

UNIVERSITY OF COIMBRA

M.Sc. THESIS

Visualisation and Analysis of Geographic Information: Algorithms and Data Structures

Author:

João VALENÇA

Supervisor:

Prof. Dr. Luís PAQUETE

Eng. PEDRO REINO

Department of Informatics Engineering

May 2015

Abstract

In recent years, Geographic Information Systems (GIS) have witnessed a large increase in data availability. There is a need to process a large amount of data before it can be managed and analysed. This project aims to develop a GIS application operating through a Web platform in order to allow for a low cost and simplified integration, management and manipulation of georeferenced information. Special emphasis is given to the implementation of efficient clustering algorithms for finding a representative set of points in a map, which can be recast as a *k-center* problem. The approaches covered in this report include exact algorithms for finding minimum coverage subsets.

Keywords: Geographic Clustering, k-Center, Coverage, Branch-and-Bound, Delaunay Triangulations

Resumo

Nos últimos anos, os Sistemas de Informação Geográfica (GIS) têm assistido a um grande aumento na quantidade de dados disponíveis. De facto, existe uma necessidade de encontrar uma maneira eficiente de processar grandes quantidades de dados para que tanta informação possa ser facilmente gerida e analisada. Este projeto visa desenvolver uma aplicação GIS para uma plataforma Web, de modo a obter uma integração simples e de baixo custo que manipule e analise dados georeferenciados. Uma ênfase especial é dada à implementação de algoritmos para encontrar um conjunto representativo de pontos num mapa, que pode ser formulado como um problema de *k-center*. As abordagens descritas neste relatório incluem algoritmos exactos para encontrar o subconjunto ótimo.

Palavras-chave: Clustering Geográfico, k-Center, Cobertura, Branch-and-Bound, Triangulações Delaunay

Contents

1	Introduction	1
2	Concepts, Definitions and Notation	3
2.1	Definitions of Coverage	3
2.2	Algorithmic Concepts	5
2.2.1	Branch-and-Bound	5
2.2.2	Approximation Algorithms	5
2.3	Geometric Concepts and Geometric Structures	6
2.3.1	Nearest Neighbour Search and Point Location	6
2.3.2	k -Dimensional Trees	6
2.3.3	Median of Medians	7
2.3.4	Voronoi Diagrams	9
2.3.5	Delaunay Triangulations	10
2.3.6	Hilbert Curves	15
3	State of the Art	17
3.1	Integer Linear Programming Formulation	17
3.2	Local Search Approach	18
4	Optimal Minimum Coverage Algorithms	20
4.1	Naïve Branch-and-Bound	20
4.1.1	Branching	20
4.1.2	Bounds	22
4.2	Delaunay Assisted Branch-and-Bound	22
4.3	Algorithm Comparison	24
4.3.1	Time Complexity	24
4.3.2	Experimental Results	25
5	Future Work	29
5.1	Integration in a Visualisation Framework	29
5.2	Heuristic Approaches	29

5.3	Approximation Algorithms	30
-----	------------------------------------	----

List of Figures

1.1	Example of a Representative Set	2
2.1	Different assignments for the same set of points	4
2.2	Example of a Voronoi Diagram	9
2.3	Example of a Delaunay Triangulation	10
2.4	Overlap of a Voronoi Diagram and its Delaunay Triangulation	11
2.5	Illustration of the Half-Edge Structure	13
2.6	Example of the Greedy Routing Algorithm	14
2.7	Illustration of the Walking Algorithm	15
2.8	First Five Orders of Hilbert Curve Approximation	16
4.1	CPU-time for different values of K with varying values of N	26
4.2	CPU-time for different values of N with varying K	27

List of Tables

4.1	Time complexities for the various operations in a uniformly distributed point set	25
4.2	Machine specifications	25

Chapter 1

Introduction

In recent years, there has been a large increase in both the quantity and availability of geographic data. This new availability of such large quantities of data has prompted a similar rise on the number of applications to capture, store, manipulate and analyse this data. A lot of these applications share the need to visualise the geographic information in such a way that it can be easily understood by a human. This is usually done by displaying points of interest on a map so that their relative position or direction can be easily interpreted without much thought by the user.

One obstacle when representing large amounts of geographic data is that the sheer volume of points to display can be overwhelming for a human, as well as computationally intensive to render for a machine. As such, there is a need to develop and implement a viable way to reliably calculate and display a subset of geographic points, whilst keeping a degree of representability of the larger set, so that as little information as possible is absent when the representative subset is shown.

The purpose of this project is to research and develop a real-time algorithm that can analyse geographic data provided by a geographic information system (GIS) infrastructure developed and maintained at Smartgeo. More precisely, the developed algorithm will have to be able to aggregate and select geographic points according to a given a set of criteria. Figure [1.1](#) shows an example of a representative subset of an original, larger set.

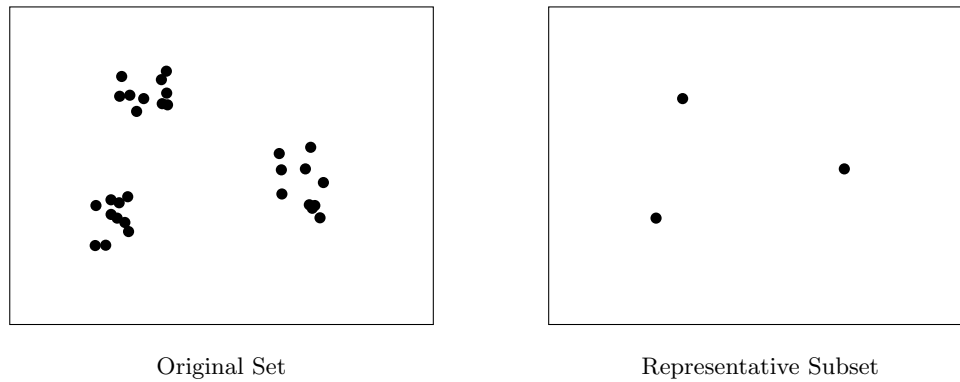


FIGURE 1.1: Example of a Representative Set

This thesis aims to research, develop, and analyse different algorithms to choose a representative subset of geographic points, whilst being able to dynamically change that set of points via zooming or panning over a geographic region containing a large amount of geographical data. In case the optimal solution algorithms prove to be too slow, heuristic approaches will be implemented. Heuristic algorithms will have their solution quality and speed benchmarked against implicit enumeration algorithms. The algorithm that is deemed the best will be implemented in the web framework via the *WFS* and *WMS* web mapping standards.

This report is organized as follows: Chapter 2 - [Concepts, Definitions and Notation](#) defines the base theoretical concepts, such as a notion of representativeness, as well as some useful structures used in the algorithms. Chapter 3 - [State of the Art](#) analyses previous related work. Chapter 4 - [Optimal Minimum Coverage Algorithms](#) describes the implicit enumeration algorithms implemented so far, as well as an analysis on their time and space complexities. Chapter 5 - [Future Work](#) describes the direction of the heuristic algorithms to be developed in the second half of the thesis.

Chapter 2

Concepts, Definitions and Notation

This chapter gives an introduction to the base concepts used further in this report. The chapter starts by establishing the formal definition of the problem at hand. Then it proceeds to detail the algorithmic and geometric concepts to be used in the different approaches described in the following chapters.

2.1 Definitions of Coverage

Representativeness consists of finding a subset of points in a larger set. The subset chosen should be able to keep some specified properties of the original set, such as density, or general distribution. As such there can be many ways to define representativeness. For the purposes of this thesis, we will use the definition of *coverage*. Given a set of points N in \mathbb{R}^2 , we must choose a subset, P , that best matches our definition of representativeness. The size of P , however, is constrained to a size k , which specifies how many points can be displayed in a section of a map.

