Towards Q-analysis Integration in Discrete Global Grid Systems: Methodology, Implications and Data Complexity

by

Veniamin Bondaruk

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Science
in
Geography

Waterloo, Ontario, Canada, 2020

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Spatial data is characterized by rich contextual information with multiple characteristics at each location. The interpretation of this multifaceted data is an integral part of current technological developments, data rich environments and data driven approaches for solving complex problems. While data availability, exploitation and complexity continue to grow, new technologies, tools and methods continue to evolve in order to meet these demands, including advancing analytical capabilities, as well as the explicit formalization of geographic knowledge.

In spite of these developments Discrete Global Grid Systems (DGGS) were proposed as a new comprehensive approach for transforming scientific data of various sources, types and qualities into one integrated environment. The DGGS framework was developed as the global data model and standard for efficient storage, analysis and visualization of spatial information via a discrete hierarchy of equal area cells at various spatial resolutions. Each DGGS cell is the explicit representation of the Earth surface, which can store multiple data values and be conveniently recognized and identified within the hierarchy of the DGGS system.

A detailed evaluation of some notable DGGS implementations in this research indicates great prospects and flexibility in performing essential data management operations, including spatial analysis and visualization. Yet they fall short in recognizing interactivity between system components and their visualization, nor providing advanced data friendly techniques. To address these limitations and promote further theoretical advancement of DGGS, this research suggests the use of Q-analysis theory as a way to utilize the potential of the hierarchical DGGS data model via the tools of simplicial complexes and algebraic topology. As a proof of concept and demonstration of Q-analysis feasibility, the method has been applied in a water quality and water health study, the interpretation of which has revealed much contextual information about the behaviour of the water network, the spread of pollution and chain affects.

It is concluded that the use of Q-analysis indeed contributes to the further advancement and development of DGGS as a data rich framework for formalizing multilevel data systems and for the exploration of new data driven and data friendly approaches to close the gap between knowledge and data complexity.

Acknowledgements

First and foremost I would like to express gratitude and appreciation to my supervisor, Dr. Steven A. Roberts, for his professional guidance, expertise in the field and commitment to this research. Without a doubt this work would not have been be possible without him.

Particular recognition goes also to Dr. Colin Robertson for his valuable suggestions, constructive feedback and input, which had a significant impact on the direction and quality of this research.

I am also thankful to the examining reader, Dr. Rob Feick, for taking the time to review this work and validate its quality, as well as for sharing his expert knowledge and committing to this integral responsibility.

Thank you to the Global Water Futures Programme and the Global Water Citizenship Project for funding this research.

I wish to thank Dr. Brent Doberstein, Dr. Richard Kelly and the Department of Geography and Environmental Management at the University of Waterloo for providing administrative support, academic guidance and financial assistance. Certainly, their hard work and efforts were vital and meant a lot to me.

I also want to thank all my colleagues and fellow researchers in The Spatial Lab at Wilfrid Laurier University. It was a pleasure working with them and I am glad our paths have crossed.

Finally, I want to express special gratefulness to my family and church for taking on this journey with me, their endless love and encouragement without measure. I love them a lot.

Dedication

For the glory of Christ Jesus.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

With the growing significance and impact of the data driven and data friendly approaches on the spatial analysis in Geographic Information Systems (GIS) (Comber & Wulder, 2019; Graham & Shelton, 2013; Miller & Goodchild, 2015), as well as social, scientific and economic dynamics influenced by the data rich culture and data accessibility (Johnson et al., 2018), it is important to address big data analysis and challenges associated with it. In a general sense, attempts have been made to describe big data traits not only in terms of the volume, velocity and variety characteristics (Kitchin & McArdle, 2016), but also digital availability and complex integrated environments (Barrett et al., 2018; Johnson et al., 2018), which has led to the evolution of new data platforms (Li et al., 2015), software tools (Sowkhya et al., 2018), sensors (Arza-García et al., 2019) and computing architectures (Helmi et al., 2018). These developments were put in to practice to meet additional demands for data optimization, adaptability, extensibility, scalability and flexibly (Barrett et al., 2018; Kitchin & McArdle, 2016), which are argued to have essential role for gaining better conceptual understanding of big data in the current data rich age.

One such data storage infrastructure technologies Discrete Global Grid Systems (DGGS) were found to be a prominent and comprehensive approach suitable for integrating large data quantities of various sources, types and qualities, as well as providing analytical capability necessary for its interpretation (OGC, 2017). Implications of DGGS have been officially recognized by the Open Geospatial Consortium (OGC) and the corresponding OGC DGGS abstract specification, developed to outline the fundamental features and requirements necessary for the core DGGS data model.

While the emergence of DGGS have provided the GIS community with great advantages in terms of data storage, access, interpretation and visualization, much of the operational and functional capability of these systems lack completeness, as well as integration with other methods, technologies and data standards (Bondaruk et al., 2019). In particular, multi-relational interaction complexity of data networks and exchange of information through the multilevel hierarchical DGGS data model requires more explicit articulation and theoretical developments to meet the requirements for complex data organization, management and analysis. Additional attention should also be given to the visualization component, since with the complex data

model, such as DGGS, it can be challenging to properly convey DGGS embedded multivariate data, while preserving interaction complexity and assisting interpretation.

In the light of these developments the research has turned to the study of complexity and complex systems science (Johnson, 2014) with the attempt to address the challenges related to the DGGS analytical capability, data complexity and interpretation, as well as to promote further advancement of DGGS data model and improvement of its OGC abstract specification. One of the objectives of complex systems science is to provide methods and frameworks capable of generalizing information and interacting components from different study areas or domains, in order to observe system behaviour, its connectivity structure and discover new applications. In fact, it is possible to view DGGS data model as one of such complex systems, due to the same properties of multilevel and discrete dynamics, network connectivity and globality, which are compatible with the notion of complexity (Johnson, 2014, pp. 6–10).

In the context of complex systems science, this research specifically emphasises advantages and application capability of Q-analysis theory (Atkin, 1972, 1974, 1980; Johnson, 2014), as a method suitable to address some of the challenges related to DGGS analytical capability, data complexity and interpretation, including its compatibility with the hierarchical DGGS data model and its effective visualization. Specifically, Q-analysis utilizes concepts of algebraic topology and relational algebra in order to retain as much contextual information as possible about interactivity between components, and explicitly model system behaviour.

Although, Q-analysis originated from the branch of mathematics algebraic topology, some attempts for its use in the areas of spatial analysis and GIS (Jiang & Omer, 2007; Omer & Goldblatt, 2017; Roberts et al., 2001), as well as big data and complexity (Maletić & Zhao, 2017) have been made in the past. These and similar studies have recognized the implications of Q-analysis and its fundamental concept of simplicial complexes to explore structural properties of geographic systems through chains of connected components, and modelling of hierarchical datasets. These implications also include graphical representation of simplicial complexes in order to enhance the overall analysis and its interpretation. In other words, Q-analysis' theoretical method was found useful to facilitate the relational structure between various system components and provide additional insight into the system complexity, its behaviour and visualization.

1.1 Research Scope

The goal of the research is to contribute to the long term theoretical advancement of DGGS and improvement of their analytical capabilities for the search of new and more explicit data driven and data friendly solutions, as well as understanding of complexity in the data rich DGGS environment and architecture.

To achieve the goal, the study has identified the following research objectives:

- Perform a detailed assessment of some of the notable DGGS implementations, their functional operability, performance and conformance with the OGC abstract specification;
- Search for the new data driven methods and analytical techniques compatible with the hierarchical DGGS data model, as well as to determine the extent to which such methods are suitable to handle data complexity and interpretation;
- Develop a corresponding methodology capable of formalizing spatial knowledge within DGGS framework and remain broad enough for use across various datasets, domains and applications;
- Present a case study validating the effectiveness of the developed methodology and its significance for the theoretical advancement of DGGS and improvement of their analytical capabilities.

This research strives to provide a scientific argument and a procedural framework which permits the use of Q-analysis and simplicial complexes within the hierarchical DGGS data model.

1.2 Thesis Outline

The outline of this thesis consists of six chapters:

Chapter 1, the current chapter, introduces the problem statement, research scope and provides necessary background information which underlines the significance, scientific relevance and motivation of the undertaken research.

Chapter 2 elaborates on the main concepts introduced in Chapter 1 and provides in-depth background review of DGGS and Q-analysis, their historical developments, specific technical details and applications.

Chapter 3 conveys information regarding key methodological developments and arguments necessary for the use of Q-analysis within hierarchical DGGS data model. The methodology provides the core implementation procedures for running the Q-analysis, including derivation of the necessary components, such as definition of cover sets, matrix construction, creation of simplicial complexes and their visualization.

Chapter 4 is a case study chapter dedicated specifically to apply the proposed methodology from Chapter 3 in a real-world scenario, in order to demonstrate its analytical capability and its contribution to the development of DGGS.

Chapter 5 communicates the main results related to the assessment of DGGS, theoretical advancement and interpretation of the output from the hierarchical Q-analysis performed on the case study from Chapter 4.

Chapter 6 summarizes the main outcomes and makes final conclusions in accordance with the research goal and objectives. In addition, potential future work and next steps forward to improve the research are also discussed and concluded here.

Chapter 2 Literature Review

The following chapter provides necessary background information for the research scope covered in this thesis. In particular, two main subjects: DGGS and Q-analysis are reviewed in detail here and organized accordingly. Section 2.1 provides introductory information on DGGS as a comprehensive approach to model the Earth. Section 2.2 outlines key criteria necessary for implementation of a fully functional DGGS. Section 2.3 describes some of the most notable DGGS implementations in the industry. Section 2.4 is a transition section which reflects on the need for new data handling techniques, such as Q-analysis, to be explored within a DGGS context, and draws some parallels for its use. Sections 2.5 and 2.6 describe key concepts and advantages of Q-analysis as a language of structure. Lastly, Section 2.7 lists important Q-analysis applications which greatly contributed to the direction of this research.

2.1 Introduction to DGGS

DGGS are hierarchical tessellations of regular shaped polygons that form sets of equal area cells to partition and reference the Earth's surface. Traditionally, geospatial information is referenced on a continuous space via the projected Cartesian or ellipsoidal coordinate systems, which sometimes leads to a discrepancy between sampled data and its location precision (OGC, 2017). DGGS, on the other hand, offer a discrete way for spatial referencing using cell units as its base rather than coordinates taken from underlying continuous space. By its definition DGGS is a hierarchical tessellation of nested cells fully covering the Earth's surface, which means that each cell's location is fixed and associated with a specific area. Such implementation appears to be superior since information associated with those cells is also explicitly associated with a geographic space with a fixed level of accuracy and attribute mapping at various spatial resolutions.

DGGS provide a spatial structure on which data values of many types and formats, such as vector and raster, can be integrated, visualized and analysed in one uniform and homogeneous environment (Mahdavi-Amiri, Alderson, et al., 2015; Mahdavi-Amiri et al., 2016). Given the hierarchical nature and nesting properties of DGGS cells, multiple data sources can be

aggregated or distributed accordingly with arbitrary positional precision and accuracy, even though higher precision might not be necessary for practical purposes and can be limited by existing hardware and software specifications (e.g., memory management or float number handling). All these properties allow for consistent and unique representation of spatial data, as well as the development of external methods for spatial analysis that could operate on DGGS structure independently (Purss et al., 2016).

The process of generating a DGGS involves several design parameters, which can be chosen according to one's needs or application purposes. Five of such design choices have been outlined and properly articulated, which include a choice of a base regular polyhedron; a fixed orientation of the polyhedron relative to the Earth's surface; a hierarchical partition method of the polyhedron; a method of transforming polyhedron's faces onto the Earth's surface; and a method for referencing and assigning data to grid cells (Sahr et al., 2003). The goal of the following subsections is to review some of these design parameters in more details and outline key components which are necessary to consider when dealing with a DGGS.

2.1.1 Base Polyhedron

Due to the Earth's curvature and its roughly spherical shape, projecting its surface onto a planar space necessarily results in some distortion in shape, size, distance, or orientation (Olson, 2006; White et al., 1992, 1998). For a cell-based spatial referencing system of the Earth, equal-area cell coverage is a desirable property (i.e., one not achieved by geographic coordinates). In order to maintain as much as possible area and shape properties on a global scale, a spherical approximation of the Earth can be modelled as a Platonic solid (Song et al., 2002), which is then itself tessellated, and the tessellated cells projected back onto the spherical earth. Platonic solids are also known as regular polyhedra, which are the three-dimensional solid geometric spaces formed with one kind of a regular polygon with equal sides and interior angles (e.g., square or triangle). There are only five Platonic solids or regular polyhedra (Figure 2.1) (Wells, 2012; White et al., 1992). All DGGS that aim to develop equal-area cell structure on the Earth take this approach (Sahr et al., 2003) with one of the five Platonic solids as the geometric base of the grid system (Figure 2.1).

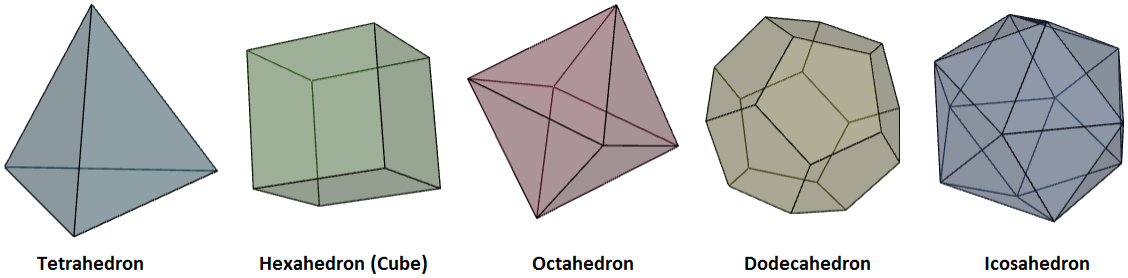


Figure 2.1. The five Platonic solids to be used as the initial tessellation of DGGS. The figure was generated via the Polyhedra Viewer (Nat, 2018).

