Investigation into Alternative Representations for Intelligent Geoinformation Systems in Data-rich Environments

David Breitkreutz and Ickjai Lee
School of Business, Cairns Campus, Smithfield, Queensland 4870, Australia
Email: {David.Breitkreutz, Ickjai.Lee}@jcu.edu.au

Abstract—Whilst a geographical information system may use many different representations of areal data, the vector format is the most commonly utilized structure. Although some researchers consider it efficient, others comment on its drawbacks. The potential exists for the implementation of another format with improved attributes and capabilities. This paper presents a method of creating a hybrid Voronoi approach to areal representation. This representation builds the positive qualities of the vectorized model into the Voronoi diagram to improve upon functionality within a geoinformation environment. This alternative representation aims to offer notable performance improvements whilst maintaining comparable accuracy to its original counterpart. A number of important applications are summarized, as are potential extensions of this work.

Index Terms—Representation, Voronoi diagram, Geoinformation

I. INTRODUCTION

Representing information within aggregated areal regions is commonplace within both geography and Geographic Information Systems (GIS). This concept of areal representation is extremely popular as it enables sizable amounts of information about a region to be summarized in different manners, allowing analysis of these areas to be tailored specifically to the application. This is the reason many organizations use this methodology to maintain their ever-increase stores of data [1]. Not only does areal representation provide an effective modeling environment for generalized regional data, but also inherently ensure that individual data elements are sufficiently obfuscated such that no single object can be identified.

This feature is extremely useful in circumstances where the privacy of personal details is paramount. One of the most significant benefits of areal representation is that different resolutions of summation can demonstrate different information easily. Consider the example of gathering census data from a number of households in a country. Summarizing information from a number of neighboring homes will provide a model of the demographics of a street. Performing the same collation of street summaries would yield details about a suburb. The same procedure follows for the encompassing town or city, or state or province. If individual data points were considered separately, calculations would be unnecessarily complicated, and each data point would be needed in each query. Thus, utilizing the aggregated areal approach represents improved memory usage, and less processing time required for calculations.

Many works have discussed methods of how to improve operations that execute upon areal data sets [2]-[5], but none approach the problem of performance by attempting to remold the underlying data structure. This paper presents an alternative structure which utilizes the Voronoi diagram in order to re-represent the original polygon structure more effectively, and to allow it to work better in data-rich environments. The issue that this paper will address is that of efficiency of the standard areal representation. The term standard refers to the vectorized areal format which uses polygons as its geometric primitive; this data structure is discussed in detail in Section II. The issue of efficiency concerns both the time complexity required to process varying operations upon the data set, and the amount of memory required to do so. Compared to the standard structure, the Voronoi diagram, the dual graph of the more widely known Delaunay triangulation [6], represents regions differently - it creates regions that are closest to specific points, given their relationship with other surrounding points. This is a beneficial quality as it means that entire regions can be created from a small number of generator points. In addition, the inherent creation of a Delaunay triangulation from a Voronoi diagram brings with it an automatic topology, a structure that would otherwise take considerable computational time if calculated separately. Some researchers have highlighted the potential of further integrating Voronoi diagrams within GIS [7]-[11], given that it is considered one of the most fundamental constructs defined by a discrete set of points [6]. However, whilst this type of diagram commonly works with other data formats [12], it is not typically used to represent regional data itself.

The alternative areal representation proposed in this paper aims to decompose the original polygonal data set such that it can be represented reasonably accurately within a Voronoi diagram. This new format is thereby referred to as a hybrid approach, as it considers the original vector regions, and utilizes the Voronoi and Delaunay constructs to maintain the resulting optimized data set. This paper presents an algorithm for converting an existing polygonal data set into the suggested hybrid...
Voronoi format. A number of different flavors of hybrid structure are also discussed, each utilizing a different method of polygon decomposition or reduction in order to determine the most suitable. A number of operations typically used within GIS compare the performance of the original structure to that of the hybrid solutions.

As a benchmark, the tests also examine the performance of the resulting Centroid-based Voronoi Tessellation (CVT), in which the centroid of each polygon is selected as the representative point for the Voronoi diagram; this is detailed further in Section III-B.

From the results that have been found, this hybrid areal representation appears to offer significant performance benefits for operations typically used within GIS. The results show the hybrid typically outperforms the original vector representation, and is comparable in terms of accuracy. This paper demonstrates the results of this experimentation in detail, and describes the circumstances to which this new format is most applicable.

