *Java Library for Machine Learning and Data Mining Algorithms* [16]. In the experiments, the clustering algorithms maximize the following fitness function $q(R)$ — see also Eqn 1:

$$q(R) = \sum_{r \in R} (i(\{z_{di}, z_{si}\}, r) \times size(r)^{\beta}) \tag{6}$$

For the purpose of simplification, we will use z for $i(\{z_{di}, z_{si}\}, r)$ in the rest of the paper. In the experiments, the interestingness threshold is set to be $z_{\text{th}} = 0.15$



(a) Gamma-ray spectrometer data

(b) Rampart craters

(c) Clustering solutions with β=1.01

(d) Clustering solutions with β=1.2

Fig. 4. Grayscale background depicts elevation of the Martian surface between longitude of -180^o to 180^o and latitude -60^o to 60^o. Legend indicates z value for each cluster. Objects not belonging to regions identified are not shown (better viewed in color).

and two different β values are used: $\beta = 1.01$ is used for finding stronger hotspots characterized by higher values of z even if the sizes are small, and $\beta = 1.2$ for identifying larger but likely weaker hotspots. Table 1 summarizes the experimental results. Fig. 4c shows the correspondent clustering results using $\beta = 1.01$. And Fig. 4d demonstrates that larger (but weaker) hotspots are identified for $\beta = 1.2$. Objects (craters) are color-coded according to the z values of clusters to which they belong. The hotspots are in the locations where objects are coded by either deep red or deep blue colors. In the red-coded hotspots the two variables have values from the same-side wings of their distributions (high-high or low-low). In the blue-coded hotspots the two variables have values from the opposite-side wings of their distributions (high-low or low-high).

Which clustering algorithm produces the best region discovery results? In the rest of section, we evaluate the four clustering algorithms with respect to

Table 1. Parameters of clustering algorithms and statistical analysis

	SPAM	**SCMRG**	**SCDE**	**MOSAIC**
$\beta = 1.01/\beta = 1.2$				
Parameters	$k = 2000/k = 807$	None	$\sigma = 0.1/\sigma = 1.2$	None
q(R)	13502/24265	14129 / 34614	14709/39935	14047/59006
# of clusters	2000/807	1597/644	1155/613	258/152
Statistics of Number of Objects Per Region				
Max	93/162	523/2685	1258/3806	4155/5542
Mean	18/45	15/45	25/49	139/236
Std	10/25	31/201	80/193	399/717
Skewness	1.38/1.06	9.52/10.16	9.1/13.44	6.0/5.24
Statistics of Rewards Per Region				
Max	197/705	743/6380	671/9488	3126/16461
Mean	10/46	9/54	12/65	94/694
Std	15/66	35/326	38/415	373/2661
Skewness	5.11/4.02	13.8/13.95	10.1/19.59	6.24/4.69
Statistics of $\sqrt{z}$ Per Region				
Max	2.7/2.45	2.85/2.31	2.95/2.94	1.24/1.01
Mean	0.6/0.57	0.74/0.68	0.95/0.97	0.44/0.40
Std	0.38/0.36	0.31/0.26	0.47/0.47	0.24/0.22
Skewness	1.14/1.34	1.58/1.88	1.28/1.31	0.73/0.40

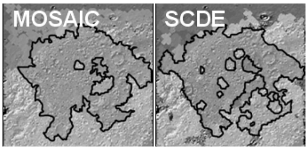
statistical measures, algorithmic consideration, shape analysis, and scientific contributions.

Statistical Measures. Table 1 is divided into four sections. The first section reports on the overall properties of clustering solutions: the parameters used by the clustering algorithms, the total reward and the number of regions discovered. The remaining three sections report on statistics of three different properties: region size, its reward on the population of the constituent regions, and $\sqrt{z}$, the square root of the interestingness of regions. The SPAM algorithm requires an input parameter k, which is chosen to be a value that is of the same order of magnitude as the values of k yielded by the SCMRG and SCDE algorithms. Due to its agglomerative character the MOSAIC algorithm always produces a significantly smaller number of clusters regardless of the size of its input provided by the SPAM clustering solution. Thus the MOSAIC is separated from the other solutions in the table.