Out of the five available options, the icosahedron has the largest number of faces (20) and the smallest areal proportion for each face (1/20) (White et al., 1992), which results in the least overall distortion (White et al., 1998). Yet, none of the Platonic solids tessellate the sphere surface perfectly. The areas and interior angles of recursive partitions of regular shapes cannot be all equal, thus either or both equal area and shape properties cannot be preserved completely (White et al., 1992). In order to minimize such distortions it was proposed to use even smaller regular shaped faces and a semiregular polyhedron instead, such as truncated icosahedron with hexagon and pentagon faces at its base (Figure 2.2) (Snyder, 2006; White et al., 1992).

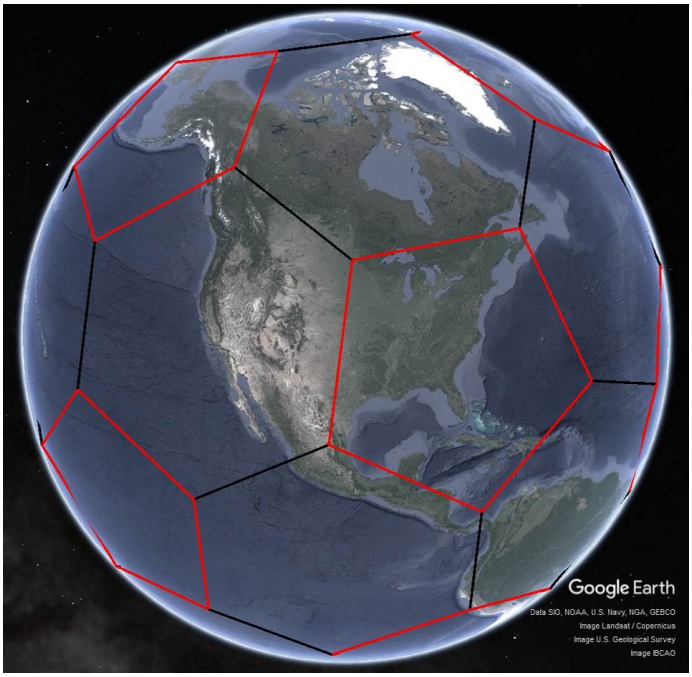


Figure 2.2. A truncated icosahedron projected onto the Earth's surface using the dggridR library (Barnes, 2016). The solid is composed of 12 pentagon (red) and 20 hexagon (black) faces.

A truncated icosahedron is a better approximation of a sphere, which can be conceptualized by truncating an icosahedron solid at the 12 vertices with 1/3 the length of each of the edges radiating from the vertex to create pentagons. In other words, the centers of each pentagon are aligned with 12 icosahedron's vertices. The rest of the solid partition is composed of 20 hexagons remaining from original triangular faces of the icosahedron (White et al., 1992). The truncated icosahedron has 32 faces in total, 12 of them pentagons and 20 hexagons. This means that further partition of the truncated icosahedron requires handling of two regular shapes (i.e., hexagon and pentagon) which adds complication for construction of a DGGS and global sampling. Nevertheless, it is considered acceptable since the number of pentagons is always equal to the number of icosahedron vertices (12) and this remains constant at each level of grid resolution (Sahr et al., 2003).

Additionally, studies have shown that the truncated icosahedron is the most effective in preserving equal area and shape properties compared to other regular polyhedra (Snyder, 2006; White et al., 1992), and does not violate defined OGC criteria (see Section 2.2). As a result, the truncated icosahedron has been widely integrated in order to design hexagon-based DGGS (Brodsky, 2018; Bush, 2017; Sahr, 2018). Nevertheless, other polyhedra choices for discretization of the sphere, such as octahedron (Górski et al., 2005) or hexahedron (Gibb, 2016; Veach, 2017), have also been used and are claimed to be effective based on their practical applications in astronomic data analysis (Górski et al., 2005), compliance with OGC (Gibb, 2016) and analysis of Google's collection of geographic data (Veach, 2017).

2.1.2 Polyhedron Registration

A fixed polyhedron registration relative to the Earth's surface is the next step in designing DGGS. It includes permanent assignment of a polyhedron's vertices in specific locations and orienting its connected edges in particular directions across the globe while preserving the equal area and shape properties as best as possible. As mentioned earlier, projecting polyhedron onto a sphere results in distortion. It is worth noting that distortion is also variable and tends to alternate with changes of the projection angle and distance. This indicates that distortion is also directly related to the choice of a polyhedron, since projection angle and distance is not the same for all

Platonic solids. In the literature, the resulting distortion of transforming a planar space onto a spherical space is also referred to as angular distortion (Mahdavi-Amiri, Samavati, et al., 2015; Snyder, 2006) or line distortion (Tong, Ben, Liu, et al., 2013). In a general sense, distortion increases further from a centroid of a polyhedral face and reaches its maximum at the vertices, since the angular distance is the largest in those places (Figure 2.3). As a result, the initial registration or placement of the polyhedral vertices is of great importance, since those areas will be impacted by the angular distortion the most.

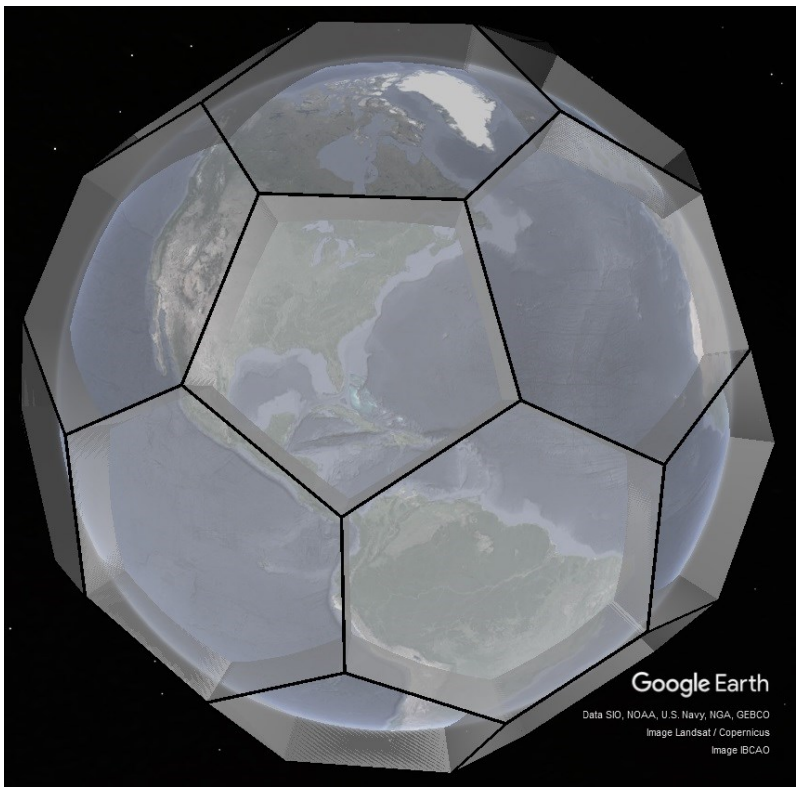


Figure 2.3. Illustration of the angular distortion as a result of projecting the truncated icosahedron onto the surface of the Earth. Each vertex point is the place of the maximum distortion.

There are many possible ways to specify the registration of a base polyhedron, each having their advantages and disadvantages. In a nutshell, the base polyhedron's registration can be specified by assigning latitude and longitude coordinates to one of its vertices and the azimuth angle to the next neighbouring vertex (Sahr et al., 2003). This approach works for all Platonic solids and can be used to derive locations of the remaining vertices. The method appears to be extremely flexible and has advantage of meeting specific needs for large variety of applications. In

particular, if the focus is on a specific geographic region or continent, as opposed to the whole globe, then places of maximum distortion could be repositioned away from paces of interest or regions important for the analysis.

Some other orientation choices which have been successfully used in the past include a hexahedron based DGGS. It can be oriented in a way that centroids of the top and bottom faces are places at the North and South Poles respectively, and the remaining four faces are aligned along the equator (Gibb, 2016, p. 4). Similarly, an icosahedron can be also oriented by placing two of the vertices at the North and South Poles, to provide partial symmetry along the prime meridian by matching its orientation with the orientation of one of the projected edges. This orientation tends to be one of the most popular since icosahedron is oriented in a familiar way and partially matches the usual latitude and longitude grid (Sahr et al., 2003, p. 124). The icosahedron can be also oriented in a way to preserve symmetry around the equator, as well as by mirroring the well-known Fuller's Dymaxion map projection (Leslie, 2001). By Fuller's definition, all vertices of icosahedron are places in the water in order to minimize the distortion on land. In the past, Fuller's Dymaxion orientation was known as the only one with such property (Sahr et al., 2003, p. 125). A recent study, however, developed a functional methodology which allows one to generate multiple orientations that maximize distances of all vertices away from landmasses, waterbodies or points of interest (e.g., Poles of Inaccessibility) (Barnes, 2019). The method is capable of generating several orientations for various polyhedra, which further grants some flexibility in minimizing effect of distortion in areas of interest. As illustrated in the study, multiple orientations that place all vertices in water are now possible for cuboctahedron, icosahedron, octahedron and tetrahedron (Barnes, 2019). If necessary, an icosahedron can be converted to the truncated icosahedron in order to minimize areal and shape distortion even further.

In summary, in order to meet specific application needs a polyhedron orientation can be specified by manually assigning one of the vertices to specific latitude and longitude coordinates and choosing the azimuth angle to its neighbouring vertex (Sahr et al., 2003). Regardless of the registration choice, to maximize the effectiveness of a DGGS it is important to recognize its core use cases in early development stages and adjust accordingly. Therefore, it is advised to explore different options ahead of time in order to select or design a DGGS with the required properties.

2.1.3 Hierarchical Tessellation

Hierarchical partitioning is one of the fundamental properties of DGGS, which is concerned with attribution of spatial data at various scales. In order to construct hierarchical partitions on the sphere, a polyhedron must have a recursive geometric tessellation using a base cell shape applied to its faces. These cells shapes are largely limited to the three main regular polygons that have been used in the past, such as squares, triangles and hexagons (OGC, 2017). Each of them have their own benefits and disadvantages, but the requirements for choosing the appropriate partition method depends very much on the application as well as considerations of adjacency and congruency properties (Peterson, 2017). To elaborate, adjacency concerns the connectivity properties and arrangement of cells in relation to their neighbours by sharing either an edge or a point. There can be uniform adjacency, when all neighbouring cells are attached similarly, or non-uniform adjacency, when cells do not have the same neighbourhood connectivity amongst each other. Similarly, congruency indicates the tessellation property amongst shapes, where congruent implies that a shape can be tessellated into self-similar shapes and fit perfectly within their own shape boundary without an overlap (e.g., square divides into four smaller squares), and non-congruent implies that shapes cannot do this (e.g., a hexagon does not divide evenly into any number of smaller hexagons). Additionally, the number of shapes that can subdivide a coarser resolution parent is known as the aperture level. For example, if a shape can be tessellated into four self-similar shapes, its aperture is four. Aperture is the key parameter choice in the definition of a DGGS, and was varied in the DGGS explored here.

While square grids might be the most familiar and popular choice due to the wide variety of applications in satellite remote sensing as well as compatibility with current hardware and software (Peterson, 2017), their topology is not directly compatible with a truncated icosahedron – one of the best approximations of the sphere in preserving equal area and shape properties. At the same time, it was suggested to use a diamond shape as a square alternative and to partition the icosahedron with it instead, if necessary (White, 2000). The diamond shape might be considered as one of the popular tessellation choices, since they are recognized as the skewed transformation of squares and retain close geometric similarity with them. Diamonds partially inherent many algorithms that already exist for squares, such as addressing (White, 2000). On the contrary, diamonds cannot be directly transferrable from squares if neighbour distance is one

of the required computational components, since neighbourhood distance differs for both shapes (White, 2000). Both, squares and diamonds are congruent, but not uniform adjacent (Figure 2.4).

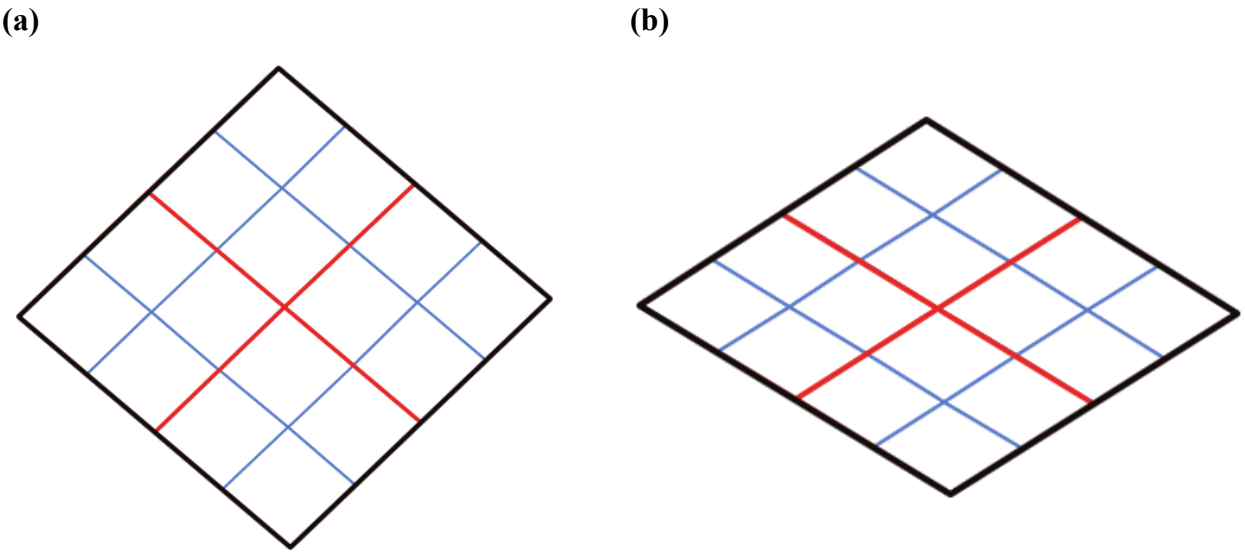


Figure 2.4. Illustration of (a) square based and (b) diamond based grid properties generated with S2 (S2Geometry, n.d.) and dggridR (Barnes, 2016) software. Note the similarity in their topologies but distance differences between neighbours.