II. RELATED WORK

The standard, vectorized areal data structure offers a well-rounded approach to the management of regional data [5]. Its ability to maintain highly detailed and accurate information within a typically small file size leads to it being far more useful than other formats when displaying areal data with definite boundaries. Because of these benefits, this format finds itself being the most commonly used structure for managing areal data consisting of straight-edged regions. Whilst some researchers [13], [14] describe the benefits of this structure, others have provide critical viewpoints to this format [14], [15]. One such example is that the vector representation imposes clear-cut boundaries on areal regions. Thus, there exists the possibility of an introduction of a degree of inaccuracy, which may affect data when precision is critical. These drawbacks affect the usage of the standard areal representation, and its overall performance.

The most commonly utilized alternative to the standard data structure is that of the raster; considered by cartographers to be a tessellation of square shaped objects [5]. In a rasterization, data being recorded is continuous, and accordingly, all elements of the tessellation require a value. Raster representations are most useful when non-discrete features - such as elevation, temperature, and ground covering types - need to be represented [14]. Due to geometrical requirements [5], there exist three types of regular rasterized tessellations: equilateral triangles, squares, and regular hexagons. Although commonly used as backdrop maps behind other layers of data within GIS, the raster model's usage as an areal data representation is notably limited. Many previous works [2], [5], [14] have discussed limitations pertaining to decreased accuracy, lack of topological support, in terms of regional relationships and connectivity, and its inability to efficiently use data storage space. Whilst the raster format excels in some applications, such as digital imaging, it cannot easily perform analytical GIS queries, such as the point-in-polygon or nearest neighbor operations. Comparatively, the vector-based approach inherently offers this ability, given its geometrical primitive is the polygon. At the same time, the standard vector approach offers far superior storage efficiency than that of a rasterization.

However, the existing literature relating to this topic loosely follows the same paradigm being investigated. Some works investigate the operations that will be performed on a data set, such as methods of population approximation [16], yet these are concerned with unrealistic environments, being free of both processing and time constraints. Whilst such algorithms may give accurate results, time- and mission-critical applications rely upon data structures and operations that can be executed quickly; this is the focus of this paper. Due to their complexities, these structures become inefficient as data sets grow. Because of this, they demonstrate that the original data structure is far less useful in data-rich environments.

Another issue with the existing literature relates specifically to the proposed hybrid solution: this notion of utilizing the Voronoi diagram, a non-geometric structure, to act within GIS as the data structure itself. Whilst some have discussed the issues with decomposing a Voronoi diagram into a set of Voronoi polygons [4] and others about the process, none has gone so far as to contextualize these polygons within any such geographic system. The research at hand, however, aims to investigate this potential and determine if a suitable alternative areal representation exists, and if so, how it can be best calculated and implemented.

Fig. 1 illustrates different areal representations of urban suburbs of Townsville, Queensland in Australia. In its original form, a satellite areal image as shown in Fig. 1(a) offers very little in the way of metadata about the represented areas. It is captured by the Landsat satellite in 2003 [17]. Although this image is able to depict the landscape and its physical features in detail, logical and political divisions are notably absent. Thus, in order to display and manage this extra information, another type data representation is required; the most common of which is the vectorized structure, whereby areal information is represented using a set of straight-edged, polygonal objects. In comparison to the original image, Fig. 1(b) is able to depict Townsville’s suburbs, overlaid in this instance for ease of correlation. Vectorized data most frequently records and reproduces ownership divisions of land parcels, postal code areas, and similarly formed areas. It is suited to these usages because the well-defined nature of such shapes requires the allocation of edges and vertices within a given space. Because the vector structure utilizes a continuous, rather than discrete, coordinate space, the construction and location of such regions is straightforward. Fig. 1(c) and Fig. 1(d) show alternative representations: raster representation and Voronoi representation, respectively.

III. VORONOI AREAL REPRESENTATION
A. Voronoi Diagram

Voronoi diagrams, also known as Thiessen polygons or Voronoi tessellations, are a way of representing areas closest to specific points, given their relationship with other surrounding points. The polygons within a Voronoi tessellation are termed proximal polygons, as they are the areas of closest proximity to a given point within a set $P$ of points [5]. Each tile of the polygon structure covers the region closest to their generator point, as is visible in Fig. 2.

As previously highlighted, the Voronoi diagram is the dual graph of the Delaunay triangulation, a graph that is concerned with constructing connections from each point to its nearest neighbors. As is evident in Fig. 2(a), Voronoi regions sharing a common boundary will have an edge connecting their generator points within the respective Delaunay triangulation; the dual nature of these graphs is thereby evident. This triangulation inherently produces triangle shapes that are “as nearly equilateral as possible” [5], and the resultant connections between points are paths to their closest neighbors. Refer to [6] for more details.