For any point in N not in P , there must be a point in P that best represents it. In a geographic or geometric plane, this notion of representativeness may be defined by the distance, i.e. the point p that best represents q is the closest point closest to q , under some notion of distance. Although the points represent geographic locations, the metric that would measure their distance on the surface of the globe, the geodesic distance, will not be used in this thesis. The triangular inequality property does not apply to geodesic distances as a sphere (or an approximation of thereof) is not an Euclidean space, and it would add an unnecessary layer of complexity to computing the coverage.

Because of this, in this thesis, the coordinates of the points are the planar projection of the geographic coordinates to their location counterparts, as implemented by the WMS and WFS web mapping standards. Therefore, the Euclidean norm will be used as the spatial distance notion.

Finding the most representative set P in N will mean that every point in N will be assigned to the point in P closest to it. This definition of representativeness is referred to as *coverage* and the points in P are called *centroids*.

The coverage value of a given centroid is defined by the circle around that centroid with the radius defined by the distance between itself and the farthest non-centroid point assigned to it. The coverage value of a subset is determined by the highest coverage value of its points. It can thus be more formally described as:

$$\max_{n \in N} \min_{p \in P} \|p - n\| \quad (2.1)$$

where N is the initial set of points in \mathbb{R}^2 , P is the centroid subset and $\|\cdot\|$ is the Euclidean distance. The most representative subset, however, is the one with the minimum value of coverage. This means all points will be assigned to the closest centroid, minimising the coverage of all centroids and avoiding overlapping coverage areas whenever possible. We can then finally define our problem as the minimising the coverage:

$$\min_{\substack{P \subseteq N \\ |P|=k}} \max_{n \in N} \min_{p \in P} \|p - n\| \quad (2.2)$$

This is known in the field of optimisation as the *k-centre* problem, and is an example of a facility location problem [5]. Figure 2.1 shows two possible centroid assignments, each with its own value of coverage, d .

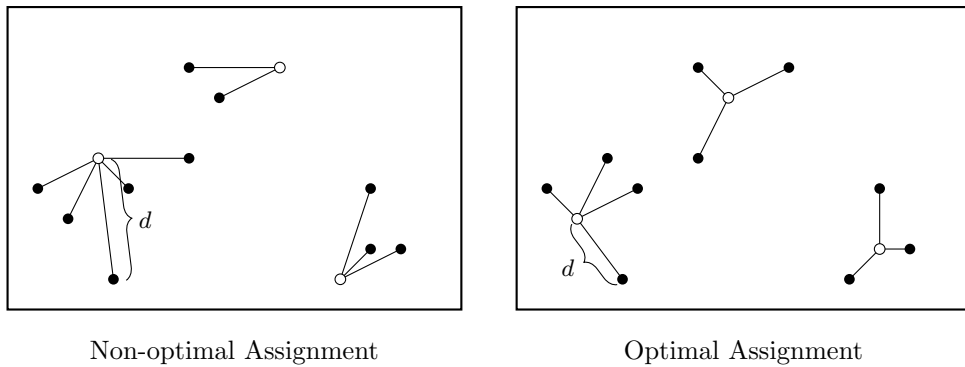


FIGURE 2.1: Different assignments for the same set of points

For the 1-dimensional case, the minimum coverage value can be calculated in polynomial time [14]. However, for any other number of dimensions it is a *NP-hard* problem, and cannot be solved in polynomial time [9].

2.2 Algorithmic Concepts

2.2.1 Branch-and-Bound

Minimising coverage, as shown, is a *NP-hard* combinatorial problem [9]. One possible way of solving the problem is to use implicit enumeration algorithms such as *Branch-and-Bound* algorithms. These algorithms solve the problems by recursively dividing the solution set in half, thus *branching* it into a rooted binary tree. At each step of the subdivision, it then calculates the upper and lower bounds for the best possible value for the space of solutions considered at that node. This step is called *bounding*. In the case of a minimisation problem, it would be the upper and lower bounds for minimum possible value for the objective function in the current node. These values are then compared with the best ones already calculated in other branches of the recursive tree, and updated if better.

The bounds can be used to *prune* the search tree. This can be done when the branch-and-bound algorithm arrives at a node where the lower bound is larger than the best calculated upper bound. At this point, no further search within the branch is required, as there is no solution in the current branch better than one that has already been calculated. In the case that the global upper and lower bound meet, the algorithm has arrived at the best possible solution, and no further computation must be done.

These algorithms are very common in the field of optimisation and can be very efficient, but their performance depends on the complexity and tightness of its bounds. Tighter bounds accelerate the process, but are usually slower to compute, so a compromise has to be made in order to obtain the fastest possible algorithm.

2.2.2 Approximation Algorithms

Optimal solution algorithms, even very optimized ones, are oftentimes still too inefficient to be used in any practical, time constrained application. One possible strategy to solve an *NP-hard* problem is to use an *approximation algorithm*.

Approximation algorithms do not compute the optimal solution to a given problem. Instead, for the sake of time efficiency, these algorithms are designed to compute a

solution that differs from the optimal by a given factor. For example, a 2-approximation algorithm for a minimising problem will not compute any solution that is more than double the optimal value for any given input. By compromising the quality of the solution, the algorithms can end within polynomial time.

2.3 Geometric Concepts and Geometric Structures

In the following, we explain some geometric concepts that are used in the following chapters, in order to simplify the explanation of more complex algorithms in further chapters.

2.3.1 Nearest Neighbour Search and Point Location

A common concept in computational geometry is point location. A point location algorithm finds the region on a plane that contains a given point p . Depending on the nature and shape of the regions, point location algorithms may differ in approach. In this thesis, most point location problems consist of finding the closest centroid to a given point, i.e. a nearest neighbour search algorithm.

Given a point p , a *nearest neighbour search* algorithm returns the closest point to p in a given set. Since we need to find the closest centroid to a given point in order to find the correct coverage value, this operation will be one of the most used, and so we need a fast and flexible way of determining which of the centroids is closest, in order to reduce computational overhead.

Point location algorithms direct the nearest neighbour search to smaller regions, bypassing any regions that are too distant from p , thus reducing the number of calculations necessary to get the proper point location. Common structures used for point location are *k-d trees* as described in Cambazard et al. [4]. A *k-d tree* partitions the space using a divide and conquer approach to define orthogonally aligned half-planes. This approach takes $\mathcal{O}(\log n)$ time to achieve point location queries. However, a *k-d tree* needs to be periodically updated in order to keep its efficiency and cannot be constructed or deconstructed incrementally without considering this overhead.

2.3.2 *k*-Dimensional Trees

A *k-dimensional tree*, or *k-d tree*, is a space partitioning structure used for point location and nearest neighbour queries. A *k-d tree* is a binary tree, which each of whose nodes

represents an axis-aligned hyper-rectangle. Each node specifies an axis and splits the set of points based on whether their coordinate along that axis is greater than or less than a particular value. The axis chosen to split each subgroup is chosen via a rotation system, which in the 2-dimensional space means that each level alternates between the vertical and horizontal axis.

During construction, the splitting point is chosen to be the point whose relevant coordinate best divides the group into two subgroups. The best possible choice for the splitting point at each level is the point which has the median of the relevant coordinate in the group. This ensures that each node divides the number of points in half for so that the resulting tree becomes balanced, and that each point is reachable by performing $\mathcal{O}(\log n)$ operations.

Construction

Constructing a k -d tree requires selecting a pivot, which divides the set into two groups: the points whose relevant coordinate is smaller than the pivot, the points whose relevant coordinate is larger than the pivot. The same process is then repeated recursively for each of the groups, alternating the relevant coordinate between both axes. Each recursive call fixes one pivot, which ideally will contain the median value of the relevant coordinate. The time complexity of the construction of a k -d tree relies on the pivot selection function. Calculating the median usually requires sorting a list of n and then picking the $n/2$ th element. Since sorting algorithms typically take $\mathcal{O}(n \log n)$, building a k -d tree would take $\mathcal{O}(n^2)$ time. To achieve better time complexity and performance the median of medians algorithm described in 2.3.3 can be used.