Triangular shapes might also appear to be the most natural choice for the hierarchical partition of a polyhedron since three out of five Platonic solids have a triangular face at their bases, including the icosahedron. Similar to squares, a triangular tessellation is congruent and can be partitioned into self-similar shapes without overlap. A triangular tessellation is supported by a number of graphical rendering algorithms (Mahdavi-Amiri, Samavati, et al., 2015), as well as hierarchical computational models, such as quaternary triangular mesh (QTM), which is noted to be effective for spatial access and data visualization on a sphere (Dutton, 1999). Turning two adjacent triangles into a diamond shape makes it possible to implement fast and highly efficient addressing and neighbour finding algorithms for spatial databases and geometric operations. Such transformations also allow for integration of the square-based algorithms onto triangular grids, since diamonds are compatible with triangles as well as squares (Bai et al., 2005). As a standalone shape, however, triangles tend to be fairly complicated since they have the worst non-uniform adjacency out of the three available shapes. In particular, a single triangle has 12 neighbours with 3 of them having an edge connection and 9 containing a vertex connection. Out

of the 9 vertex connections 6 of them have angular attachment and 3 have straight attachment (Figure 2.5). Furthermore, each type of adjacency has a different distance from the central cell and a non-uniform orientation (e.g., triangles pointing upwards or downwards), which makes it difficult for implementation and handling of spatial operations, such as adjacency analysis, spatial query and data update (Bai et al., 2005).

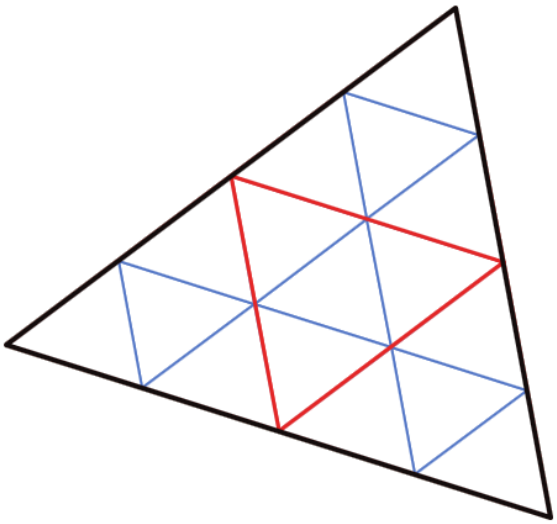


Figure 2.5. The adjacency and aperture properties of a triangular shape. The hierarchical partitions of these grids were generated via the OpenEAGGR software (Riskaware Ltd., 2019).

Hexagons have also received much attention in the literature due to a number of advantages, and appear to becoming one the common partition choices for DGGS (Ben et al., 2018; Mahdavi-Amiri et al., 2019; Sahr, 2008, 2013, 2019; Tong, Ben, Liu, et al., 2013; Tong, Ben, Wang, et al., 2013). Hexagons have been used successfully in data sampling and modelling for both image processing in computer vision (Li, 2013; Middleton & Sivaswamy, 2005, p. 10) and ecological simulations (Birch et al., 2007). Hexagons are attached similarly to each other by sharing only one edge of the same length, giving them the uniform connectivity property. In total, there are six neighbors and all of them are equidistant – located at the same distance from the central cell (Middleton & Sivaswamy, 2005, p. 2). Because of this property, hexagon-shaped cells have six different directions which can be used to model spatial objects with curved boundaries more effectively. A hexagon's six cell edges consistently model a finer resolution of angles to neighbouring cells, compared to squares or triangles.

In terms of shortcomings, hexagons are non-congruent and cannot be refined into self-similar shapes without overlap. In other words, fine resolution hexagons cannot be tessellated and fit perfectly within the lower resolution shape boundary. And since it is also impossible to complete the sphere with hexagons alone, non-hexagon faces (e.g., pentagons) must be introduced at each vertex of a polyhedron. This is why a truncated icosahedron is formed with 12 pentagon cells that replace original vertices of the icosahedron. This number remains constant regardless of resolution or hierarchical level of a DGGS (Sahr et al., 2003). To mitigate these limitations researchers came up with ways to partition a hexagon cell with 3, 4 and 7 additional partial hexagons or aperture levels (Figure 2.6). The goal of these partition methods is to provide different ways of dividing a cell into smaller hexagons and to have equal area tiles completely covering a sphere on multiple levels of resolution. The downside, however, is the necessity to consider both shapes (i.e., hexagon and pentagon) during algorithm implementations and data handling processes for hexagon-based DGGS.

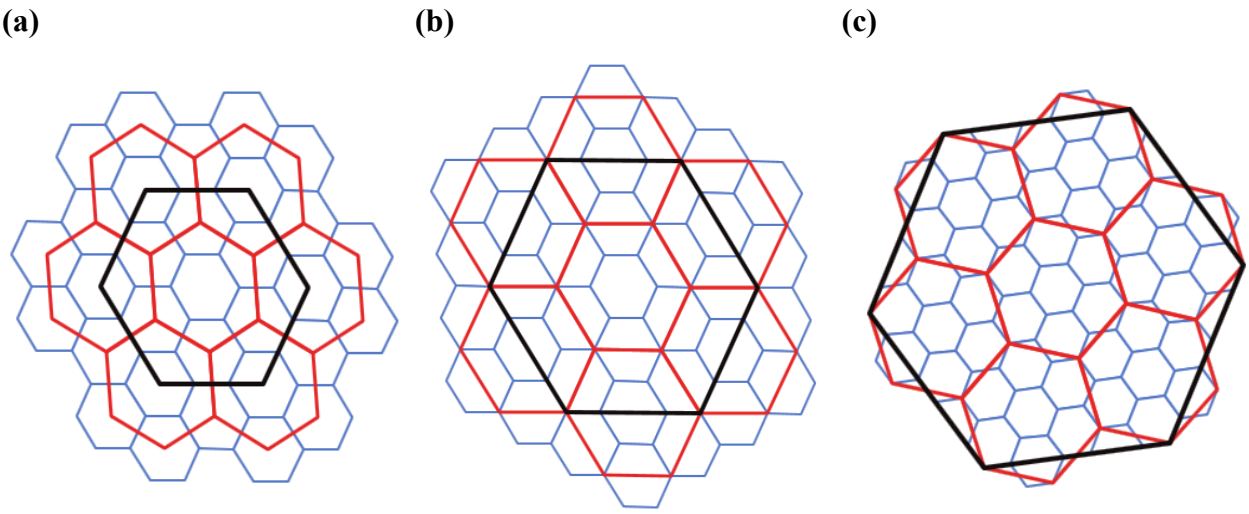


Figure 2.6. Hierarchical partition of space with (a) aperture 3, (b) aperture 4 and (c) aperture 7 hexagon levels. The hierarchy for aperture 3 and 4 were generated with the dggridR library (Barnes, 2016) and aperture 7 with the H3 library (Uber Technologies Inc., 2019).

2.2 OGC Abstract Specification

Even though the DGGS concept itself was in development since the 1980s (Dutton, 1984), the official formalization of DGGS occurred in 2017 with the release of an OGC abstract specification. At that time DGGS were defined as a data model framework for handling different data types via hierarchical tessellation of the globe. The system design must follow a set of criteria in order to support rapid access, storage, conversion and visualization of spatial information including core algorithms for data analysis. The emphasis of such systems was put on the repeatable representation of measurements as opposed to repeatable results for navigation applications (OGC, 2017). It was decided to review and summarise the core criteria defined by OGC to outline necessary components for a functional DGGS (Table 2.1).

Table 2.1. The following table summarizes core criteria that must be preserved in a DGGS, as outlined by OGC abstract specification (OGC, 2017).

Criteria	OGC Requirement	Criteria Statement
1	Model	This requirement outlines definition of DGGS conceptual data model which includes: • reference frame elements – structural items that provide a spatial reference system along with tessellation rules on which functional algorithms can operate internally (criteria 2-13); • functional algorithm elements – operations that allow location of cells, assignment and retrieval of data, run algebraic operations, and provide conversion methods for integration and use of DGGS externally (criteria 14-18).
2	Area	The grid system must guarantee domain completeness that covers the entire globe.
3	Overlap	Initial tessellation must ensure position uniqueness without any overlapping or underlapping cells.

Table 2.1. Continued.

Criteria	OGC Requirement	Criteria Statement
4	Tessellation sequence	Each global grid tessellation must form a joined system of hierarchical tessellations with progressively finer spatial resolution.
5	Area preservation	At each successive tessellation, the grid must preserve same domain completeness (criterion 2) and position uniqueness (criterion 3).
6	Shape	The cell shapes must be of simple polygon, which have the following properties: • Edges must only meet at vertices; • Only 2 edges can meet at each vertex; • Have the same number of edges and vertices; • Must always enclose a region with measurable area.
7	Equal area precision	At each level of grid refinement, equal area precision must be defined as the ratio of cell area uncertainty to cell area. The cell area uncertainty can vary across implementations, stability of equal area approximations, storage architecture or precision of reference frame parameters.
8	Equal area	At each level of grid refinement all cells must enclose equal areas within the specified level of precision (criterion 7). In the case of a grid system which consists of more than one cell geometry, each must retain equal area property separately and preserve a constant ratio between different shapes throughout the hierarchical structure.
9	Initial tessellation	At the base of each DGGS must be a polyhedron mapped on the Earth's surface to produce the initial tessellation, with each cell type representing specific surface area of equal size.

Table 2.1. Continued.

Criteria	OGC Requirement	Criteria Statement
10	Refinement	A DGGS must specify a method by which parent cells are tessellated into child cells. It is also recommended to indicate maximum number of refinements, maximum possible resolution, its limitation and precision used by the reference frame.
11	Addressing	A DGGS must use one or many spatial referencing methods to assign a unique spatial reference (i.e., index) to each DGGS cell.
12	Spatial reference	A DGGS must define a unique index to address each DGGS cell.
13	Cell centroid	The location of each DGGS cell must be defined by their centroids, which also allows for dual representation of a DGGS as cell grids and point lattices.
14	Quantization	A mechanism for assigning and retrieving of data to and from individual DGGS cells must be defined. Different methods for associating spatial data with DGGS cells may be used.
15	Cell navigation	Methods for hierarchical and neighborhood navigation across DGGS domain must be provided.
16	Spatial analysis	Methods for performing simple spatial analysis operations across DGGS domain must be provided.
17	Query	Methods for receiving, interpreting and processing data queries received from external client applications by internal DGGS quantization (criterion 14) and algebraic (criteria 15, 16) operations must be provided.
18	Broadcast	Methods for translating data query results from internal DGGS operations in formats that are suitable to broadcast to external client applications must be provided.

2.3 DGGS Implementations

There are several open source implementations of DGGS that can be used for practical applications of geospatial data analysis. In this section four open source implementations are reviewed in detail: R binding for dggridR (Barnes, 2016), JavaScript binding for H3 (Uber Technologies Inc., 2019) and Python bindings for OpenEAGGR (Riskaware Ltd., 2019) and S2 (S2Geometry, n.d.). The criteria for selecting the aforementioned libraries included: affiliations to large tech companies or academic institutions, a strong and active development team, claims for delivered functionality or unique properties (e.g., aperture, adjacency, congruency), an open source license, well-documented software, a large user support base, and popularity. While these criteria were considered during the selection process, not all of them were met exhaustively by each library mentioned here.

The dggridR library is an R wrapper for its DGGRID parent library, which is a Unix-based software package designed and developed by Kevin Sahr in C++ at Southern Oregon University (Sahr, 2018). The dggridR is an R package developed by Richard Barnes which allows construction of DGGRID grids within the statistical programming language R (Barnes, 2016). The software is fairly versatile since it is able to handle several grid systems. In particular, DGGS based on hexagons with 3, 4 and mixed 4-3 apertures, and triangles and diamonds with aperture 4 are supported (Sahr, 2018). This is the largest variation of shapes and apertures amongst the reviewed implementations. The library includes a large array of methods for working with vector geometry including native intersection operations between grid objects and shapefiles. However, it lacks methods for direct navigation across grid hierarchy and cell neighbours. On the other hand, the software is well-documented according to the global standards for packages in the R programming language.

H3 is another geospatial solution to hierarchical partition and spatial indexing on the sphere. Developed by Uber, H3 has been actively used as one of the tools for Uber's own operational needs, which includes dynamic optimization of ride prices and quantitative analysis of geographic data for decision making purposes, as well as visualization. Written natively in C, the H3 library has also a large selection of available bindings for other programming languages. These include but are not limited to C#, JavaScript, Python and R. It is worth noting that not all

bindings are at the same stage of development, and some are missing certain functionality. However, H3 undergoes a rapid development process and there is continuous enhancement of the project. The software has a number of built-in functions which permit conversion of points, lines and polygons into grids with unique spatial identifiers at various grid resolutions, as well as methods that permit moving across a grid system and identifying the neighbourhood of specific cells. One of the unique features of H3 is the integration of hexagon-based aperture 7 grid partitions, which allows for easier navigation through the hierarchy at the cost of the reduced precision of a cell and the area it covers (Brodsky, 2018).

Open Equal Area Global GRid (OpenEAGGR) is another software library implementation that models the Earth's surface as hierarchical layers of equal area tiles. One of the key differences noted for OpenEAGGR is that it claims to be OGC compliant (Riskaware Ltd., 2019), which means its development was completed with core OGC criteria in mind. OpenEAGGR was natively written in C++; however, other bindings for C, Java and Python, and integration with external applications (e.g., PostgreSQL and Elasticsearch) are also available to use. The grid partitioning method incorporates both aperture 3 hexagon and aperture 4 triangle hierarchical models with the ability to assign spatial data to individual cells (Bush, 2017; Riskaware Ltd., 2019). Unfortunately, in the last couple of years the development of this project has declined, which has led to lack of technical support and difficulty for use in practical applications. OpenEAGGR provides a rich variety of spatial analysis functions for operations with vector geometry and shape comparison.

S2 is yet another geospatial library that was developed and introduced by Google. S2 strives to model data onto a three-dimensional sphere using a hierarchical partition of squares with aperture 4 (S2Geometry, n.d.). S2 does not seem to be explicitly defined as DGGS, but happens to contain certain DGGS characteristics, such as discrete representation of space, spatial indexing, integration of basic vector type data, as well as the ability for spatial operations (e.g., intersection, union) and query (e.g., distance measurement, neighbor search) (Veach, 2017). In the past, the library seems to have been used by Google internally, such as in Google Maps web-based services (Perone, 2015), as well as externally for data modelling by Foursquare (Titlow, 2013), city studies by Sidewalk Labs (Kreiss, 2016) and even location-based game applications,

such as Pokémon GO (Zeroghan, 2019). The S2 library is written in C++, but also available to work with Go, Java and Python programming language bindings.