B. Voronoi Areal Representation

To provide the best approximation of a vectorized data structure, the suggested Voronoi areal representation utilizes the centroid of each original polygon as a representative point. This new areal representation

Figure 1. Different Areal Representations of Townsville, Queensland, Australia: (a) Lansat satellite image; (b) Vector representation; (c) Rasterized representation; (d) Voronoi representation.

Figure 2. Examples of underlying structures: (a) The Voronoi diagram of 7 points; (b) The Voronoi diagram and its dual Delaunay triangulation.
thereby implements a CVT to maintain the decomposition of its associated vectorized data structure. The difference between these two structures can be seen in Fig. 3, an example of the suburbs of the city of Townsville, Queensland, Australia, and the CVT is shown in Fig. 3(b). The procedure used to convert the vector data set to the CVT is described in Algorithm 1. It should be noted that insertion of a single point into a Voronoi diagram takes \( O(\log n) \), where \( n \) is the number of points present. This algorithm requires that each polygon be simple and will take \( O(vn \log n) \) time complexity in the worst case, where \( v \) is the total number of vertices of all polygons and \( n \) is the number of polygons.

### Algorithm 1 Decomposition of vector data set to Voronoi diagram

1) For all polygons in given set of data regions:
   a) Calculate centroid of polygon;
   b) Insert centroid into Voronoi diagram;
   c) Associate new Voronoi node with data from original polygon;
2) Remote original data set and node associations.

Whilst the algorithm can be executed upon any arbitrary set of simple polygons, the closest replication of the original structure is seen when polygons are convex and have few vertices. Hence, this areal representation will have difficulty matching the original shapes, given it uses only the centroids of sets of points. The procedure presented here is akin to that of compressing an image; the result resembles that of the original, but some loss of accuracy is experienced. The process results a reduction storage space needed, and thus improving speed with less data present. The forthcoming section details the notion of the hybrid approximation is presented, an alternative data structure that attempts to overcome this issue of accuracy.

### C. Hybrid Voronoi Areal Representation

In order to improve upon the Voronoi areal representation presented previously, a type of polygon partitioning is applied to the source polygon prior to being reduced. The time complexity of this algorithm can vary greatly, depending upon which partitioning technique is used. The main method this research focuses upon is a \( y \)-monotone sweep-line algorithm, given its speed and results compared to other techniques investigated. For further details upon this technique, the interested reader is referred to Berg et al. [18]. Therefore, the worst-case time complexity for a \( y \)-monotone hybrid representation is \( O(v^2p \log v \log p) \) time complexity, where
Algorithm 2 Decomposition of vector data set to hybrid Voronoi diagram

1) Create Voronoi diagram and Voronoi node grouping.  
2) For all polygons in a set of regions;  
   a) Execute given portioning technique;  
   b) Create grouping for hybrid nodes;  
   c) For all polygon partitions;  
      i. Calculate centroid of partition;  
      ii. Insert centroid into hybrid Voronoi diagram and current grouping.  
   d) Insert grouping into original node group container;  
   e) Associate group with data from original polygon.  
3) Remove original data set and node associations.

This procedure is evidently more complicated than the CVT representation in terms of time complexity, this extra process is required to better approximate the shape of each given polygon. This algorithm is more resilient to arbitrarily shaped polygons, including those that are concave and of irregular shape, than the centroid-based approached mentioned previously. The CVT method works effectively only when very simplistic, concave polygons with few vertices are present, compared to this hybrid approach that handles larger and more complicated regions. As a comparison, this paper also presents a randomized hybrid Voronoi model, whereby the representative points of a region are selected using a pseudo-random number generator. An example of a five-point randomization is demonstrated in Fig. 4(b). Irrespective of the precise locations of these chosen points, the results in the forthcoming Section IV demonstrates the usefulness of using several points in a region, rather than its centroid alone. This algorithm differs slightly in that partitioning is not required, and only points need to be generated for each original polygon. The resulting time complexity is thereby greatly reduced to become $O(cnlogcn)$, where $n$ is the number of original polygons and $c$ is the selected number of points required.