Even though the median of medians algorithm does not necessarily return the actual median, the query complexity in a k -d tree constructed using this still achieves the $\mathcal{O}(\log n)$. This occurs because the median of medians always outputs a value between the 30th and 70th percentile. This guarantees that at each level of the k -d tree the group of points covered by each of the children nodes is substantially smaller than the parent node by a constant factor, and there will never be a redundant node that covers the same set of points the parent does. At each level of a given query, each decision discards at least 30% of the points, maintaining the $\mathcal{O}(\log n)$ time complexity. [?]

2.3.3 Median of Medians

Efficiently constructing a balance k -d tree depends on an efficient method to pick the point that divides the hyper-rectangle in two. One way to reasonably quickly find a

value close to the median is to find the median of a sample. To ensure the quality of this sample, is to gather the medians of smaller subsets which can be quickly calculated. This algorithm is an example of a *selection algorithm* [?] and is known as the *median of medians* algorithm [?].

The median of medians algorithms works as follows. Any starting array S consisting of n arbitrary values is split into $n/5$ sub-arrays, each containing at most 5 elements (the last array might have less, depending on whether n is divisible by 5 or not). For each of the sub-arrays, the median can be calculated in constant time, since for 5 values it can be done in at most 6 comparisons, which for the whole array S takes $6n/5$ comparisons. After finding all the sub-arrays' medians and gathering them in a new array F , the algorithm then is called recursively for F until only one value M remains. M is then used to partition the input into two sub-groups: elements smaller than M and elements larger than M . The two subgroups are then concatenated in increasing order and with M in between them, and the algorithm is recursively called again for the group that contains the $n/2$ th point of the newly concatenated list. Whenever the list has less than a given number of elements, the median is calculated via brute-force, to avoid infinite recursion. This value will be the value returned by the initial recursive call of the function.

As stated above, this algorithm only returns a value close to the real median. Despite this, it can be proven that for any array S , the value M will always be between the 30th and the 70th percentiles. At each recursive stage, the values in F larger than M are discarded. This means that out of the $n/5$ values for any given vector, $n/10$ will be larger by definition, since M is picked as the median. For each value in F larger than M , there will also be two other values that are larger than M , since each value in F was chosen as a median out of 5 different values. This means that the number of values greater than M will be at most $3n/10$. Similarly, by a symmetric proof, there will also be $3n/10$ values in S smaller than M . This also means that the second recursive call will at worst have $7n/10$ elements, which is a constant fraction of the input. This property is essential in proving the linear complexity of the algorithm.

Analyzing the time complexity $T()$ of this algorithm requires analyzing separately both recursive calls of the algorithm. The first recursive call occurs in a list of size $n/5$, and takes $T(n/5)$ time. The second recursive call occurs in a list with $7n/10$ elements, which takes $T(7n/10)$. Finding the median for a group of 5 elements requires a constant number of comparisons. These comparisons can be arranged in such way that only 6 are necessary for a group of 5 elements. This means that the algorithm has a constant

factor of $6/5$ for calculating a median on its smallest division. $T(n)$ is then given by:

$$T(n) \leq 6n/5 + T(n/5) + T(7n/10) \quad (2.3)$$

If $T(n)$ has, in fact, linear time complexity, then there is a constant c such that:

$$\begin{aligned} T(n) &\leq 6n/5 + cn/5 + 7cn/10 \\ &\leq n(12/5 + 9c/10) \end{aligned} \quad (2.4)$$

If $T(n)$ is to be at most cn , so that the induction proof is valid, then it must be true that:

$$\begin{aligned} n(6/5 + 9c/10) &\leq cn \\ 6/5 + 9c/10 &\leq c \\ 6/5 &\leq c/10 \\ c &\leq 12 \end{aligned} \quad (2.5)$$

This proves that $T(n) \leq 12n$, or any larger constants than 12 multiplied by n comparisons.

2.3.4 Voronoi Diagrams

Voronoi diagrams [10] partition the space into regions, which are defined by the set of points in the space that are closest to a subset of those points. Definitions of distance and space may vary, but in our case we will consider the \mathbb{R}^2 plane and the Euclidean distance. Figure 2.2 shows a partitioning of a plane using a Voronoi Diagram for a set of points:

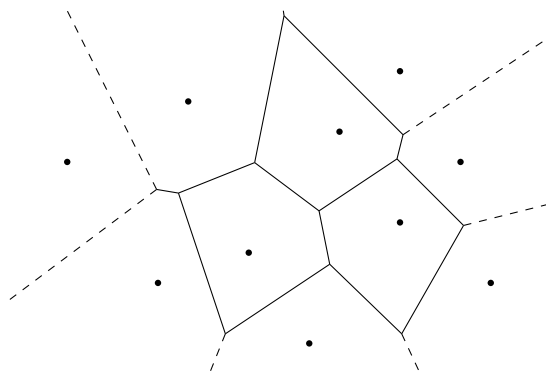


FIGURE 2.2: Example of a Voronoi Diagram

Dashed lines extend to infinity. Any new point inserted in this plane is contained in one of the cells, and its closest point is the one at the centre of the cell. Each edge is the perpendicular bisector between two neighbouring points, dividing the plane in two half planes, containing the set of points closest to each of them. The construction of Voronoi diagrams can be done incrementally, but in order to obtain fast query times, one needs to decompose the cells into simpler structures.

2.3.5 Delaunay Triangulations

Another useful structure for geometric algorithms is the Delaunay triangulation [10]. A Delaunay triangulation [6] is a special kind of triangulation with many useful properties. In an unconstrained Delaunay triangulation, each triangle's circumcircle contains no points other inside its circumference.

A Delaunay triangulation maximizes the minimum angle of its triangles, avoiding long or slender ones. The set of all its edges contains both the minimum-spanning tree and the convex hull. The Delaunay triangulation is unique for a set of points, except when it contains a subset that can be placed along the same circumference. Figure 2.3 shows the Delaunay triangulation of the same set of points used in Figure 2.2:

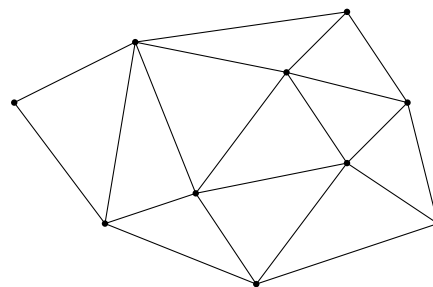


FIGURE 2.3: Example of a Delaunay Triangulation

More importantly, the Delaunay triangulation of a set of points is the dual graph of its Voronoi Diagram. The edges of the Voronoi diagram, are the line segments connecting the circumcentres of the Delaunay triangles. When overlapped, the duality becomes more obvious. Figure 2.4 shows the overlapping of the Voronoi diagram in Figure 2.2 and the Delaunay triangulation in Figure 2.3. The Delaunay edges, in black, connect the points at the centre of the Voronoi cells, with edges in blue, to their neighbours.

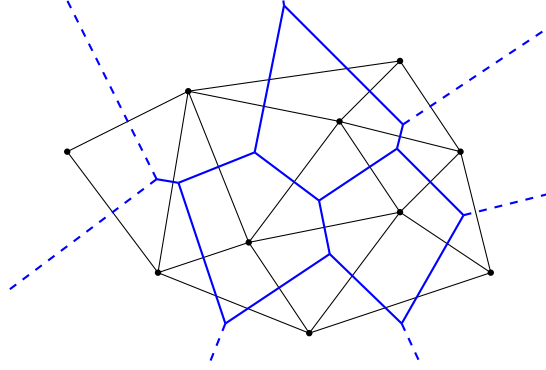


FIGURE 2.4: Overlap of a Voronoi Diagram and its Delaunay Triangulation

Unlike its counterpart, the Delaunay is much simpler to build incrementally. It is also easier to work with, whilst still providing most of the Voronoi diagram's properties, including the ability to calculate both point location and nearest neighbour searches.