Some other notable DGGS include HEALPix, an octahedron-based implementation initially designed to measure cosmic microwave background anisotropies (Górski et al., 2018), and PYXIS, a web-based platform that provides a user interactive environment to search, process and share data on a virtual globe (PYXIS innovation inc., 2017). However, in this review, these models and implementations were not explored for the various reasons mentioned at the beginning of this section including limited available documentation and difficulty accessing the source code.

2.4 DGGS Analytics and Q-analysis

The development of DGGS began in the 1980s as an analytical framework for working with global terrain data (Dutton, 1984). This framework evolved into a spatial reference system and has been integrated as a data structure for consistent storage, use and analysis of spatial information globally, as well as its corresponding auxiliary information (e.g., attribute data) (OGC, 2017; Purss et al., 2016). In a general sense, DGGS can be described as a data warehouse which combines various data sources and serves as a global analytical system to provide better insight and understanding of the complex science systems, as well as simplify data handling procedures. Considering the fact that spatial operability or existence of functional algorithms are the core criteria for DGGS abstract specification (Table 2.1), new and refined methods for the data driven models must be constantly explored in order to account for the increasing data volumes and demands for understanding systems complexity (Miller & Goodchild, 2015).

On its own, the term complexity is not a self-explanatory concept, but it can be defined as a descriptive characteristic of systems with properties related to network connectivity, multiple subsystem dependencies, discrete dynamics, multilevel dynamics, globality and many other components not mentioned here (Johnson, 2014, p. 6). It appears that these descriptive characteristics are complementary to the definition of DGGS making it a subject of a complex systems science. Therefore, DGGS require formalization of more descriptive and rich methods

for data handling and insight across its domain when other conventional methods fall short in capturing system components, drivers and interactions explicitly (Johnson, 2014, pp. 7–9).

One of such methods suggested and applied in this study is called Q-analysis – a technique used to improve interpretation capability of complex systems by exploring their structural characteristics and connectivity between components (Atkin, 1974). It is interesting to discover that the need for DGGS has been indirectly recognized at the early stages of Q-analysis development, when the requirements for discretization of space and hierarchical structure of systems were acknowledged (Atkin, 1974, p. 117). Additionally, others have identified that the geographic division of space requires a scientific approach for proper understanding of its underlying structure (Johnson & Wanmali, 1981, p. 262), and the need for Q-analysis methodology to be extended for a wider variety of datasets and complex networks (Maletić & Zhao, 2017, p. 2). These and other incentives have been explored and addressed throughout this study.

2.5 Q-analysis Concepts

The concept of Q-analysis was introduced back in 1970s by the English mathematician Ronald Atkin from the University of Essex (Atkin, 1972, 1974; Atkin et al., 1971), with intention to enhance the interpretation perspective of complex systems and provide methods to describe their structural characteristics, connectivity and relationship between their various components. In the literature, Q-analysis is also commonly known as Polyhedral Dynamics and considered as a language of structure originating from the branch of mathematics referred to as algebraic topology (Atkin & Casti, 1977; Casti, 1975a, 1975b; Casti et al., 1979). In fact, much of Atkin's original thoughts on Q-analysis were based on the work of Dowker in relational algebra (Atkin, 1972; Dowker, 1952).

One of the reasons why the concept is described as "polyhedral" is explained by the fact that it can use a three-dimensional polyhedron to represent and visualize a multidimensional connectivity structure geometrically (Atkin, 1972, p. 152). While it is interesting to recognize that both DGGS and Q-analysis modelling take advantage of such geometric objects as

polyhedra, it is worth noting that they are not necessarily used in the same way nor directly related.

The main objective of Q-analysis is to provide a different point of view into complex data and system structures to study the connectivity of individual components (Atkin & Casti, 1977, p. 2). It is achieved by using rather a rich relational approach to examine connectivity instead of the functional approach, which is argued to constrain relations and lead to the loss of information, pattern and structure (Gould, 1980, p. 177).

2.5.1 Hierarchical Cover Sets

At the core of the Q-analysis theory are the notions of sets and cover sets (Atkin, 1980; Beaumont & Gatrell, 1982; Gould, 1980; Gould & Johnson, 1980). In a mathematical sense a set is a collection of elements, which is defined by some rules designed to determine whether an element belongs to the set or not. Such rules, as Gould states, must be carefully considered and be well-defined, due to potential membership ambiguity or existence at various generalization levels (Gould, 1980, p. 173). To clarify these statements, consider a political map of the world. Theoretically, it can be defined by the sets of sets at various hierarchical arrangements to match the order of countries' administrative divisions, such as sets of continents, countries, provinces, regions, counties, districts, municipalities, neighbourhoods, etc. Practically, however, such definitions can lead to the number of difficulties. For example, elements from different levels of aggregation cannot be members of the same set as they cover different spatial extents and might lead to the logical difficulties (Gould, 1980, p. 176). At the same time, a member of the municipalities set can belong to more than one district, which causes some ambiguity in membership definition.

This leads to an important concept in Q-analysis – a cover set, such that a set element which belongs to a higher and more general level of hierarchy can cover elements from different sets at the lower level. This is a very important property of Q-analysis since it provides a way to define the relationship and connectivity between elements in an unconstrained manner. Covers are opposite from the traditional tree-based hierarchy – a partition set, which tends to have a strict definition of sets and does not account for more than one element membership at the next

broader hierarchical level (Figure 2.7) (Atkin, 1980, p. 387). Gould states that all partitions can be considered as covers, but not otherwise (Gould, 1980, p. 174).

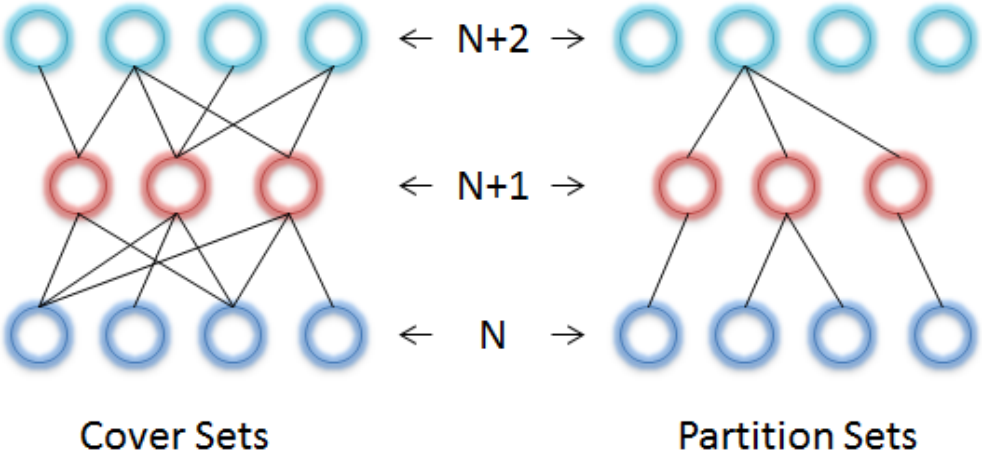


Figure 2.7. A comparison between the cover and partition approaches for defining relationship between elements across the hierarchy of sets on N, N+1 and N+2 levels. The figure also compares the richness of the connectivity structures of cover versus partition sets.

2.5.2 Relational Thinking

Another critical component of Q-analysis is the idea of relation, which can be formed between two arbitrary sets A and B as their Cartesian product at some fixed level of hierarchy N. The output is a new relation ($\lambda \subseteq A \times B$), where each element is a pair (a, b) such that $a \in A$ and $b \in B$, and λ is a relation which associates elements of A with elements of B based on some rule (e.g. intersection). Similarly, the relation ($\lambda^{-1} \subseteq B \times A$) can be represented as the inverse of λ. To illustrate the concept with example, let $A = \{a_1, a_2, a_3\}$ be a set of polygons and $B = \{b_1, b_2, b_3, b_4, b_5\}$ – a set of points, which are λ related if the element of A contains the element of B. Correspondingly, they are λ^{-1} related if the element of B is in the element of A. The Cartesian product between these two sets can be further stored in an array of pairs, such that:

$$\begin{aligned} A \times B = \{&(a_1, b_1), (a_1, b_2), (a_1, b_3), (a_1, b_4), (a_1, b_5), \\ &(a_2, b_1), (a_2, b_2), (a_2, b_3), (a_2, b_4), (a_2, b_5), \\ &(a_3, b_1), (a_3, b_2), (a_3, b_3), (a_3, b_4), (a_3, b_5)\} \end{aligned}$$

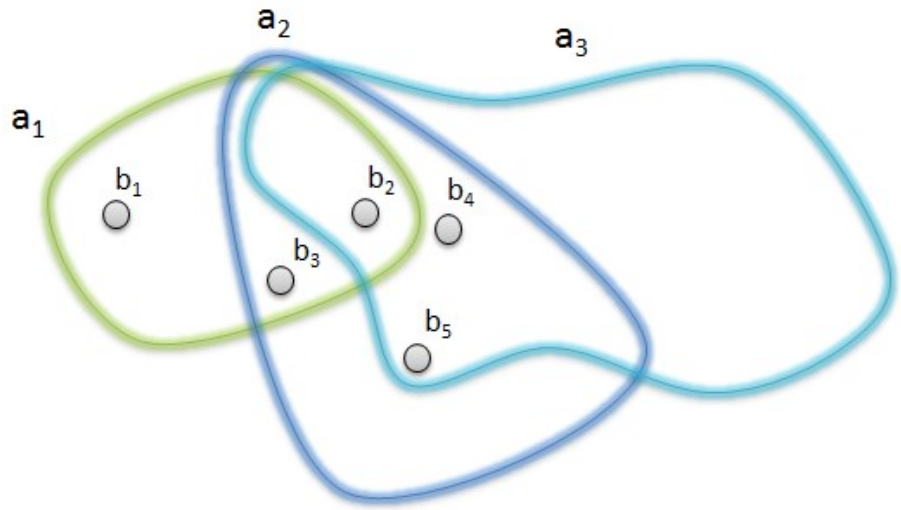
The λ relation, on the other hand, equals to the array of pairs which is a subset of the A x B product, where each element is a binary value indicating that the relation is true or equal to 1 (Figure 2.8), such that:

$$\lambda = \{(a_1, b_1), (a_1, b_2), (a_1, b_3), (a_2, b_2), (a_2, b_3), (a_2, b_4), (a_2, b_5), (a_3, b_2), (a_3, b_4), (a_3, b_5)\} \subseteq A \times B$$

This array may be represented as the incidence matrix Λ, and formally summarized via the following notation:

$$[\Lambda]_{i,j} = \begin{cases} 1, \text{if } (a_i, b_j) \in \lambda \\ 0, \text{if } (a_i, b_j) \notin \lambda \end{cases}$$

(a)



(b)

λ	b_1	b_2	b_3	b_4	b_5
a_1	1	1	1	0	0
a_2	0	1	1	1	1
a_3	0	1	0	1	1

(c)

λ^{-1}	a_1	a_2	a_3
b_1	1	0	0
b_2	1	1	1
b_3	1	1	0
b_4	0	1	1
b_5	0	1	1

Figure 2.8. The figure illustrates (a) sets A = {a1, a2, a3} (polygons) and B = {b1, b2, b3, b4, b5} (points), and their Cartesian product A x B reflected in the (b) incidence matrix λ and (c) transpose incidence matrix λ^{-1}. Value of 1 indicates that elements are related and 0 – they are not related.

At the same time, a conjugate relation ($\lambda^{-1} \subseteq B \times A$) can be generated by transposing the incidence matrix Λ which can be used to achieve a different view point of the data structure and provide additional analytical capabilities (Atkin, 1972, pp. 151–152, 1974, pp. 22–26; Atkin & Casti, 1977, pp. 3–5; Casti, 1975a, p. 3). The presented approach is the basis for the relational thinking, and considered superior and more adaptable to the variety of applications compared to the functional thinking (Gould, 1980, p. 176).

One of the examples for the notion of functional thinking is the well-known linear regression analysis, which is mostly concerned with estimating some variable Y with explanatory variable X via its linear approximation. Considering Y and X are two sets at some hierarchical level N a special form of the relation can be established via some function f, where each element in Y is related to only one element in X, such that $Y = f(X)$. In this sense, the relation is highly constrained since the function might not be capable to sufficiently account for the interconnectivity between all elements, which is often necessary to describe complex systems. As a result, valuable data structure information will be lost in the process. For this matter, a relation must account for more extensive connectivity definition between the elements of two sets, providing grounds for many-to-many mapping structure. Therefore, Gould stresses that it is critical to begin with defining data connectivity in a broader relational thinking to preserve as much original information as possible, since all functions are relations but not all relations are functions. In fact, it is relation that defines a structure (Gould, 1980, p. 179). To illustrate this concept the earlier example with polygons and points is appropriate to use here as well (Figure 2.9 cf. Figure 2.8).

Certainly, the whole concept of a relation which defines an incidence matrix is not limited to the binary representation only. In fact, it is appropriate to use other alternative data types, including real numbers or nominal data, which can be used to relate sets and form structure accordingly. In this case, the notion of slicing can be introduced in order to generate a binary matrix. In a nutshell, by choosing some slicing parameter θ, one can convert data into the familiar binary format (1 or 0) if a data value is greater than the θ parameter, for example (Atkin, 1974, pp. 35–38; Beaumont & Gatrell, 1982, pp. 20–23; Gould, 1980, pp. 180–181; Johnson, 1981a, p. 74). In this sense, Q-analysis is also rather flexible, as it is possible to define a separate set of θ parameters individually chosen for each element or element pair in the array.