To seek for feature-based hotspots of shallow ice and deep ice, the solution that receives high value of $q(R)$ and provides more clusters with the highest values of $\sqrt{z}$ is the most suitable. This is the solution having a large value of skewness for the reward and $\sqrt{z}$ properties. Skewness measures the asymmetry of the probability distribution, as the large value of skewness indicates existence of hotspots (more extreme values). In addition a suitable solution has larger values of the mean and the standard deviation for the reward and $\sqrt{z}$ properties, as they indicate existence of stronger hotspots. The analysis of Table 1 indicates that SCDE and SCMRG algorithms are more suitable to discovery hotspots with higher values in z. Furthermore, we are interested in evaluating the search capability, how the top n regions are selected by the four algorithms. Fig. 5a illustrates the average region size with respect to the top $99^{th}, 97^{th}, 94^{th}, 90^{th}, 80^{th}, 60^{th}$ percentile for the value of interetingness z. Fig. 5b depicts the average value of interestingness per cluster with respect to the top 10 largest regions. We observe that SCDE can pinpoint stronger hotspots in smaller size (e.g., $size = 4$ and $z = 5.95$), while MOSAIC is the better algorithm for larger hotspots with relatively higher value of interestingness (e.g., $size = 2096$ and $z = 1.38$).

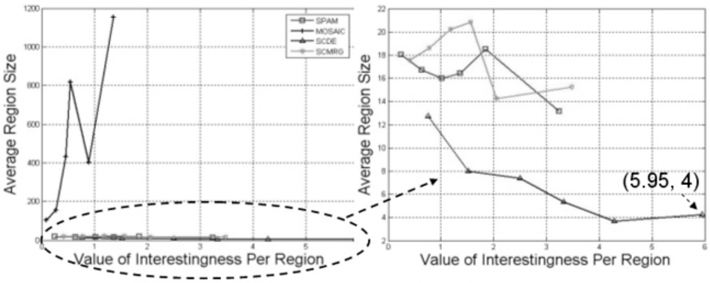
Algorithmic Considerations. As determined by the nature of the algorithm, SCDE and SCMRG algorithms support the notion of outliers – both algorithms evaluate and prune low-interest regions (outliers) dynamically during the search procedure. Outliers create an overhead for MOSAIC and SPAM because both algorithms are forced to create clusters to separate non-reward regions (outliers) from reward regions. Assigning outliers to a reward region in proximity is not an alternative because this would lead to a significant drop in the interestingness value and therefore to a significant drop in total rewards.

The computer used in our experiments is Intel(R) Xeon, CPU 3.2GHz, 1GB of RAM. In the experiments of $\beta = 1.01$ the SCDE algorithm takes $\sim 500s$ to complete, whereas the SCMRG takes $\sim 3.5s$, the SPAM takes $\sim 50000s$, and the MOSAIC took $\sim 155000s$. Thus, the SCMRG algorithm is significantly faster than the other clustering algorithms and, on this basis, it could be a suitable candidate to searching for hotspots in a very large dataset with limited time.

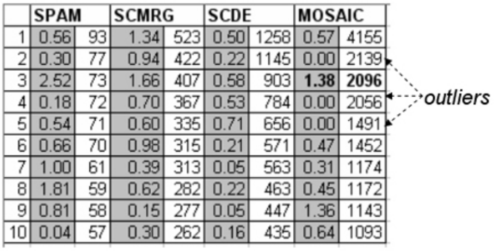


Shape Analysis. As depicted in Fig. 4, in contrast to SPAM whose shapes are limited to convex polygons, and SCMRG whose shapes are limited to unions of grid-cells, MOSAIC and SCDE can find arbitrary-shaped clusters. The SCMRG algorithm only produces good solutions for small values of β, as larger values of β lead to the formation of large, boxy segments that are not effective in isolating the hotspots. In addition, the figure on the right depicts the area of Acidalia Plantia on Mars (centered at $\sim -15^{o}$ longitude, -40^{o} latitude). MOSAIC and SCDE have done a good job in finding non-convex shape clusters. Moveover, notice that both algorithms can discover interesting regions inside other regions – red-coded regions (high-high or low-low) are successfully identified inside the blue-coded regions (low-high or high-low). It thus makes the hotspots even "hotter" when excluding inside regions from an outside region.