Construction

There are many algorithms to construct a Delaunay triangulation. The particular conditions of our approach to the coverage problem impose some restrictions to the choice of the algorithm to use. Building a Delaunay triangulation can be done incrementally. Starting with a valid triangulation, points can be added, creating and updating existing triangles. An efficient way to do so is to use the Bowyer-Watson algorithm [3].

Starting with a valid Delaunay triangulation \mathcal{T} , we find the triangle t with vertices a, b and c that contains the vertex to insert v using a point location algorithm, such as the line walking algorithm described in the previous section. We then follow algorithm 1 for each vertex v to be included in the triangulation:

Algorithm 1 Bowyer-Watson Algorithm

```

1: function INSERTVERTEX( $v, a, b, c$ )
2:   DeleteTriangle( $a, b, c$ )
3:   DigCavity( $v, a, b$ )
4:   DigCavity( $v, b, c$ )
5:   DigCavity( $v, c, a$ )

1: function DIGCAVITY( $a, b, c$ )
2:    $d \leftarrow \text{Adjacent}(b, c)$ 
3:   if  $d \neq \emptyset$  then
4:     if inCircle( $a, b, c, d$ ) then
5:       DeleteTriangle( $w, v, x$ )
6:       DigCavity( $a, b, d$ )
7:       DigCavity( $a, d, c$ )
8:     else
9:       AddTriangle( $a, b, c$ )

```

The algorithm starts by removing the triangle t that contains the new vertex v , and recursively checks adjacent triangles whose circumcircle contains v using the *DigCavity* function. Any triangle that contains v in its circumcircle (calculated with the *InCircle* function), violates the Delaunay rule, and must also be deleted and have its sides recursively checked, until no adjacent triangles violate the Delaunay rule. Whenever the *DigCavity* function reaches a set of three points whose circumcircle does not contain v , it creates the triangle by creating counter-clockwise half-edges between those three points (*AddTriangle*). For inserting n points, this algorithm has an expected time complexity of $\mathcal{O}(n \log n)$ and is described in more detail by Shewchuk [12].

Deconstruction

Deconstructing a Delaunay triangulation usually consists of reversing the construction algorithm to remove points from the triangulation. However, as we will explain in a later chapter, in our case the deconstruction has to be incremental. Since the first point to be removed from the triangulation is necessarily the last one to be inserted, we can use a simpler approach. At each step of the construction, all created and removed edges and triangle from the triangulation can be stored in a LIFO structure, or a stack. When the last inserted point is to be removed, recreating the previous state of the triangulation is only a matter of rolling back and retrieving the information from the stack. This also

means no geometrical calculations have to be performed, and the old edges and triangles are quickly put back in place, with no new memory allocation needed.

Half-Edge Structure

A useful structure to use when building and managing triangulation meshes is the half-edge structure. The half-edge structure represents one orientation of each edge in the triangulation. This means that for each pair of points (p_i, p_j) connected in a triangulation \mathcal{T} , there are two directed half-edges: one represents the edge from p_i to p_j , and the other represents the opposite direction, connecting p_j to p_i . They both contain information about the triangle that they face, and thus, are part of. Triangles are defined by three half edges. All the half edges in the triangle share two of the vertices of the triangle, and are all sorted in a counter-clockwise order. Figure 2.5 further illustrates the concept of the half-edges.

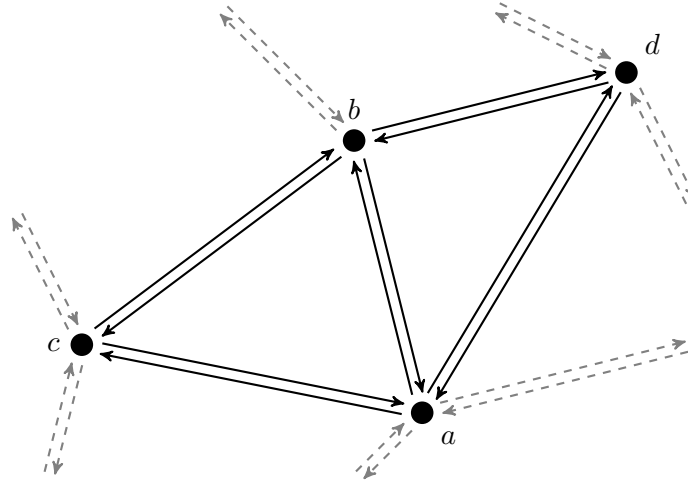


FIGURE 2.5: Illustration of the Half-Edge Structure

This structure makes it easier to store the changes to the triangulation at each step, since they contain the information about the triangles themselves. This means that only the half edges need to be stored in the stack (for construction and deconstruction) with no need to manage the triangles directly. The half-edge structure helps to obtain the triangulation neighbours for any vertex v , since it keeps all the edges starting at any given point easily accessible. All neighbours to any point v are the end points to the half-edges starting at v . This property is useful when efficiently implementing the greedy routing algorithm described in the following Section.

Greedy Routing

In order to quickly calculate the nearest neighbour to a point in a set, one can make use of the Delaunay triangulation with Greedy Routing [2]. Consider a triangulation \mathcal{T} . In order to find the closest vertex in \mathcal{T} to a new point p , start at an arbitrary vertex of \mathcal{T} , v , and find a neighbour u of v whose distance to p is smaller than the distance between p and v . Repeat the process for u and its neighbours. When a point w is reached such that no neighbours of w are closer to p than w is, the closest point to p in \mathcal{T} has been found. In the following, we show that the greedy routing algorithm is correct:

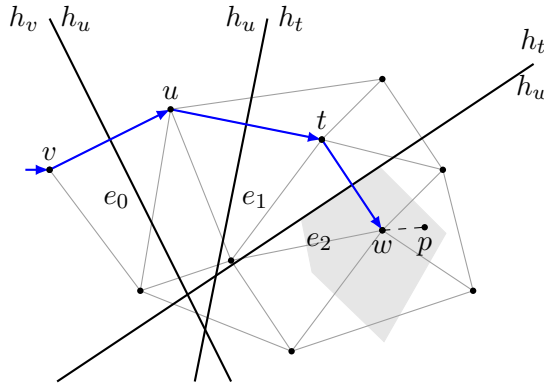


FIGURE 2.6: Example of the Greedy Routing Algorithm

In Figure 2.6, the search for the closest vertex to p starts at point v . From there, point u , which is closer to p than v is, is found. The step is repeated, following the blue path until point w is reached. Since no neighbour of w is closer to p than w is, then p must be within the Voronoi cell of point w (shaded light grey).

Theorem 2.1. [2] *There is no point set whose Delaunay triangulation defeats the greedy routing algorithm.*

Proof. For every vertex v in a triangulation \mathcal{T} , let the perpendicular bisector of the line segment defined by v and any neighbour u be called e if there is at least one neighbour of v , u closer to p than v is. The line e intersects the line segment (v, p) and divides the plane in two open half planes: h_v and h_u . Note that the half plane h_u contains p . Delaunay edges connect the Voronoi neighbours and their bisectors define the edges of the Voronoi cells, which are convex polygons. Repeating the process recursively for u , if a point w is found, whose neighbourhood contains no points closer to p than itself, then p is contained within all possible open half planes containing w , defined by w and all its neighbours. Point p is then by definition located in point w 's Voronoi cell. This means that w is the point in \mathcal{T} closest to p . \square

Line Walking

Another point location algorithm to consider is the line walking algorithm [1]. This algorithm finds a triangle t in a triangulation \mathcal{T} that contains a given point v . Starting at any triangle s , with the geometrical centre m , if point v is not contained in s , then the line segment (v, m) intersects a finite set of triangles. The line segment (v, m) intersects two edges of each triangle in this set, with the exception of s and t where (v, m) only intersects one edge each. By iterating through each triangle choosing the neighbour triangle that contains the next edge that intersects (v, m) , triangle t can be found in $\mathcal{O}(n)$ time.