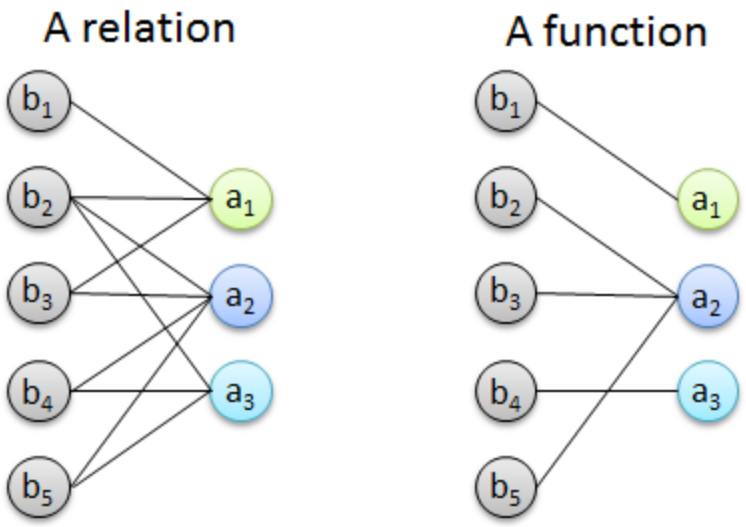


Figure 2.9. This figure illustrates the relational and functional approaches for capturing connectivity between A and B sets. It is clear that all connectivity outlined by the functional approach is also reflected by the relational approach. The opposite, however, is not true since much information has been omitted for the functional approach. The functional approach illustrated here does not follow a specific function, its connectivity is chosen arbitrary within the problem's domain limits (i.e., if B element is in A element).

2.5.3 Geometric Representation

In consideration of the core Q-analysis concepts it is appropriate to recognize that previously discussed relation can be also given a geometric form to explore its structure and connectivity patterns. For this reason, the use of simplicial complexes was found to be a suitable mathematical theory (Dowker, 1952), as it is capable of modelling complex network systems via multidimensional geometric objects – the polyhedra (Atkin, 1974, pp. 26–27).

Considering the earlier defined relation ($\lambda \subseteq A \times B$) along with its corresponding incidence matrix (Figure 2.8a, b), each polygon element of the set A can be represented as a common geometric feature (e.g., point, line and polygon) or a polyhedron known as simplex, while the elements of set B are the vertices of the simplicial complex. For example, elements a_1 and a_3 in set A are 2-dimensional simplices (i.e., polygon) defined by three vertices from the set B, whereas element a_2 is a 3-dimensional simplex (i.e., tetrahedron) defined by four vertices (Figure 2.10a cf. Figure 2.8a, b). Formally, these simplices can be symbolized as $\sigma_2(a_1) = \langle b_1, b_2, b_3 \rangle$,

$\sigma_3(a_2) = \langle b_2, b_3, b_4, b_5 \rangle$ and $\sigma_2(a_3) = \langle b_2, b_4, b_5 \rangle$ in the $K_A(B; \lambda)$ simplicial complex. Note, the subscript of the simplex notation indicates its dimensionality, and it is always one unit less than the total count of vertices a particular simplex has.

The representation of the ($\lambda^{-1} \subseteq B \times A$) relation carries the same principle but with the set B elements being simplices and set A elements being their vertices this time (Figure 2.10b cf. Figure 2.8a, c). In the literature, the transposed relation is commonly known as the conjugate relation which respectively forms the conjugate simplicial complex $K_B(A; \lambda^{-1})$ (Atkin & Casti, 1977, p. 27).

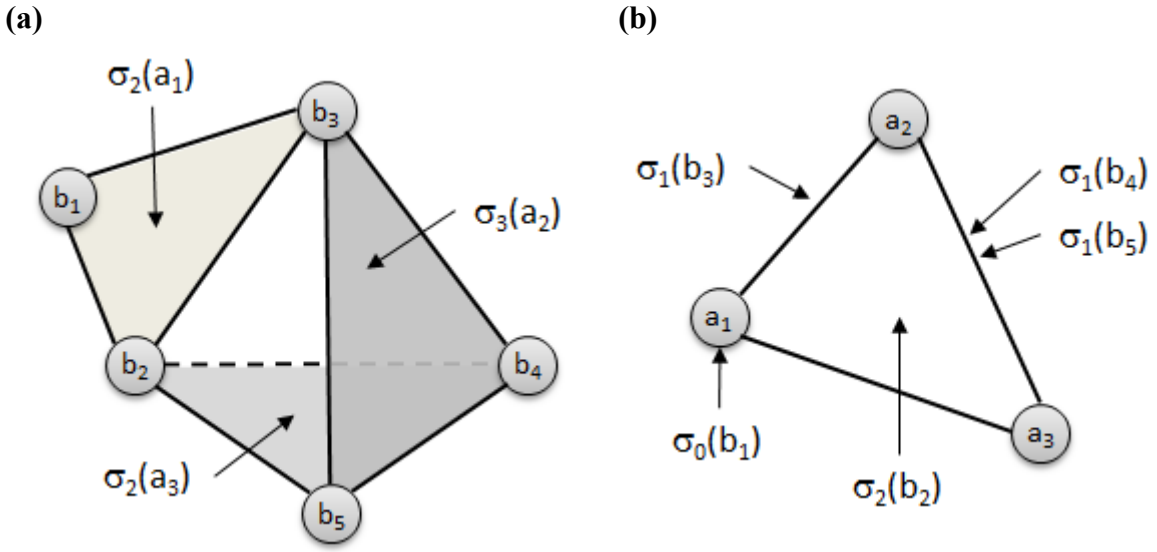


Figure 2.10. Geometric representation of (a) $K_A(B; \lambda)$ and (b) $K_A(A; \lambda^{-1})$ simplicial complexes. Each simplex is marked with a σ notation and represented by corresponding vertices which together form a geometric space when combined.

One must also keep in mind that dimensionality of greater than three is impossible to visualize directly. Therefore, corresponding representation is limited to the three-dimensional space for visualizing the high dimensional connectivity between vertices. On the other hand, this outlines a significant advantage of Q-analysis and algebraic topology as an effective approach for expressing highly connected, interacting and multidimensional data to alleviate the constrains of the graphical representation (Atkin, 1972, p. 166; Gould, 1980, p. 180).

With this understanding it is now possible to have a closer look into the structural characteristics of both simplicial complexes to explore their connectivity features. For instance, given $\sigma_2(a_1) = \langle b_1, b_2, b_3 \rangle$ and $\sigma_3(a_2) = \langle b_2, b_3, b_4, b_5 \rangle$ simplices it is clear that they are both connected via the $\langle b_2, b_3 \rangle$ vertices (i.e., line). This implies that $\sigma_2(a_1)$ and $\sigma_3(a_2)$ are related to each other in terms of sharing common features, properties or characteristics, which can be transmitted to other connected simplices through the network of connected chains. Formally this is defined such that $\langle b_2, b_3 \rangle$ is the face shared between $\sigma_2(a_1)$ and $\sigma_3(a_2)$, which also makes simplices q-connected or 1-near via the 1-dimensional simplex (i.e., line) (Atkin, 1972, p. 155, 1974, p. 29; Beaumont & Gatrell, 1982, p. 16). Surely, as the structure gets more complicated such connectivity can form a chain of connected faces representing communication pattern of joined simplices in various dimensions across a simplicial complex. In fact, it is this q-connectivity which gets explored by Q-analysis to find distinct connected components or groups of simplices in a simplicial complex (Casti, 1975b, pp. 7–8).

The direct implications of such analysis are further revealed within the scope of this study; nonetheless, it is already possible to see how connectivity structure can be obtained and explored via a simplicial complex and the use of Q-analysis. To reiterate, a simplicial complex is a multidimensional geometric space composed of multidimensional simplices and vertices which define and connect these simplices together. When connected, the simplices form a structure which further gets explored by Q-analysis.

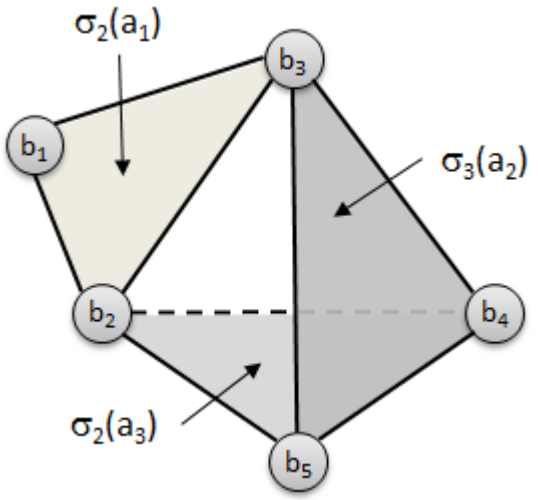
2.6 Hierarchical Backcloth and Traffic

The theoretical rationale of Q-analysis is further based on the concepts of backcloth and traffic. By taking into account the earlier defined $K_A(B; \lambda)$ and $K_B(A; \lambda^{-1})$ simplicial complexes, it can be said that both of them are used to describe a multidimensional geometric space called a backcloth (Atkin, 1972, p. 165). This geometric structure of the backcloth provides grounds where some phenomena or activities can take place and has a general name – traffic (Atkin, 1980, p. 384). Similarly to the backcloth, traffic is also defined by a vertex set and represented geometrically. The vertex set B, in this case, is replaced with a new vertex set C to characterize some activity (e.g., air pollution, car sales, life expectancy), such that new relations ($\mu \subseteq A \times C$)

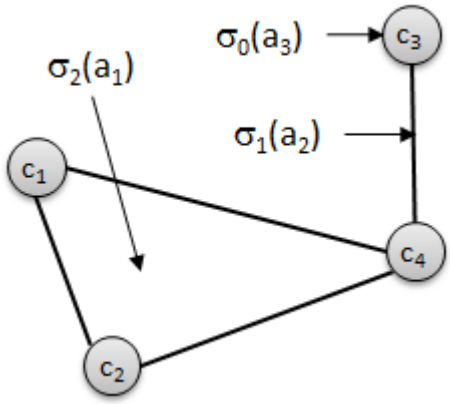
and ($\mu^{-1} \subseteq C \times A$) along with the corresponding $K_A(C; \mu)$ and $K_C(A; \mu^{-1})$ simplicial complexes are formed (Johnson, 1983b, p. 470).

Atkin has further identified that there is also a strong relationship between geometric composition of backcloth and traffic, such that backcloth can either permit or restrict traffic based on its dimensionality (Atkin, 1972, pp. 163–165, 1980, pp. 384–386). In practice, this is viewed as a mapping of traffic-based simplices, known as cosimplices, onto the vertices of the backcloth and vice versa (Atkin, 1972, pp. 154, 163; Beaumont & Gatrell, 1982, p. 24; Gould & Johnson, 1980, p. 184). Hence the notion that backcloth can either permit or restrict traffic since absence of backcloth connectivity (i.e., which is defined by vertices) will make it impossible for traffic simplices (i.e., cosimplices) to be attached to the backcloth (Figure 2.11). In other words, traffic requires certain backcloth connectivity structure in order to exist on such structure. In this sense Q-analysis presents a noteworthy property, the ability to analyse high dimensional data (i.e., traffic) in conjunction with nonetheless high dimensional space (i.e., backcloth).

(a)



(b)



(c)

λ	b_1	b_2	b_3	b_4	b_5
a_1	1	1	1	0	0
a_2	0	1	1	1	1
a_3	0	1	0	1	1

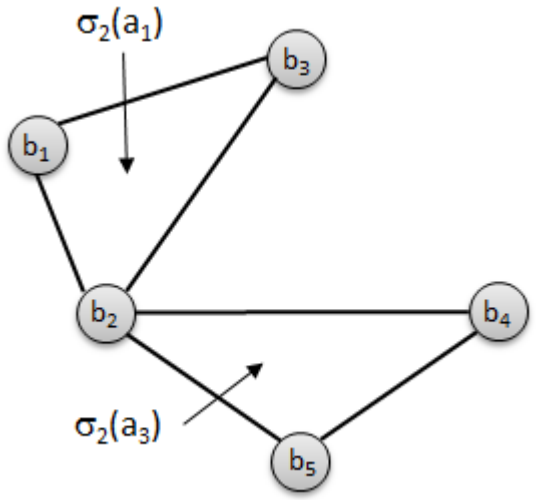
(d)

μ	c_1	c_2	c_3	c_4
a_1	1	1	0	1
a_2	0	0	1	1
a_3	0	0	1	0

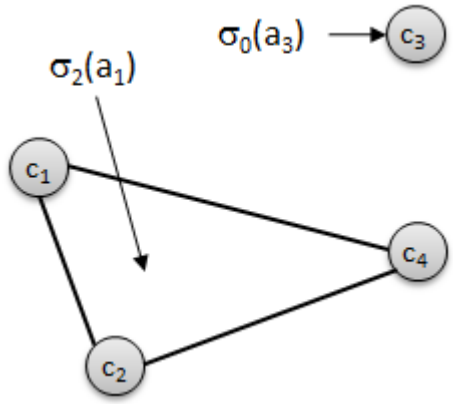
Figure 2.11. Illustration of (d) arbitrary traffic related incidence matrix μ mapped onto (c) the backcloth related incidence matrix λ, such that every element in set A have associated vertices that form (a) backcloth $K_A(B; \lambda)$ and (b) traffic $K_A(C; \mu)$ simplicial complexes.

There are two possible ways to observe changes in the traffic geometry: by observing changes to the traffic structure while the backcloth remains constant, and by observing changes to the traffic as a result of some structural changes in the backcloth. In the literature they are also referred as Newtonian and Einsteinian views respectively (Gould, 1980, pp. 182–183). For example, in the earlier case, imagine the backcloth being formed as a result of a relation between residential areas (set A) and location of some government regulated electric meters (set B). The traffic structure, in this case, can be defined based on the measure of electricity consumption (set C) at each residential area. Supposedly if the backcloth structure remains unchanged then it is safe to assume that energy consumption is a subject to change depending on time of the day, week or even season. Thus, changes in traffic are caused by some external factors or forces, called t-forces (Gould, 1980, p. 182). In the latter case, imagine that the backcloth structure has been modified, such that residential area a_2 in the set A is no longer part of the backcloth. The geometry of the backcloth will be changed, which also result in the structural changes of traffic (Figure 2.12 cf. Figure 2.11).

(a)



(b)



(c)

λ	b_1	b_2	b_3	b_4	b_5
a_1	1	1	1	0	0
a_3	0	1	0	1	1

(d)

μ	c_1	c_2	c_3	c_4
a_1	1	1	0	1
a_3	0	0	1	0

Figure 2.12. Example of (b) changes in traffic structure as a result of (a) structural changes in the backcloth. The figure also reflects corresponding changes in the (c) backcloth λ and (d) traffic μ incidence matrices.