IV. EXPERIMENTS

A. Experimental Setup

To determine the viability of the newly suggested areal representations, each structure needs to be compared against a number of different operations, and the results compared for speed and accuracy. In order to execute this testing as objectively as possible, a test environment has been designed, being written in C++ and programmed using both the Library of Efficient Data types and Algorithms (LEDAl) and the Computational Geometry Algorithms Library (CGAL). The operations selected for analysis are two of the most commonly utilized within GIS: point-in-polygon, also known as point intersection, and nearest neighbors. The point-in-polygon operation involves testing which region a given point is associated with. This operation is extremely important because it allows the discovery of how objects and areas interact, and reveals meaning by demonstrating this relationship. Consider the previously-mentioned application of a census. If a query was being performed on a given household, it would be necessary to determine which suburb or city the data is part of. The nearest neighbor operation, however, is concerned with which regions or objects surround another. Again, this operation is of the utmost importance with GIS because it demonstrates associations between centers of information, and also allows deductions to quickly occur based upon this information. A practical usage of this is within emergency management, such as the occurrence of a forest fire. When such an event takes place, the relevant authorities would need to evacuate areas nearest to the fire as a way of preserving human life. Without the ability to directly intersect a point with an area, or know which regions border others, the system itself would become far less useful. Also, the processing overhead to otherwise determine the same results would be significant, in addition it may jeopardize the application of the system is being used for.

In order to accurately compare the same operation against different data structures, results produced were returned in a common format, able to be checked against the original structure. The process bench-marked alternative formats against the original structure. In terms of performance, the number of CPU cycles, the smallest unit of time measurable on a computer processor, were recorded for each operation. Whilst these units cannot be compared across computer hardware, they clearly illustrate performance experienced. Finally, with respect to testing the accuracy of each areal representation, the results of each operation will be compared to that of the same function executed upon the original polygons. For the point-in-polygon operation, the result will either be correct or incorrect, and the nearest neighbor comparison will yield a percentage based upon correct responses, and those expected to be returned.

B. Data Preprocessing

A method of improving the original data representation is by reducing the complexity of the source data set. Whilst a trivial solution to this issue, the typical reasoning to why data sets are difficult to manage is due to the number of vertices and edges that form the boundaries of the regions described. The number of vertices present within a data set severely impacts upon the number of operations that could be performed to complete polygon partitioning.

In order to demonstrate the effectiveness and accuracy of the each given areal representation, two different data structures are utilized: the real-world data set, as demonstrated previously, and a simplified version. The experimental tests compare these two different sized data structures. A comparison of both is shown in Fig. 3(a)
and Fig. 5(a). Both the simplified and real-world data sets have 26 regions, but rather than consisting of 6645 vertices as the real-world set does, the simplified version contains only 162. As evidenced in Fig. 5(a), the regions contained within this structure are far less complicated than those in the real-world data set shown in Fig. 3(a). This reduction in complexity indicates that all operations performed upon this data structure stand to benefit from an improvement in performance, and likewise, for any derivative alternative areal representations. What the development of this reduced data set aims to achieve, however, is an improvement to performance experience from the usage of these alternative representations.

C. Experimental Results

The experiments detailed in Section IV-A have been performed and results have been charted in Fig. 6 and 7. Both the simplified and real-world data sets have 26 regions. For the original vectorized structure, the simplified data set consists of 162 points, whilst the more complex set has 6645 vertices. For the CVT structure and randomized hybrid, both data sets have the same number of points: 26 and 104 respectively, given that they use constants. Finally, the simplified $y$-monotone hybrid structure has 33 data points whilst its real-world counterpart has 319.

The point-in-polygon operation results demonstrate that for the simplified data set, whilst the performance of each structure was comparable, the original data structure performed, on average, faster than both hybrid solutions, but slower than the CVT. In comparison, the CVT performed 4.18% faster, whilst the $y$-monotone and randomized hybrids performed 38.96% and 115.45% slower respectively. In this situation, the accuracy of each alternative representation was approximately 65% (CVT: 65.46%, Hybrid ($y$-monotone): 65.06%, Hybrid...
(randomized): 66.67%) of the result of the vectorized structure. In contrast, for the real-world data set, the performance difference is evident, given the considerable increase in data points present for the vectorized polygons. In this situation, the original data set was clearly out-performed, with the CVT being 95.30% faster, y-monotone hybrid being 81.98% faster, and randomized hybrid 81.97% quicker. These increases in speed do come without significant decrease in accuracy, with the resulting accuracy being approximately 65% (CVT: 65.86%, Hybrid (y-monotone): 64.65%, Hybrid (randomized): 66.67%).