This algorithm was described by Amenta et al. [1], and is illustrated in the following figure, where the dark shaded triangles represent the starting and finishing triangles, and the light shaded triangles the path the algorithm takes to find the final triangle that contains the vertex v .

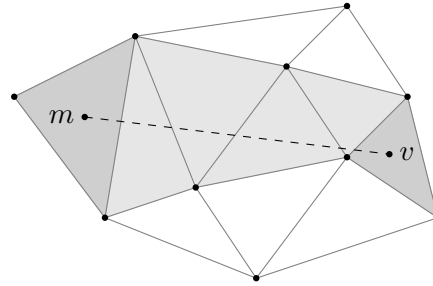


FIGURE 2.7: Illustration of the Walking Algorithm

After finding this triangle, the Bowyer-Watson algorithm described in Section 2.3.5 can be used to update the new triangulation, which now includes v .

2.3.6 Hilbert Curves

Most of the point location algorithms aforementioned have linear time complexity, and most of the worst case scenarios include searching across the plane. These occur when the starting search position is random and does not make use of the spatial organisation of the data. In order to fully take advantage of these approaches, the points should be sorted in such a way that the distance between consecutive points is minimised.

Hilbert curves are a kind of fractal space-filling curves [11] that generally minimize the Euclidean distance between points close on the curve.

True Hilbert curves map a 2-dimensional space in a 1-dimension line. This line has an infinite length, which makes mapping 2-dimensional points to it infeasible. Instead,

discrete approximations are used. Since the true curve is fractal, the approximations are defined by the number of fractal steps it takes in order to reach them. Figure 2.8 demonstrates the first few orders of approximation:

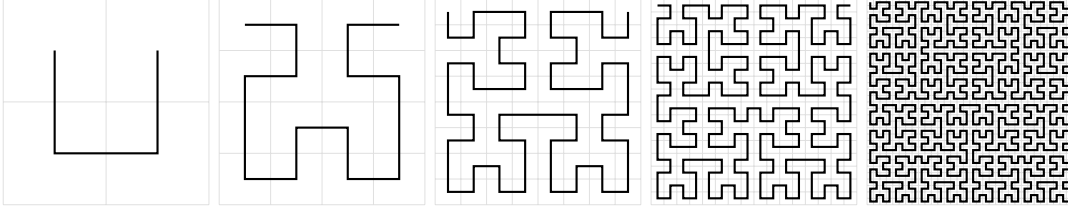


FIGURE 2.8: First Five Orders of Hilbert Curve Approximation

Since the coordinates of the points in our problem are continuous rather than discrete, the points must first be mapped into a square grid with tile size and number appropriate to the Hilbert approximation chosen. In order to sort an array of 2-dimensional points to follow a Hilbert approximation, each point should be assigned the 1-dimensional coordinate of the square tile that contains that point. The array is then sorted using the square coordinates along the Hilbert approximation as a key. This means that there are cases where more than one point will share the same discrete approximation coordinates, but this has little effect on the performance of the point location algorithms, as long as the grid is fine enough to separate most of the points. The space must be partitioned into a grid of 2^n squares in height and width and the grid must contain all points.

Chapter 3

State of the Art

The *k-centre* problem is a well known problem in the optimisation field of study. As such, several approaches have been explored over the years. In the following, two important approaches are described: an Integer Linear Programming approach and an Incremental Approach.

3.1 Integer Linear Programming Formulation

A simple and straight-forward approach to the problem is to model it in integer linear programming as follows:

$$\text{minimise } D \tag{3.1}$$

$$\text{subject to } \sum_{j=1}^N y_j = k \tag{3.2}$$

$$\sum_{j=1}^N x_{ij} = 1 \quad i = 1, \dots, N \tag{3.3}$$

$$\sum_{j=1}^N d_{ij} x_{ij} \leq D \quad i = 1, \dots, N \tag{3.4}$$

$$x_{ij} \leq y_j \quad i = 1, \dots, N; j = 1, \dots, N \tag{3.5}$$

$$x_{ij}, y_j \in \{0, 1\} \quad i = 1, \dots, N; j = 1, \dots, N \tag{3.6}$$

In this formulation, $y_j = 1$ if point j is a centroid and $y_j = 0$ if it is a non-centroid; $x_{ij} = 1$ if the point i is assigned to the centroid j , and $x_{ij} = 0$ otherwise; d_{ij} is the Euclidean distance between points i and j . Constraint (3.2) ensures that k centroids

are chosen. Constraint (3.3) limits the assignment of one point to more than one centroid. Constraint (3.4) ensures that all active distances are lower than the limit we are minimising. Constraint (3.5) limits points to being assigned only to centroids, where $y_j = 1$. Constraint (3.6) defines both x_{ij} and y_j as binary variables, in order to properly represent selection and assignment.

It is worth noting that this formulation minimises the objective function by selecting the best possible set of centroids, but it only minimises the maximum coverage. This way, only the farthest point from its centroid has the guarantee that it is connected to its closest centroid. Every other point, however, can be linked to any centroid so long as it is closer to it than the distance defined by the objective function, since $d_{ij}x_{ij}$ only has to be lower than D , but not include the lowest possible values. Likewise, it can also produce the result where one centroid is assigned to another centroid as opposed to itself, as long as they are close enough together and it does not affect the coverage of the whole set. These cases have no effect on the outcome of the final coverage value or the centroid selection, but are rather counter-productive, since we want to minimize the coverage of all centroids, with minimal overlapping of the covered areas.

In order to best display the results, a simple post-processing step can be applied, where each point will be strictly assigned to the closest centroid. This can be easily computed in $\mathcal{O}((N - k)k)$ time in case there is a need for a clearer display of the assignment.

Other more elaborate formulations can be used. Elloumi et al. [7] explore a new formulation to obtain tighter bounds in the LP relaxation. They also limit the values that the solution can take by enumerating all different values of distances between points and sorting them in decreasing order.

3.2 Local Search Approach

Cambazard et al. [4] solve the k -centre problem heuristically using local search. Their method describes algorithms to incrementally insert and remove centroids from a set of points, and update the centroid assignment only in the geometrical area surrounding the changed centroid. This method allows for small modifications on an already valid solution, until a similar solution is deemed optimal is found. It can be used as a fast way to calculate neighbour solutions to be used in heuristic approaches to the k -centre problem.

In order to minimise computation time, Cambazard et al. [4] maintain the selected centroids in a k - d tree. Using a k - d tree reduces the number of comparisons needed for the step of point location in the algorithm. To keep the k - d trees from losing efficiency

in insertions and removal of points, the trees need to be balanced from time to time. Calculating the optimal time interval to balance the *k-d tree* has to be done *a priori* and can affect the performance of the algorithm.

The structures and incremental procedures can be used with enumeration algorithms in order to obtain the global optimum solution.

Chapter 4

Optimal Minimum Coverage Algorithms

This chapter describes two possible algorithms that solve the coverage problem described in Section 2.1. Both algorithms use an incremental branch-and-bound approach for implicit enumeration of the centroid subsets.

The first algorithm is a naïve implementation of a branch-and-bound algorithm and uses simple loops over arrays for point location queries. The second algorithm builds and uses Delaunay triangulations to achieve more efficient point location queries.