A concluding consideration for performing a valid Q-analysis is the importance of establishing a proper hierarchical backcloth and traffic schemes, such that elements at one hierarchical level relate to the elements of another hierarchical level in the non-partition, cover-based manner. This requirement has a close connection with the notion of cover sets (see Section 2.5.1), except it is applied on both backcloth and traffic structures comprehensively. In other words, a hierarchical distinction must be made clear for both pieces of geometry, such that backcloth or traffic sets at higher level of generalization (e.g., N+1 or M+1) must cover those at the lower level (e.g., N or M). To identify and make such distinction is the critical part and subject to a successful Q-analysis at various levels of hierarchy (Gould & Johnson, 1980, p. 181; Johnson, 1983a, p. 342). A significant contribution and advancement of these theoretical developments were made by Johnson and Gould with the extensive study on the structure of hierarchical backcloth and traffic (Johnson, 1983a, 1983b) and the work on structural and hierarchical complexity in the context of television programme and policy (Gould & Johnson, 1980; Johnson, 1978). Their theoretical and application-based work has outlined important concepts and rationale for defining a hierarchical relation between well-defined sets, as well as mapping of hierarchical traffic onto the hierarchical backcloth.

2.7 Q-analysis Applications

Historically Q-analysis has been applied in a large variety of applications, including such diverse areas as chess (Atkin, 1974, pp. 46–64), soccer (Gould & Gatrell, 1979), lake ecosystems (Casti et al., 1979), transportation (Johnson, 1981b, 1986), market systems (Johnson & Wanmali, 1981), economic systems (Sonis & Hewings, 2000), entrepreneurial networks (Bliemel et al., 2014), big data and complexity (Maletić & Zhao, 2017), and even shopping mall movement patterns (Omer & Goldblatt, 2017). Yet amongst the most notable applications which had a significant impact on the direction of this research were studies in urban analysis of a town (Atkin, 1974, pp. 105–140), television programme complexity (Gould & Johnson, 1980) and agriculture and communication (Gaspar & Gould, 1981).

In his urban analysis work Atkin has recognized the limitation of a two-dimensional map to obtain and accurately describe physical properties and functioning of a town or community. For

this reason, the study suggested to cover the town area with the set of lozenges (i.e., areal units) of arbitrary size and shape in order to cover important features of the town (e.g., trading centers, residential property, amenities, etc.). At the same time, it was also stressed that these areas must be small enough to aggregate into the larger scale areas, if needed. The association between lozenges and town features created a basis for the urban community in the area and formed a backcloth. The connectivity analysis of the urban community structure demonstrated that Q-analysis can be used as a decision making tool for urban planning and can promote development of a well-balanced and functioning community, including visual appearance and aesthetics of the town (Atkin, 1974, pp. 105–140). Similar research in this field have also identified the usefulness of Q-analysis to interpret complex data structures and gain better understanding of urban systems (Gatrell, 1981; Spooner & Batty, 1981).

A study on television policy and complexity has clearly identified the difficulty of having the well-defined television programme structure at various generalization levels. The reason of defining such complex structure was meant to promote scientific and policy-making development in this research area. As a result, a clear distinction of program content and its categorical arrangement (e.g., education, entertainment, miscellaneous) was specified. The study recognized distinct hierarchical structures for both backcloth and traffic and addressed the use of cover sets accordingly. It also demonstrated practically how backcloth can be mapped with traffic by associating a backcloth simplex with a traffic simplex. Such mappings were defined as hierarchical, which means that a backcloth simplex at level N was possible to map with a traffic simplex at any hierarchical level, such as N+2 for example (Gould & Johnson, 1980). These and other developments of this study outlined a valuable theoretical background for examining other large-scale hierarchical structures and properly accounted for their descriptive components and features.

Lastly, another great study in the field of agriculture and communication which used Q-analysis has significant interest and practical value in the context of this thesis. The study integrated the familiar concepts of cover sets and hierarchies of backcloth and traffic elements. In particular, the backcloth was formed as a relation between a set of farmers and their personal and property characteristics (e.g., literacy, presence of irrigation, access to information), which created necessary geometric structure to support traffic. Traffic, in this case, was formed as a relation

between farmers and their productions in various fields of agriculture (e.g., livestock, orchards, grains). Both backcloth and traffic were carefully defined in a hierarchy of sets, such that elements of "age, literacy, education" at level N, became "personal characteristics" at level N+1, for example. What was unique about this study is a consideration for the communication backcloth between farmers who were seeking for advice and those who were prepared to give advice, as well as comparison between backcloths formed of only younger and older farmers. The study revealed that geometric structures of younger and older farmers were noticeably different. In addition, the difference between type of agricultural occupation they practice and level of education they have was also observed. The study also combined both backcloth and traffic elements in a single geometric structure to explore some distinct characteristics of farming in the study area (Gaspar & Gould, 1981); however, some procedural aspects of merging two structures together remained unclear and require additional clarification.

As a concluding remark, it should be recognized that Q-analysis theory and methodology does not stop here, but continues to evolve into more explicit and prominent theoretical methods to meet the needs for the complex data organization, management and analysis. This has led to advancement and theoretical formalism of hypernetworks, which are meant to provide interpretation tools for the multilevel backcloth-traffic systems, unambiguous definition of such multilevel systems and relational mapping, as well as comprehensive use of simplices, simplicial complexes and Q-analysis. A significant advancement in this area was performed by Johnson and his work on hypernetworks in complex systems (Johnson, 2014). At the same time, it is also important to mention that there are other aspects of Q-analysis theory which were not covered in this review as they falls outside of the study scope, these were such concepts as eccentricity or dynamics (i.e., space time analysis).

Chapter 3 Methodology

The main goal of the following chapter is to outline a detailed methodology for using Q-analysis in a DGGS for the broad variety of applications. The chapter begins with exploring DGGS in a general context, its operational capability and industry standards in Section 3.1. Section 3.2 formalizes the use of DGGS as one complex system, the implications of which can be further interpreted mathematically via corresponding Q-analysis. Section 3.3 explores importance of cover sets and their practical application in DGGS to form a connected structure. The chapter ends with Section 3.4 outlining the final steps necessary for the successful use of Q-analysis as well as its implementation in JavaScript.

3.1 Exploring DGGS

The ability to discretize geospatial data for analytical purposes and presenting the produced outcome via available DGGS implementations is an important part of effective communication of results. Due to the differences and variation in DGGS functional availability to perform such operations, a method for user-friendly handling of spatial data was explored. For the most part available DGGS implementations use their own data type structures for storing data. Often, these formats are not straightforward to work with for dynamically modifying and assigning multiple data values. This section focuses on exploring how the integration of spatial data for various types can be achieved using the GeoJSON data format for embedding, processing and visualizing of geographic data, due to its simplicity, versatility and universality (Butler et al., 2016).

3.1.1 Embedded DGGS

In particular, the general procedure of such an approach is theoretically scalable to any DGGS implementation as long as two essential functions are provided: methods for converting geographic data into grid cells and extracting coordinate information for each of their vertices. For example, methods for converting a coordinate location of a point into a hexagon grid cell to obtain the coordinates of its vertices must be provided. The acquired geometry can be used to

construct a GeoJSON polygon feature following its standard encoding format to populate its coordinate information. During this stage additional data can be embedded in the feature type or the current one can be modified. The ID value encoded in the feature can be directly read and interpreted by DGGS libraries in order to determine each cell's resolution, search for its neighbours, navigate through the hierarchy or used in spatial analysis, as long as these two methods are supported (Figure 3.1).

```
3   {
4   "type": "FeatureCollection",
5   "features": [{
6         "geometry": {
7               "type": "Polygon",
8               "coordinates": [[
9                     [longitude1, latitude1],
10                    [longitude2, latitude2],
11                    [longitude3, latitude3],
12                    [longitude4, latitude4],
13                    [longitude5, latitude5],
14                    [longitude6, latitude6],
15                    [longitude1, latitude1]
16              ]]
17        },
18        "type": "Feature",
19        "properties": {
20        "fill": "#298178",
21        "fill-opacity": 0.5
22        "ID": 1
23        "city": "valueA"
24        "population": "valueB"
25        "land_type": "valueC"
26        }
27  }]
28  }
```

Figure 3.1. A sample GeoJSON encoding format of a hexagon feature type. Note the potential use of the "coordinates" field on lines 8-16, and "properties" field on lines 19-26 to include necessary information associated with the hexagon cell (e.g., city name, city population).

To provide a concrete example, the study further considered implementation of an intersection algorithm using only the primary functionality for cell generation and geometry extraction. Two geometry types of line and polygon representing a street network and a water feature respectively

were indexed and integrated into a DGGS hierarchy. Using their generated cell IDs and land cover types, the intersection region between two spatial features was determined and assigned a different value to reflect the intersection property, while also preserving their original attribute values. In the meantime, all geographic features along with their attribute information were written to a GeoJSON file format as noted earlier. Once all data were stored in the GeoJSON file format, the intersection output can be easily visualized using a 3rd party desktop application (e.g., QGIS) or a web-based service (e.g., Leaflet), or even converted into other file formats, such as SHP, KML, WKT or CSV, if necessary. The presented approach is a clear illustration of how basic data integration, analysis and visualization functionality can be achieved if such methods are not directly provided by a DGGS. The presented methodology was implemented and applied for the duration of this study.

3.1.2 Scalability

Although the proposed methodology for data handling and display might be useful and convenient to apply for individual use cases or even small datasets, it is crucial to consider application with much larger datasets for particular DGGS implementations, aperture levels and shapes. With this in mind, individual aperture-shape pair combinations were considered for this analysis based on library implementation and pair availability. In particular, aperture 3 hexagon (A3H), aperture 4 hexagon (A4H), aperture 4 triangle (A4T) and aperture 4 diamond (A4D) pairs were examined for dggridR library. However, only A3H, A4T pairs were tested for OpenEAGGR, aperture 7 hexagon (A7H) for H3, and aperture 4 square (A4S) for S2, due to their specific approaches to DGGS implementation. It is important to mention that hexagon with mixed 4-3 aperture implementation was not considered for dggridR library, as it generates hierarchical sequence of grids which alternates between both apertures (Sahr, 2018). Thus, given a fixed resolution the same objective can be achieved by using the aperture 3 or 4 only.

The performance of selected DGGS software packages was evaluated by varying dataset size and performing a commonly required task. Randomly generated points were used to simulate a baseline point spatial dataset at an arbitrary fixed scale. This process was then repeated five separate times and each sample was converted to its DGGS cell addresses. The first sample case

begins with 100 random features, and subsequent sample sizes were increased by a factor of 10 ending with 1,000,000 sample points in total. To reiterate, the analysis begins by sampling random latitude-longitude locations across the globe in order to generate, extract, store and save cell geometry to a GeoJSON file. The whole process was timed in seconds and compared with the performance of the four DGGS implementations considered in this study. While it is acknowledged that performance might vary significantly for different libraries according to multiple factors, such as algorithm efficiency, input/output operations, development environment, language binding or even choice of the cell shape and solid polyhedron as a base of a DGGS, the goal of the test is to observe the behaviour and performance under increasing data volumes. One of the implications of exhaustive partitioning of space provided by DGGS is the fact that datasets can be extremely large, so scalability becomes a critical requirement for operational use. The results presented in Chapter 5 indicate whether a library is able to handle large datasets, display its overall performance, and, if possible, identify major differences and similarities amongst the libraries.

3.1.3 OGC Compliance

The final component of this section for embedding DGGS and finding linkages to GIS applications includes DGGS state of development evaluations and their OGC compliance. For this matter it was decided to review the released OGC abstract specification (OGC, 2017) in order to compare its defined criteria (Table 2.1) with four selected DGGS implementations: dggridR, H3, OpenEAGGR and S2. Each criterion was explored individually for all libraries in order to determine whether it had been met by a particular DGGS implementation. The analysis also reflects on the overall compatibility of OGC abstract specification and suggests possible ways for their further refinements and enhancements.

3.2 Understanding Complexity

The process of formalizing spatial information in a standardized data infrastructure (i.e., DGGS) requires procedures capable of integrating various data sources while preserving their complex

relational structures across the hierarchy of distributed systems. These data sources can consist of many descriptive components and interactions, the retention of which are important in data driven analyses (Miller & Goodchild, 2015). In fact, these interactions can further form a multidimensional relational structure (i.e., multilevel backcloth-traffic system), the complexity of which can be difficult to interpret without utilizing tools and methods of algebraic topology (i.e., Q-analysis) (Gould, 1980). The following methodology attempts to account for the structural complexity of such spatial information by integrating the concept of relation (see Section 2.5.2) into the definition of hierarchical connectivity within DGGS data infrastructure. For this matter, a relation between DGGS cells and spatial features must be formed to define how components are held together in a hierarchical DGGS data structure.

3.2.1 Matrix Construction

The process of encoding the information requires explicit definition of the well-defined finite sets and their association in a binary matrix form as a Cartesian product. In the DGGS context, the primary set must consist of DGGS cells as the elements representing some physical space or a study area, while the elements of an associated set must contain subject matter features and its descriptive characteristics (e.g., spatial or attribute data of the subject matter) related to the primary DGGS set. At this stage, the methodology is restricted to a single level of generalization, which means that hierarchical arrangement of sets is not considered, and all members are assumed to have a similar level of generalization (i.e., same coverage and descriptive domain). Thus, a binary relation is established in order to indicate the presence or absence of association between a pair of elements from the chosen sets. That is, if a DGGS cell is described by an element from the subject matter set the relation equals to 1, otherwise the relation is 0. In this case, DGGS cells are rows and subject matter or its descriptive characteristics are columns in a corresponding binary matrix. Columns, on the other hand, can be also classified into groups to represent various geographic features (e.g., points, lines, polygons or raster scenes) in an organized manner, if required (Table 3.1).

Table 3.1. Matrix construction of a binary relation between DGGS cells (rows) and feature elements (columns) to indicate association between each pair, where 0 or 1 specify absence or presence of the relation respectively.