Scientific Contributions. Although the global correlation between the shallow ice and deep ice variables is only -0.14434 — suggesting the absence of a global linear relationship — our region discovery framework has found a number of local regions where extreme values of both variables co-locate. Our results indicate that there are several regions on Mars that show a strong anti-collocation between shallow and deep ice (in blue), but there are only few regions on Mars where shallow and deep ground ice co-locate (in red). This suggests that shallow ice and deep ice have been deposited at different geological times on Mars. These



(a) 99th to 60th percentile sorted by the value of z



	SPAM		SCMRG		SCDE		MOSAIC	
1	0.56	93	1.34	523	0.50	1258	0.57	4155
2	0.30	77	0.94	422	0.22	1145	0.00	2139
3	2.52	73	1.66	407	0.58	903	**1.38**	**2096**
4	0.18	72	0.70	367	0.53	784	0.00	2056
5	0.54	71	0.60	335	0.71	656	0.00	1491
6	0.66	70	0.98	315	0.21	571	0.47	1452
7	1.00	61	0.39	313	0.05	563	0.31	1174
8	1.81	59	0.62	282	0.22	463	0.45	1172
9	0.81	58	0.15	277	0.05	447	1.36	1143
10	0.04	57	0.30	262	0.16	435	0.64	1093

outliers

(b) z values of the top 10 regions sorted by region size

Fig. 5. Search capability evaluation

places need to be further studied by the domain experts to find what particular set of geological circumstances led to their existence.

4 Conclusion

This paper presents a novel region discovery framework for identifying the feature-based hotspots in spatial datasets. We have evaluated the framework with a real-world case study of spatial distribution of ground ice on Mar. Empirical statistical evaluation was developed to compare the different clustering solutions for their effectiveness in locating hotspots. The results reveal that the density-based SCDE algorithm outperforms other algorithms inasmuch as it discovers more regions with higher interestingness, the grid-based SCMRG algorithm can provide acceptable solutions within limited time, while the agglomerative MOSAIC clustering algorithm performs best on larger hotspots of arbitrary shape. Furthermore, our region discovery algorithms have identified several interesting places on Mars that will be further studied in the application domain.

Acknowledgments

The work is supported in part by the National Science Foundation under Grant IIS-0430208. A portion of this research was conducted at the Lunar and Planetary Institute, which is operated by the USRA under contract CAN-NCC5-679 with NASA.

References

1. Wang, W., Yang, J., Muntz, R.R.: STING: A statistical information grid approach to spatial data mining. In: 23rd Intl. Conf. on Very Large Data Bases (1997)
2. Koperski, K., Han, J.: Discovery of spatial association rules in geographic information databases. In: Egenhofer, M.J., Herring, J.R. (eds.) Procs. of the 4th Intl. Symp. Advances in Spatial Databases, vol. 951, 6–9, pp. 47–66 (1995)
3. Shekhar, S., Huang, Y.: Discovering spatial co-location patterns: A summary of results. In: Jensen, C.S., Schneider, M., Seeger, B., Tsotras, V.J. (eds.) SSTD 2001. LNCS, vol. 2121, Springer, Heidelberg (2001)
4. Eick, C.F., Vaezian, B., Jiang, D., Wang, J.: Discovering of interesting regions in spatial data sets using supervised clustering. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, Springer, Heidelberg (2006)
5. Choo, J., Jiamthapthaksin, R., Sheng Chen, C., Celepcikay, O.U., Giusti, C., Eick, C.F.: MOSAIC: A proximity graph approach for agglomerative clustering. In: The 9th Intl. Conf. on Data Warehousing and Knowledge Discovery (2007)
6. Brimicombe, A.J.: Cluster detection in point event data having tendency towards spatially repetitive events. In: The 8th Intl. Conf. on GeoComputation (2005)
7. Tay, S.C., Hsu, W., Lim, K.H.: Spatial data mining: Clustering of hot spots and pattern recognition. In: The Intl. Geoscience & Remote Sensing Symposium (2003)
8. Kulldorff, M.: Prospective time periodic geographical disease surveillance using a scan statistic. Journal Of The Royal Statistical Society Series A 164, 61–72 (2001)
9. Silverman, B.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, Boca Raton (1986)
10. Karypis, G., Han, E.H.S., Kumar, V.: Chameleon: Hierarchical clustering using dynamic modeling. IEEE Computer 32(8), 68–75 (1999)
11. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Chichester (1990)
12. Gabriel, K.R., Sokal, R.R.: A new statistical approach to geographic variation analysis. Systematic Zoology 18, 259–278 (1969)
13. Jiang, D., Eick, C.F., Chen, C.: On supervised density estimation techniques and their application to clustering. In: Procs. of the 15th ACM Intl. Symposium on Advances in Geographic Information Systems (2007)
14. Feldman, W.C.: Global distribution of near-surface hydrogen on mars. J. Geophys. Res. 109, E09006 (2004)
15. Barlow, N.G.: Crater size-distribution and a revised martian relative chronology. Icarus 75(20), 285–305 (1988)
16. Data Mining and Machine Learning Group, University of Houston: CougarSquared Data Mining and Machine Learning Framework (2007), `https://cougarsquared.dev.java.net/`