4.1 Naïve Branch-and-Bound

A more sophisticated approach to the problem is to use a branch-and-bound method. In this approach, the assignment of non-centroids to their correct centroids is built incrementally. At each step of the recursive tree, one of the available points is considered. A decision is then taken of whether the point is a centroid or a non-centroid. According to which decision is taken, the objective function and the centroid assignment is updated accordingly. This is done until all the centroids have been chosen.

4.1.1 Branching

As stated above, branching the tree involves updating the assignment between new points and/or new centroids, as well as updating the objective function. The following procedures explain in detail how to do so.

Inserting a Centroid To insert a centroid c , the established non-centroids which are closer to c than their current centroid must be checked, and change their assignment to c . Since non-centroids only change assignment to centroids closer to them, inserting a centroid means that the objective function either decreases in value or stays the same. After inserting a centroid, if the farthest non-centroid is reassigned, all non-centroids must be checked to see which one now maximises the objective function. This step compares all non-centroids to the new centroid c , taking $\Theta(N)$ time.

Inserting a Non-Centroid Inserting a non-centroid n only requires finding which of the current centroids is the closest to n . Updating the objective function is a matter of testing whether the distance between n and its centroid is larger than the current maximum. Inserting a non-centroid cannot produce a better objective function, since it will either decrease or maintain the current value. This step compares the distance between point n and all centroids, taking $\Theta(k)$ time.

After a branch is fully calculated, it is necessary to backtrack to the parent state, either by removing a centroid, or a non-centroid.

Removing a Centroid Removing a centroid c means redistributing the values assigned to c to their respective closest centroids in the remaining set.

The value function either increases or maintains, since the distance for all points previously assigned to c will increase, potentially above the current value for the objective function. Removing a centroid c means comparing all non-centroids assigned to c to all the other centroids. This step takes $\Theta(NK)$ time to execute. Alternatively, if the assignment state is saved before inserting the centroid, recovering it requires only retrieving the state, which means, in the worst and best cases copying an array of size N , which takes $\Theta(N)$ time at the expense of additional $\Theta(N)$ memory space.

Removing a Non-Centroid In order to remove a non-centroid n , we only need to update the objective function. If point n maximises the objective function, the second farthest point from its centroid, the new maximiser, must be found. Removing a non-centroid can either decrease or maintain the value of the objective function. Removing a non-centroid n means that the next farthest point from its centroid must be found. This can be done by checking all distances between the non-centroids and their respective centroids, taking $\Theta(N)$ time. Alternatively, one can save the previous value for the objective function, as well as the maximiser. Retrieving the previous value can be done in $\mathcal{O}(1)$ time at the expense of additional $\Theta(1)$ memory space.

4.1.2 Bounds

At all steps in the branching, the lower bound for the value of the objective function in the current branches is calculated. If the lower bound is larger than an already calculated upper bound, then there is no purpose in further exploring the current branch. In a minimisation problem, the upper bound can be the best solution found until that point in time.

Lower Bound After each insertion, centroid or non-centroid, one can assume that, the best case scenario, all the points not yet inserted will be centroids. This would hypothetically decrease the value the most. If this value is larger than the best value found, then there is no possible assignment that will improve the current solution in the current branch, and the branch can be pruned.

4.2 Delaunay Assisted Branch-and-Bound

Most of the operations in the branch-and-bound approach described in Section 4.1 have at least linear time complexity for both the best and expected cases. We can speed these up by implementing incrementally built Delaunay triangulations, which can be used to accelerate point location queries. To aid the calculations, the points are pre-processed and sorted by a Hilbert Curve approximation of a sufficiently high order.

Inserting a Centroid In order to take advantage of Delaunay triangulations, each time a centroid is chosen, it must be included in the Delaunay triangulation. This means that the triangulation must be updated. Inserting a point in a triangulation with K vertices using the Bowyer-Watson algorithm described in Section 2.3.5 takes an estimated $\mathcal{O}(\log K)$ for a uniformly distributed set of points [12]. After a centroid c is included in a Delaunay triangulation, it is possible to know which other centroids are its Voronoi neighbours. This is due to the duality between Delaunay triangulations and Voronoi diagrams. Since Voronoi diagrams partition the space into regions by distance to the centroids, we only need to check the subset of non-centroids assigned to the direct neighbours of c to find which points should change assignment to c . This property lowers the expected number of comparisons to make. Since the average number of Voronoi neighbours per centroid in any given diagram cannot exceed six [8, 10], the number of points to be compared in a uniformly distributed set of non-centroids should not include all non-centroids, but only a small fraction of them. Despite the lower number of comparisons, the worst-case time complexity still takes $\mathcal{O}(N)$ time to complete, and

in the worst case scenario it can still require a check through all non-centroids, which can all be neighbours of c . If the objective function maximiser is assigned to c , all non-centroids can be candidates to become the new maximiser, so a linear search through all the non-centroids must be done, to see which one is now the farthest away from its centroid.

Inserting a Non-Centroid Since there is a triangulation built, using the centroids as its vertices, finding the closest centroid c to a new non-centroid n is simply a matter of using the greedy routing algorithm to find c [2]. The greedy routing algorithm has a worst-case time complexity of $\mathcal{O}(K)$. This happens when the search starts from the farthest centroid from n , and all centroids are either in the direction of n , or are neighbours of the centroids that are. The last centroid returned by the greedy routing algorithm can be used to start the new query. Since the points are inserted ordered by a Hilbert curve approximation, each consecutive point should minimise the position variation from the last. This means that, ideally, each inserted non-centroid n will be close to its respective optimally positioned centroid c , and it will only need to calculate the distances to the neighbours of c in order to guarantee that c is indeed the correct centroid. The aforementioned property of the average six neighbours for each centroid means that the expected time for a query starting at the right centroid would be $\mathcal{O}(1)$. This represents the best case scenario, and is heuristically approximated by the Hilbert curves. The time complexity of inserting one non-centroid is still $\mathcal{O}(K)$ for the worst case. However, the insertion of a large number of uniformly distributed points *should* behave closer to $\mathcal{O}(\sqrt{K})$ time per point. If a rectangular area has N points, the longest path would be a diagonal. The diagonal, like the sides, will have close to \sqrt{N} number of points, in an area with a sufficiently good uniformity of points in it.

Removing a Centroid Removing a centroid c means removing it from the Delaunay triangulation and redistributing all points assigned to c across its neighbours. Since all points are inserted in the triangulation in a LIFO order, removing a point from a triangulation is a matter of retrieving the previous state. We can do this by storing all new edges and triangles in a stack upon construction, and retrieve them upon removal, without the need of recalculating anything. Since inserting a centroid c takes expected $\mathcal{O}(\log K)$ time [12], and removing it will take exactly the same higher level operations (in reverse order), it can also be done in expected $\mathcal{O}(\log K)$ time, without the need to do extra calculations. Likewise, redistributing the points assigned to c takes retrieving the previous state. Each change in assignment can be saved in a stack upon insertion, and retrieving it can be done by popping the stack. This step also takes $\mathcal{O}(N)$ time, since all points can change assignment. However, using a stack limits the number of

operations to only those that changed upon insertion, which in an uniform distribution, means an expected time complexity of $\mathcal{O}(N/K)$.

Removing a Non-Centroid Removing a non-centroid n only requires recovering the second farthest point if n is currently the farthest point, otherwise, no operations besides erasing n 's assignment, taking $\mathcal{O}(1)$ time and memory.

Bounds The same lower bound described in Section 4.1.2 can be applied in this approach. Both algorithms have the same time complexity of $\mathcal{O}(N)$ for computing the bound.