	Point Features			Line Features			Polygon Features			Raster Scenes		
	X_1	⋯	X_n	Y_1	⋯	Y_n	Z_1	⋯	Z_n	R_1	⋯	R_n
C_1	0/1		0/1	0/1		0/1	0/1		0/1	0/1		0/1
C_2	0/1		0/1	0/1		0/1	0/1		0/1	0/1		0/1
C_3	0/1		0/1	0/1		0/1	0/1		0/1	0/1		0/1
⋮												
C_n	0/1		0/1	0/1		0/1	0/1		0/1	0/1		0/1

While the binary relation is necessary for the further use of Q-analysis, the whole concept is not limited to the binary representation only. Other more practical and familiar data types of real or nominal values are encouraged to be used as well. These data types can be also converted into a binary relation matrix via the concept of slicing and corresponding slicing parameter θ (i.e., condition statement) instead (see Section 2.5.2). The slicing parameter θ, in this case, can be defined as a condition upon which the relation is evaluated to true or false (i.e., 1 or 0). For example, a DGGS cell can be related to some raster scene via a raw pixel value, which in turn can be converted into a binary value according to defined condition, such as identifying pixels with Normalized Difference Vegetation Index (NDVI) greater than zero (i.e., $\theta > 0$). Thus, a binary relation is generated according to whether expression is true or false (i.e., 1 or 0). The work with rater data type was not explored in this research; however, the presented methodology provides basis for its potential use in the future.

At the same time, similar attention may be given to the inverse binary matrix, which is generated by transposing the initial binary matrix. In this sense, the rows become columns and columns become rows, such that each feature type or spatial attribute is defined by the set of DGGS cells (Table 3.2 cf. Table 3.1). Matrix transposition provides opportunity to observe connectivity structure from a data point of view, as opposed to the spatial point of view represented by the direct binary matrix (Atkin, 1972, pp. 157–158, 1974, pp. 32–33).

Table 3.2. A transposed binary relation, with initial rows and columns exchanged. The process preserves the original association parameters of each pair.

	C_1	C_2	C_3	⋯	C_n
X_1	0/1	0/1	0/1		0/1
⋮					
X_n	0/1	0/1	0/1		0/1
Y_1	0/1	0/1	0/1		0/1
⋮					
Y_n	0/1	0/1	0/1		0/1
Z_1	0/1	0/1	0/1		0/1
⋮					
Z_n	0/1	0/1	0/1		0/1
R_1	0/1	0/1	0/1		0/1
⋮					
R_n	0/1	0/1	0/1		0/1

In a general sense, the process extends the ability of capturing and retaining multidimensional complexity of DGGS infrastructure with spatial information embedded into one standardized relationship representation. The method also facilitates discrete way of reflecting relation structure of various features and their corresponding sets for further analysis and visualization.

3.2.2 Geometric Visualization

The binary relationship between sets can also take a geometric form, the visualization of which attempts to display the complexity of the given data. The procedure takes advantage of a simplicial complex, a mathematical concept which defines elements of one set (i.e., DGGS cells) as simplices, and the related sets as simplex vertices. These simplices are geometric objects which are described by the vertices connected together to reflect connectivity found in the corresponding binary matrix. That is, simplex vertices are connected if they are related to a single row element (i.e., DGGS cell). Number of the connected vertices defines simplex dimensionality, such that a vertex by itself is a 0-dimensional simplex σ_0 (i.e., point), two connected vertices – 1-dimensional simplex σ_1 (i.e., line), three connected vertices – 2-dimesional simplex σ_2 (i.e., polygon), four connected vertices – 3-dimensional simplex σ_3 (i.e.,

tetrahedron), and so on. Thus, dimensionality of a simplex is equal to the total number of vertices minus one (Figure 3.2). While challenges of visualizing such connectivity structure due to the limited human perception to the three-dimensional space remain, algebraically there are no such limits and connectivity is preserved regardless of the dimension. Certainly, such geometric realizations must also account for possible chain connectivity, in cases when vertices are shared amongst various simplices. In this case, the whole structure is represented as a polyhedron which combines simplices in a one multidimensional structure called a simplicial complex to display connectivity of the well-defined sets (Figure 3.2).

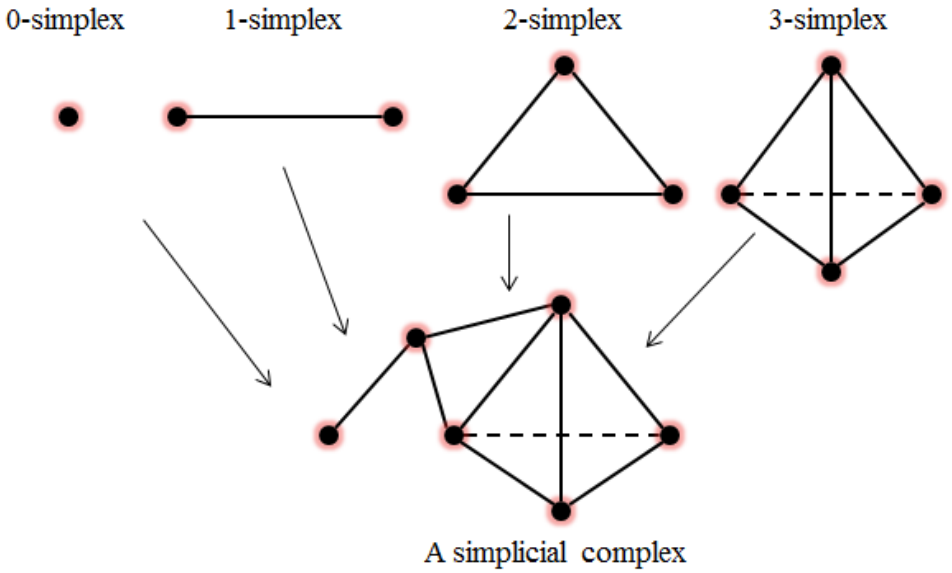


Figure 3.2. Geometric visualization of a simplicial complex as a chain of simplices with various dimensions combined into one multidimensional structure. The vertex connectivity between simplices implies sharing of similar features or characteristics.

At the same time, following the same procedure, the geometric realization can be applied to the transposed binary matrix as well. The corresponding geometric structure is called a conjugate simplicial complex, and provides additional point of view of the system complexity to gain better insight and understanding. Both simplicial complex and its conjugate make up a backcloth, a generic term which describes a relatively fixed multidimensional system structure. This structure is also able to carry traffic, which is a simplicial complex itself describing system characteristics. The existence of traffic is directly dependent on the geometric structure of the backcloth.

3.3 Q-analysis and DGGS

The process of deriving well-defined sets as the matter of managing and formalizing spatial information requires additional considerations for their descriptive domains, since it is usual for data to have descriptive characteristics at various levels of generality or hierarchy. A clear distinction of these hierarchies is the essential part, and must be properly established prior the analysis in order to avoid logical paradoxes (Gould & Johnson, 1980). Such hierarchical arrangement of data is further described by the concept of cover sets (see Section 2.5.1). That is, sets at the higher level of generalization covers sets at the lower level. The following methodology attempts to build upon the notion of cover sets and extend its application to the DGGS framework. The process requires proper articulation and logical arrangements of sets while being complementary with the embedded hierarchy of DGGS and Q-analysis theory. This also includes formalization of the established hierarchical backcloth and traffic, as the product of cover sets (see Section 2.6). Both notions are subject to having distinct hierarchical structures, the connectivity of which gets further examined by Q-analysis.

3.3.1 Defining Hierarchical Backcloth

In the theory of Q-analysis a hierarchical backcloth is understood as a relation between well-defined set elements at one hierarchical level to the set elements at another hierarchical level in the non-partition, cover-based way. Considering the abstract specification for DGGS (see Section 2.2), the idea of cover sets can be naturally established using the concepts of hierarchical tessellation and aperture of cells. That is, a clear relationship between sets of cells at various hierarchical levels can be defined via built-in methods to determine whether a cell is at the same, higher or lower hierarchical level when compared with other cells in DGGS hierarchy.

In particular, a sequence of DGGS cells can be generated on some fixed level N, where N is the numeric representation of an arbitrary real-world scale. Thus, a hierarchy of DGGS cells can be constructed to model a physical space, for example N+0 being the street level, N+1 – neighbourhood level, N+2 – municipality level, N+3 – district level, and so on all the way up to the N+k level, where k is the world level scale equivalent to the initial tessellation of the DGGS

truncated icosahedron (see Section 2.1.1). Similarly, the hierarchy can go into a negative direction too, modelling finer resolution physical features, such as N-1, N-2 and N-3 levels being buildings, rooms and house specific objects respectively. For practical purposes it might not be necessary to integrate multiple hierarchical levels at once, but the essential levels must be established according to application needs or data availability. The relationship between such hierarchical levels of cells is complementary with the definition of cover sets, such that higher level cells cover the lower level cells. Besides, set membership and relation between individual cell elements can be determined according to the conventional rules for cell tessellation and aperture (see Section 2.1.3).

At the same time, it is important to stress that aperture choice may have a significant impact on how cover sets are defined, since cells can be shared across hierarchy. For example, with a hexagon-based DGGS a single cell element can be covered by 3, 2 and 1 cell elements at the next hierarchical level, according to the definition of hexagon apertures 3, 4 and 7 respectively (Figure 2.6). A special mention should be made about aperture 7 hexagons, since their hierarchical structure might seem as partitions (Figure 2.7). That is, according to the definition of aperture 7, a hexagon at one level can be a member of only one hexagon at the next hierarchical level (Uber Technologies Inc., 2019). Nonetheless, it is still consistent with the definition of covers, since "all partitions are covers, but not all covers are partitions" (Gould, 1980, p. 174). The statement is also true for the congruent shapes, such as squares, diamonds and triangles, making all known DGGS kinds consistent with the concept of cover sets.

Consequently, given the sequence of DGGS cells at various hierarchical levels a simplicial complex $K_C^{N+k}(F, \lambda^{N+k})$ can be generated for each corresponding level, where K indicates a complex; N+k – hierarchical level for k = 1, 2, 3,…n; C – a set of DGGS cells; F – a feature set or a subject matter related to some case study (e.g., points, lines, polygons and raster scenes); λ – relation which defines association rules between C and F sets used for the binary matrix construction. Note that definition of the λ relation is hierarchy dependent to the set C instead of set F, since DGGS cells vary in size and spatial extent across hierarchy, whereas set F remains constant. Therefore, associative rules must be defined for each hierarchical level to reflect corresponding changes of the spatial extent in DGGS cells. The formal definition of this

mathematical concept was described in earlier studies, thus similar hierarchical consistency was preserved here as well (Gould & Johnson, 1980, pp. 182–183; Johnson, 1983a, pp. 355–357).

In summary, each simplicial complex consists of simplices (i.e., DGGS cells), which are described by their vertices of the associated subject matter feature set (i.e., the set F), realized geometrically, and covered by the simplices at the next hierarchical level according to the aperture definition. That is, a geometric structure of a simplex at hierarchical level N is covered by the simplex or simplices at the next N+1 level. Thus, the methodology delivers a general procedure on how a multidimensional hierarchical backcloth structure can be defined in a DGGS context with hierarchical tessellation and aperture concepts in mind, while considering Q-analysis concept of the well-defined cover sets.

3.3.2 Defining Hierarchical Traffic

In Q-analysis theory the hierarchical backcloth structure can also carry associated descriptive characteristics – traffic, a term used to describe some phenomena or activity which is directly dependent on the geometric structure of the backcloth (see Section 2.6). Similarly to the backcloth, traffic is also capable of having distinct hierarchical schemes and geometric representation. In other words, it follows a standard definition of a simplicial complex and its conjugate, including the formalization of the well-defined cover sets and their relation, binary matrix construction and corresponding association across DGGS hierarchy.

Considering the fact that backcloth and traffic have a strong contextual relationship, traffic definition must also account for the key DGGS concepts of hierarchical tessellation and aperture. Since, in a nutshell, traffic is a simplicial complex composed of simplices, which are defined by a relation between two sets (i.e., a binary matrix). One of these sets is inherited from the backcloth (e.g., DGGS cells set) in order to guarantee a common ground between backcloth and traffic, as well as to provide basis for the backcloth-traffic mapping (Figure 2.11, Figure 2.12). Whereas the other set contains elements which describe phenomena or activity. With this in mind, same DGGS principles and aperture rules used in the construction of the hierarchical backcloth hold true for the hierarchical traffic as well.

The primary reason why DGGS cells are the set which is considered for the traffic mapping is based on the need to have a common spatial object (i.e., hexagons) as a global reference for both backcloth and traffic structures. Such practice is both practical and logical considering DGGS data infrastructure incentives, the ability to have spatial versus data viewpoints (see Section 3.2.1), and demand to have multiple physical (i.e., backcloth) and descriptive (i.e., traffic) features embedded in a single DGGS cell with it being the only common ground. This, however, might be different from the conventional use of Q-analysis, where the subject matter (i.e., set F in this case) is used as a common ground between backcloth and traffic instead (Gould, 1980; Gould & Johnson, 1980; Johnson, 1983a, 1983b).

Therefore, given the same sequence of DGGS cells at various hierarchical levels a different simplicial complex $K_C^{N+k}(D^{M+k}, \mu^{(N+k, M+k)})$, can be formed to reflect the traffic structure at each hierarchical level. Note that the subject matter feature set F is replaced with the descriptive set D, which, in contrast, is capable of having a distinct hierarchical scheme M+k independent from the backcloth hierarchical scheme N+k. This resembles a double hierarchy of the new relation $\mu^{(N+k, M+k)}$, such that a cell element from the hierarchical set C^{N+k} is $\mu^{(N+k, M+k)}$ related to a descriptive element from the hierarchical set D^{M+k}. Similarly to the backcloth the μ relation must be redefined with changes to hierarchical levels in order to account for changes in spatial extent of DGGS cells, and now also for the contextual aggregation of the descriptive features.

Thus, traffic is another simplicial complex made of simplices (i.e., DGGS cells) and described by vertices (i.e., set D) to form a new geometric structure which represents backcloth descriptive characteristics. Each traffic simplex is covered by the simplex at the next hierarchical level pair. That is, a geometric structure of a simplex at level pair (N, M) is covered by the simplex or simplices at level (N+1, M+1).