![Figure 7. Nearest neighbour performance: (a) Simplified data set; (b) Real data set.](image)

Figure 7. Nearest neighbour performance: (a) Simplified data set; (b) Real data set.

The nearest neighbor operation results show that there is a notable improvement in using any of the alternative areal representations, for both types of data sets. For the simplified version, the performance benefit over the original polygon set was 86.56% for the CVT, 79.61% for the y-monotone hybrid, and 59.49% for the randomized hybrid. Again, this improvement is not without a cost of accuracy, with the CVT producing results of 50.68% accuracy, 52.97% for the y-monotone hybrid, and 48.77% for the randomized hybrid. The real-world data set, however, shows even greater performance increases when using an alternative format (CVT: 99.91% faster, Hybrid (y-monotone): 98.91% faster, Hybrid (randomized): 99.77% faster). However, these significant improvements in speed result in reduction in accuracy, such that the CVT format, y-monotone hybrid, and randomized hybrid give 46.90%, 43.75%, and 46.35% accuracy respectively.

These results confirm the original analysis of the data structures, and show that whilst accurate, the vectorized data format is typically slower than the alternatives presented, with minor exceptions. These results from both point-in-polygon and nearest neighbor operations show that as the size of the original data set increases, the performance benefit to be gained from using an alternative also increases.

Point-in-polygon operations on all data structures took \(O(n)\) time, where \(n\) is the total number of points, because a given input point needs to be compared to each of the vertices or generator points present. Thus, the performance of data structure is affected by how many points it uses. For the simple data set, the original structure had a smaller value of \(n\), and thus outperformed the Voronoi-based alternatives. However, for the complex data set, the vectorized structure has a large value of \(n\), and thus the benefit of using an alternative was evident.

The nearest neighbor operation on the vectorized data structure requires \(O(k(n-k))\) time, where \(k\) is the number of points in the selected polygon, and \(n\) is the total number of points. This is because each point within a given polygon needs to be compared against all others within the data set. Voronoi-based alternatives, however, only require \(O(1)\) (constant) time, as the Delaunay triangulation has already been calculated, and neighbors are already known.

With respect to accuracy, it appears from the testing performed that the increase in the number of data points present does not have a discernible impact upon the results. Whilst the decreases in accuracy from the original representation are notable, their true impact will entirely depend upon the application, and the emphasis the application places upon performance instead.

V. SUMMARY AND FUTURE WORK

The development of improved and more efficient representations of data is necessary to further the abilities of computerized geography. Many works have served to highlight this necessity for improvements, given the limitations imposed by the data structures that are commonly used, but none of which have served to investigate the redevelopment of the data structure itself. Improvements to the original vectorized structure are necessary as data sets continually grow in size, and overwhelm storage systems.

This paper has presented the potential for an alternative Voronoi areal representation to be utilized within GIS. It has also demonstrated that such an implementation can bring notable performance benefits to commonly used operations and functionality within the same types of systems. Whilst the usage of such a structure may incur a reduction in accuracy, for many applications, the loss is acceptable for the valuable increases in processing speed. This results in less time required waiting for results, and more time spent utilizing them in situations where consequences of inaction may be dire.

As improvements to geometrical processing occur, the possibility for the hybrid areal representation to provide even greater performance than described here also exists, and could potentially be investigated in extensions of this work.
REFERENCES


David Breitkreutz is currently a research associate at e-research center in James Cook University. He received his BSc with Honours (First class) in School of Math, Physics & IT in James Cook University, Australia in 2008. His current research areas includes business intelligence, data mining, and Voronoi diagrams.

Ickjai Lee obtained his Ph.D. in School of Electrical Engineering and computer Science from the University of Newcastle, Australia in 2002, his Honours degree in Spatial Science and graduate diploma in Computing from Curtin University of Technology, Australia, and his BSc in Department of Computer Science from the Pusan National University, Korea in 1993.

After his Ph.D., he worked as a postdoctoral research fellow at Newcastle Business School in the University of Newcastle in the areas of artificial intelligence, robot soccer and knowledge representation. Currently, he is a senior lecturer at School of Business in James Cook University, Australia. He has published around 60 articles in refereed journals and conferences. His research areas include geoinformatics, spatial data mining, Voronoi diagram, mobile-intelligence and artificial intelligence.

Dr. Lee is a lifetime member of ACM since 2008 and a member of IEEE since 2004. He is a member of editorial review board of Journal of Computer Information Systems, and is a member of program committee of several conferences. He has regularly reviewed articles for several journals from geoinformatics and information system areas.