These steps occur at each iteration of the branch-and-bound algorithm, and each is performed potentially 2^N times for both approaches. Despite having the same worst-case time complexity as the branch-and-bound algorithm described in Section 4.1, the expected time complexity for the Delaunay assisted approach is smaller. This approach should have better performance when a large number of centroids are needed. This is especially true since maintaining a valid Delaunay triangulation through all the centroid permutations, as well as the Hilbert sorting, takes a computing cost. This extra overhead will have a negative impact in the performance in the smaller instances of the problem.

4.3 Algorithm Comparison

4.3.1 Time Complexity

Each of the procedures mentioned in the previous Section are performed at each step of the recursive tree. At each step the bound is also calculated, which takes $\mathcal{O}(N)$ time. Table 4.1 shows the time complexities for each procedure in both algorithms. The values presented at the row corresponding to the average case of the Delaunay-assisted algorithm are conjectured and need to be shown in a more formal manner.

Algorithm	Insert		Remove	
	Centroid	Non-Centroid	Centroid	Non-Centroid
Naïve BB	$\Theta(N)$	$\Theta(K)$	$\Theta(N)$	$\mathcal{O}(1)$
Del. Assisted BB Average Case	$\mathcal{O}(\log K + N/K)$	$\mathcal{O}(\sqrt{K})$	$\mathcal{O}(N/K)$	$\mathcal{O}(1)$
Del. Assisted BB Worst Case	$\mathcal{O}(K + N)$	$\mathcal{O}(K)$	$\mathcal{O}(N)$	$\mathcal{O}(1)$

TABLE 4.1: Time complexities for the various operations in a uniformly distributed point set

4.3.2 Experimental Results

In this section, we analyse empirically the time spent calculating the solutions to different sizes of the problem.

Methodology and Set-up

The test cases are sets of uniformly distributed points generated with a fixed seed. Each test was repeated 10 times with different sets of points. The same sets and machine were used to test the three algorithms.

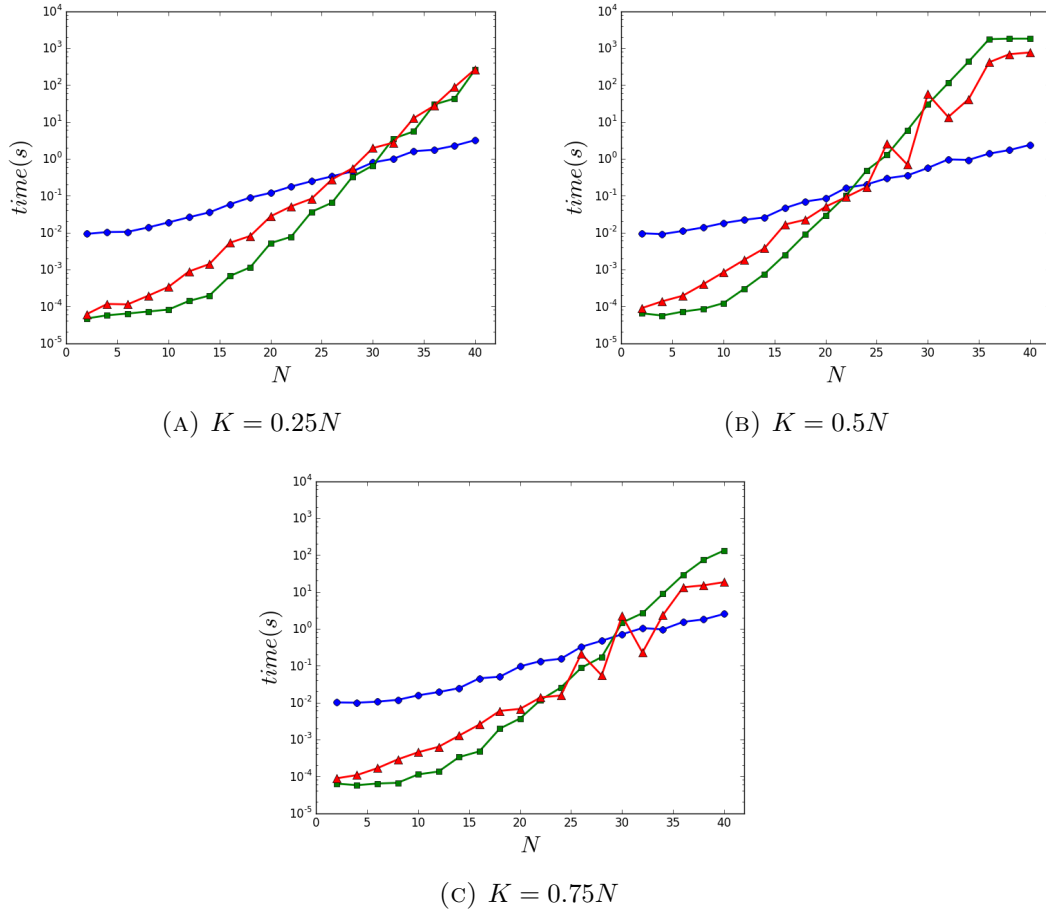
Both branch-and-bound approaches were implemented using *C++* and compiled using *g++ 4.9.2*. The integer linear programming version had the data preprocessed using *python 2.7.9* and was solved using the *GLPK LP/MIP v4.55* solver. The machine that ran the tests had the following specifications:

Operating System	Arch Linux 3.14.4 x64
CPU	Intel i7 Dual-Core, 2GHz
Memory	8 GB, 1600 MHz
Storage	Solid-State Drive, 300 MB/s (read)

TABLE 4.2: Machine specifications

Effect of N

The first test conducted analysed the variance in performance relative to changes in the value of N . It was done by changing N , with K taking fixed fractional values of N .



● – Integer Linear Programming ▲ – Delaunay Assisted B&B ■ – Naive B&B

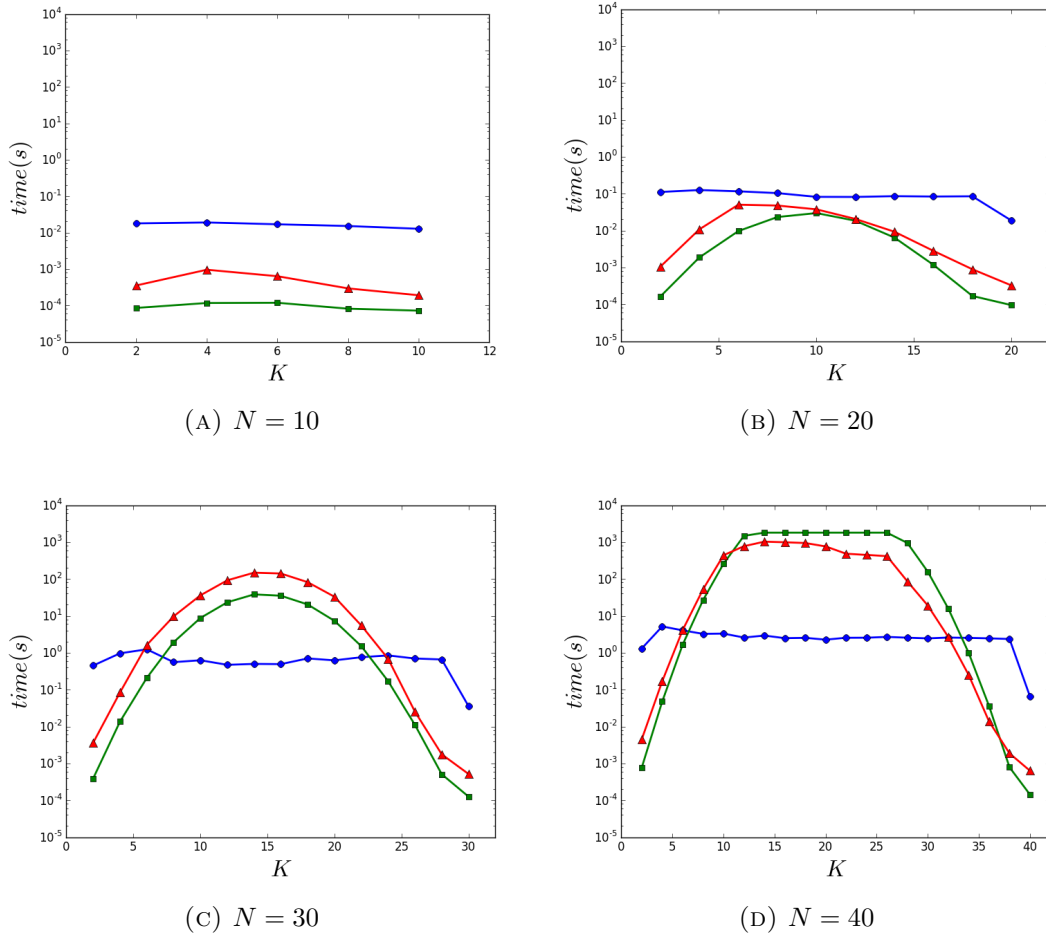
FIGURE 4.1: CPU-time for different values of K with varying values of N

Figure 4.1 shows the results of these tests. The measures were taken in seconds and account for the pre-processing steps and the solving time, but not the input reading or output writing times. The tests were stopped at the half-hour mark.

As it can be seen, the problem solving time increases exponentially with the value of N , as expected. The Integer Linear Programming Approach performs faster for larger values of N . Comparing the branch-and-bound approaches, these tests show that the Delaunay-assisted algorithm steadily approaches and surpasses the naïve branch-and-bound algorithm as the instance size grows.

Effect of K

In this experiment, we analysed the performance of the three algorithms in dependence of parameter K . We fixed N and varied K from 2 to N by steps of 2. Figure [?]]



● – Integer Linear Programming ▲ – Delaunay Assisted B&B ■ – Naive B&B

FIGURE 4.2: CPU-time for different values of N with varying K

show the results for the $N = \{10, 20, 30, 40\}$, respectively. The tests were stopped at the half-hour mark. Figure 4.2 shows the results of these tests.

The Integer Linear Programming approach seems to be the fastest approach for most cases, and seems independent to the value of K . The exceptions to this seem to be smaller values of N , as well as the smallest and largest values of K . This happens because the implicit enumeration methods only need to enumerate a very small number of combinations. As for the branch-and-bound algorithms, the Delaunay-assisted approach is slower than the Naïve implementation in these test cases. However it should be noted that for each N , the Delaunay-assisted algorithm peaks before the expected value of $K = N/2$. This is justified due to the fact that the Delaunay triangulation has an overhead which can take advantage for in larger values of K . Furthermore, the fact that the Naïve implementation had no test for the middle values of K for $N = 40$ that ended before the time-out mark is noteworthy. It is also worth noting that The

Delaunay-assisted approach showed a lot more variance between tests, often taking much lower values than the mean. However, for the tests performed, two runs had values much larger than the Naïve approach, approximating the Delaunay-assisted algorithm's mean to the Naïve approach.

The time required for each test limited the number of tests performed. Because of this, the results may not be statistically meaningful. This could mean that the Delaunay-assisted approach is only preferable for values of N and K to which neither approach is usable in real-time. Due to the small number of tests for large values of N , this result may not be statistically meaningful, but it is noteworthy.

Chapter 5

Future Work

5.1 Integration in a Visualisation Framework

The final objective of this study is to integrate an algorithm in a web application that can display the most representative set. However, the feedback time on the application needs to be as small as possible while still delivering an acceptable set of points, in order to not lose engagement from the user.

The application will display a rectangular window, showing a cut of geographical region containing a set of points. The algorithm chosen will need to be able to choose a representative set of points within the cut quickly, as well as be able to recalculate a new set points for a new cut, resulting from panning or zooming the display window over the region.

The algorithm will serve as the middleware responsible for filtering the response of a GIS server to a Web Feature Service, or WFS request. WFS lists the geographic coordinates of the points to be represented in an image by the coordinates mapped into an orthogonal plane, representing an image containing the cut of the region requested by the application. The candidate algorithms will be tested and benchmarked using data from the Open Street Map project. The project features large quantities of open source geographic data, as well as a versatile API for fetching data in the WFS standard.

5.2 Heuristic Approaches

Exact algorithms, due to their slow time performance, make them a poor choice for real-time applications. As such, heuristic algorithms that provide good but not optimal solution in faster time are more likely the most appropriate approach. Since a lot of

complex structures have already been explored and implemented in the implicit enumeration approaches, a lot of the concepts and methods can be repurposed and reused when experimenting and researching heuristic approaches.

For instance, since the two interactions with any calculated solution will be panning and zooming, some new solutions may share some points. If that is the case, calculating a solution after a small pan or zoom may reuse the previously calculated region as a starting point, only adding the new points, and removing the previously calculated ones.

An extension to this method may include calculating a larger area than the queried one, so that a small pan and/or zoom include already calculating data. After the window moves, the new adjacent regions can be calculated. This way, the user should never see the window without processed data, and the program will only calculate areas outside the vision range of the window.

5.3 Approximation Algorithms

Approximation algorithms are used to used in optimisation as a means to achieve a valid solution within a guaranteed minimum factor of quality to the optimal solution. Ideally, the approximation is optimal up to a constant factor, the smaller the better.

Approximation algorithms have been found for this problem and described by Shmoys et al. [13], which give an approximation with a factor of 3.16 to the optimal solution of the *k-centre* problem. The efficiency of approximation algorithms makes them appropriate for solving hard problems, and will be considered for the final approach.

References

- [1] N. Amenta, S. Choi, and G. Rote. Incremental constructions con BRIO. In *Proceedings of the Nineteenth Annual Symposium on Computational Geometry*, pages 211–219. ACM Press, 2003.
- [2] P. Bose and P. Morin. Online routing in triangulations. In *Algorithms and Computation*, pages 113–122. Springer, 1999.
- [3] A. Bowyer. Computing dirichlet tessellations. *The Computer Journal*, 24(2):162–166, 1981.
- [4] H. Cambazard, D. Mehta, B. O’Sullivan, and L. Quesada. A computational geometry-based local search algorithm for planar location problems. In *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*, pages 97–112. Springer, 2012.
- [5] M. S. Daskin. *Network and discrete location: models, algorithms, and applications*. John Wiley & Sons, 2011.
- [6] M. De Berg, M. Van Kreveld, M. Overmars, and O. C. Schwarzkopf. *Computational Geometry*. Springer, 2000.
- [7] S. Elloumi, M. Labbé, and Y. Pochet. A new formulation and resolution method for the p-center problem. *INFORMS Journal on Computing*, 16(1):84–94, 2004.
- [8] L. Hu, W.-S. Ku, S. Bakiras, and C. Shahabi. Verifying spatial queries using voronoi neighbors. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 350–359. ACM Press, 2010.
- [9] N. Megiddo and A. Tamir. New results on the complexity of p-centre problems. *SIAM Journal on Computing*, 12(4):751–758, 1983.
- [10] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, volume 501. John Wiley & Sons, 2009.
- [11] H. Sagan. *Space-Filling Curves*, volume 18. Springer-Verlag New York, 1994.

-
- [12] J. R. Shewchuk. *Lecture Notes on Delaunay Mesh Generation*. Department of Electrical Engineering and Computer Science, University of California at Berkeley, 1999.
 - [13] D. Shmoys, É. Tardos, and K. Aardal. Approximation algorithms for facility location problems. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 265–274. ACM Press, 1997.
 - [14] D. Vaz. Subset selection algorithms in multiobjective optimization. MSc in informatics engineering, University of Coimbra, Portugal, 2013.