3.3.3 Merging Backcloth and Traffic

Thus far, the methodology has described the key components and steps necessary to successfully define cover sets and form backcloth and traffic structures complementary with DGGS abstract specification. The following section further explores the process of merging two structures

together to form a new connected structure, such that each vertex in the backcloth is mapped with descriptive traffic vertices. The process is based on the earlier work of Johnson and his suggestions for the use of category theory to establish natural mapping relationship between backcloth and traffic sets (Johnson, 1983b, pp. 484–486).

The process requires the construction of a two-dimensional binary matrix using the general rules for matrix generation (see Section 3.2.1), given both backcloth and traffic related simplicial complexes $K_C^{N+k}(F, \lambda^{N+k})$ and $K_C^{N+k}(D^{M+k}, \mu^{(N+k, M+k)})$, as well as their corresponding binary matrices. That is, a new simplicial complex $K_C^{N+k}(F \times D^{M+k}, \rho^{(N+k, M+k)})$ is formed for each hierarchical level, where $F \times D^{M+k}$ is the Cartesian product between the subject matter F and its descriptive characteristics D, and new relation ρ as the mapping function of traffic vertices onto the backcloth. In a general sense, the connectivity structure of the entire relation is ($\rho^{(N+k, M+k)} \subseteq C^{N+k} \times (F \times D^{M+k})$), which requires backcloth elements (i.e., set F) for the traffic elements (i.e., set D) exist in each DGGS cell (Table 3.3).

Table 3.3. Construction of a two-dimensional binary matrix for a backcloth-traffic mapping of the descriptive set D onto the subject matter set F. Each element of these sets represents a vertex used to describe corresponding DGGS cells in set C.

	$F_1 \times D_1$	⋯	$F_1 \times D_n$	$F_2 \times D_1$	⋯	$F_2 \times D_n$	⋯	$F_n \times D_1$	⋯	$F_n \times D_n$
C_1	0/1		0/1	0/1		0/1		0/1		0/1
C_2	0/1		0/1	0/1		0/1		0/1		0/1
C_3	0/1		0/1	0/1		0/1		0/1		0/1
⋮										
C_n	0/1		0/1	0/1		0/1		0/1		0/1

This process is equivalent to mapping a traffic-based simplex onto one backcloth vertex at a time. Such a definition of the connectivity structure clearly demonstrates the concept outlined earlier by Atkin, which states that backcloth can either permit or restrict traffic (Atkin, 1972, pp. 163–165, 1980, pp. 384–386). The binary relation must be generated for each hierarchical level in order to account for the appropriate changes in backcloth and traffic aggregations accordingly. The mapping procedure is also flexible, which allows for additional physical features or groups (e.g., points, lines, polygons) be added to the relation, if necessary. In which case, the binary

table must reflect additional backcloth and traffic features accordingly as illustrated earlier (Table 3.1).

It is appropriate now to revisit the earlier statement and methodological choice for making DGGS cells (i.e., set C) as the common ground set between backcloth and traffic. The presented binary mapping approach is capable of defining a direct relation as a Cartesian product between subject matter set F and set D, which is said to be a descriptive set of the subject matter. This is more meaningful than having a Cartesian product between traffic elements and DGGS cells directly. Additionally, set F does not have a direct hierarchical component associated with it, which simplifies the process of backcloth-traffic mapping and the corresponding hierarchical Q-analysis.

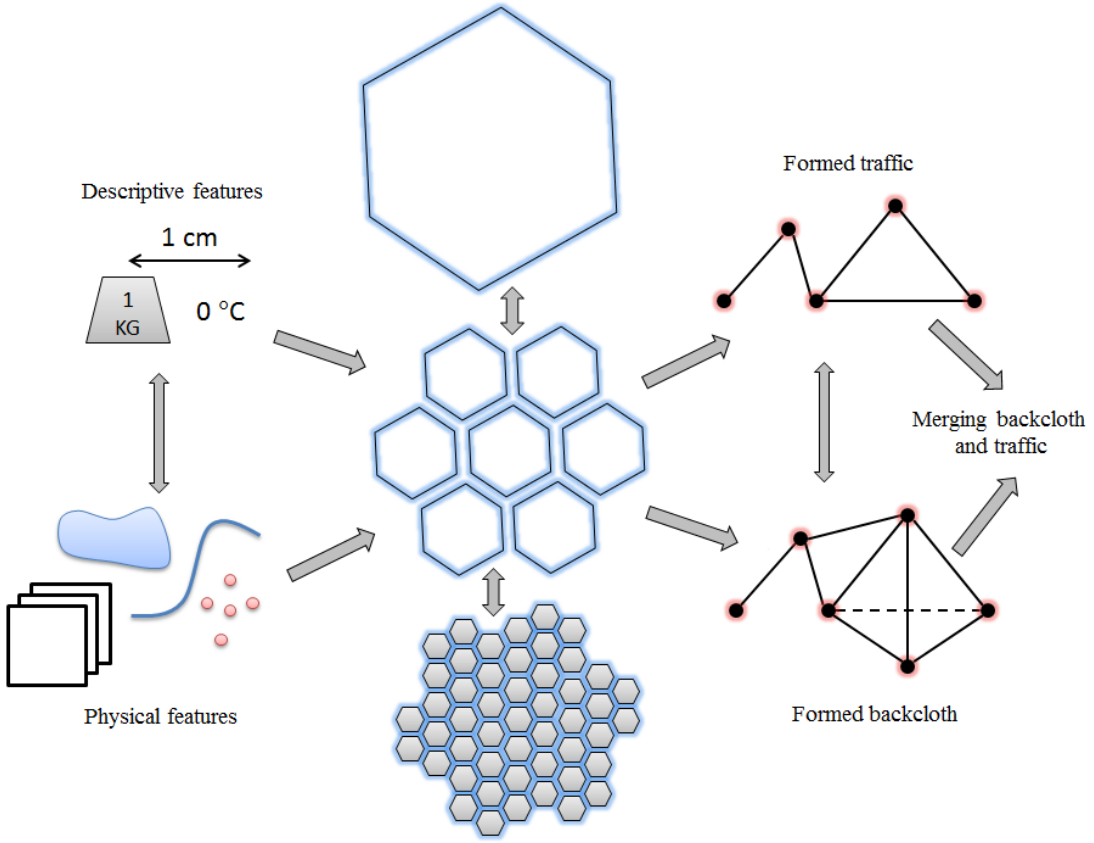


Figure 3.3. The figure describes a general workflow used to generate backcloth and traffic structures, as well as backcloth-traffic mapping for various hierarchical levels within the DGGS framework.

In summary, the entire process of formalizing spatial information in DGGS context at various hierarchical levels to form backcloth (i.e., subject matter of physical features) and traffic (i.e., descriptive features) structures, as well as the process of backcloth-traffic mapping, can be visualized in (Figure 3.3). With completion of the backcloth-traffic mapping process, the methodology further proceeds by defining a procedure for Q-analysis.

3.4 Implementing Q-analysis Algorithm

In a general sense, Q-analysis is mostly concerned with identifying so-called connected components or connected chains between sets elements to understand their structural properties and connectivity pattern (Atkin & Casti, 1977, p. 2). The technique is argued to provide the insight into the complex network systems made up of various interacting elements embedded in a matrix form or represented as a mathematical relation (Atkin, 1974, pp. 105–106). Therefore, the final part of the methodology is focused on implementing the Q-analysis algorithm in JavaScript according to the formal theoretical developments (Atkin, 1972, 1974, 1980; Beaumont & Gatrell, 1982; Gould, 1980; Johnson, 1981a). The implemented procedure here is appropriate to apply for hierarchical backcloth and traffic, as well as backcloth-traffic mapping as long as study sets are well-defined and necessary shared-face matrices are generated.

3.4.1 Producing Shared-Face Matrix

Given the two well-defined sets it is necessary to determine their cardinality. That is, number of elements in each set must be established in the (R x C) form, where R stands for the total number of rows (i.e., DGGS cells), and C for the total number of columns (i.e., physical or descriptive characteristics). Accordingly, the incidence matrix Λ is formed which relates rows and columns via some λ expression and evaluates to binary values of 1 or 0, with $\lambda_{(i,j)}$ indicating a specific (R, C) pair in Λ, such that:

$$\Lambda = \begin{array}{c|c} \lambda & C \\ \hline R & \lambda_{(i,j)} \end{array}$$

Similarly, a conjugate relation is represented as a transposition of Λ, such that:

$$\Lambda^{T} = \begin{array}{c|c} \lambda^{-1} & R \\ \hline C & \lambda^{-1}_{(i,j)} \end{array}$$

Next, the algorithm proceeds with connectivity analysis by generating shared-face matrices Γ_R of size (R x R) and Γ_C of size (C x C), such that:

$$\Gamma_R = (\Lambda * \Lambda^{T}) - \Omega_R, \text{ where } \Omega_R \text{ is all-ones (R x R) matrix, and}$$

$$\Gamma_C = (\Lambda^{T} * \Lambda) - \Omega_C, \text{ where } \Omega_C \text{ is all-ones (C x C) matrix.}$$

Both Γ_R and Γ_C matrices are symmetric around diagonals, thus only their halves are necessary to retain, including the diagonal values. The diagonal values indicate q-dimensionality of each simplex, with q being the diagonal value and (q+1) being the total number of related elements or vertices which define a simplex. The non-diagonal values, on the other hand, indicate direct connectivity between simplices. That is, two simplices directly share (q+1) vertices, where q is the value from the Γ_R and Γ_C matrices. This type of connectivity is also known as q-nearness. Two simplices can also be q-connected, such that they do not have vertices in common, but are still considered to be connected via other simplices nonetheless. Q-nearness implies q-connectivity, but not vice versa (Beaumont & Gatrell, 1982, p. 16). The final part of the algorithm is concerned with identifying a chain of connected simplices from the Γ_R and Γ_C shared-face matrices, such that all simplices are grouped into components and form connected chains according to their dimensionality and connectivity properties. For instance, given the arbitrary shared-face matrix (Table 3.4) it is possible to identify the following:

- The highest q-dimension is q = 3, which means that the corresponding $\sigma_3(C_3)$ simplex is defined by four vertices (i.e., q + 1). Thus, there are four distinct q-levels starting from the highest dimension down to zero, such as q = 3, q = 2, q = 1 and q = 0.
- At q = 3, a single component $\{C_3\}$ is formed, which consist of only one $\sigma_3(C_3)$ simplex. It is the only simplex with q-dimension being q = 3, and it shares (q + 1) vertices with itself.
- At q = 2, there are two components $\{C_2\}$ and $\{C_3\}$, but they are disconnected since they do not have (q + 1) vertices in common. In other words, they do not share a 2-

dimensional face but 1-dimensional only (i.e., they share a line). The component $\{C_3\}$ is also included at this level, because its q-dimension (i.e., q = 3) is greater than q = 2.

- At q = 1, there are two components $\{C_1\}$ and $\{C_2, C_3\}$. Here, it is possible to observe that both $\sigma_2(C_2)$ and $\sigma_3(C_3)$ simplices formed a connected component $\{C_2, C_3\}$, since at q = 1 they share a 1-dimensional face indeed. Simplex $\sigma_1(C_1)$ also enters the structure at this level as a separate component, since it is not connected.
- At q = 0, there is only a single connected component $\{C_1, C_2, C_3\}$, which forms a connected structure. A chain of connected simplices is also established, such that $\sigma_1(C_1)$ is 0-connected to $\sigma_3(C_3)$ via simplex $\sigma_2(C_2)$.
- The value of (-1) implies that corresponding simplices do not have any vertices in common, or they do not share a face.

Table 3.4. The table represents a randomly generated shared-face matrix as a worked example for the connectivity structure analysis.

	C_1	C_2	C_3
C_1	1	0	-1
C_2		2	1
C_3			3

The formal mathematical definition of q-connectivity is introduced by Atkin and restated here accordingly (Atkin, 1974, p. 178). In a general sense, two simplices $\sigma(p)$ and $\sigma(r)$ are said to be q-connected or joined by a chain of connection in some simplicial complex K, if there exist a sequence of simplices $\{\sigma(C_1),\ldots\sigma(C_n)\}$, such that:

- $\sigma(C_1)$ share a face with $\sigma(p)$
- $\sigma(C_n)$ share a face with $\sigma(r)$
- $\sigma(C_i)$ and $\sigma(C_{i+1})$ share a face of some dimension β_i, for i = 1,…(n-1)

Then, both simplices are joined by a chain of q-connectivity with q value being the minimum dimension in $\{\sigma(C_1), \sigma(\beta_1), \sigma(\beta_2),\ldots\sigma(\beta_{n-1}), \sigma(C_n)\}$.

3.4.2 Finding Connected Components

Certainly, with increasing data quantity and subsequent network complexity, the connectivity structure becomes difficult to track manually. Thus, the Q-analysis procedure for finding connected components of a given shared-face matrix has been implemented in JavaScript to automate the process (Figure 3.5), including its helper function for sets connectivity (Figure 3.4) and the testing module (Figure 3.6 cf. Table 3.4).

```
/*
Returns a connectivity dictionary of each row id in a shared-face matrix
for a specific q-value
@param  {Integer}                 qValue      A specific q-value
@param  {Object}                  qSharedFace Two dimensional matrix object
@param  {Array of Strings}        qRows       An array of row ids
@return {String: Array of Strings} qGraph     The connectivity dictionary
*/
function getQconnectivity(qValue, qSharedFace, qRows) {
      qGraph = {};

      for(i=0; i<qSharedFace._size[0]; i++) {
            let row = qSharedFace._data[i];
            let rowLen = row.length;
            let rowID = qRows[i];
            qGraph[rowID] = []; //initialize row id

            for(j=0; j<rowLen; j++) {
                  let val = row[j];
                  if((val >= qValue) && (j >= i)) {
                        qGraph[rowID].push(qRows[j]);
                  }
            }
      }

      //console.log(qGraph); //print output
      return qGraph; //return connectivity dictionary
}
```

Figure 3.4. The figure provides a helper function used to find connectivity of individual elements in a shared-face matrix to assist calculation of connected components.

```
/*
Returns connected components array of a shared-face matrix for a specific q-value
@param  {Integer}                   qValue      A specific q-value
@param  {Object}                    qSharedFace Two dimensional matrix object
@param  {Array of Strings}          qRows       An array of row ids
@return {Array of Arrays of Strings} qComponents An array of connected components
*/
function getQcomponents(qValue, qSharedFace, qRows) {
      let qComponents = getQconnectivity(qValue, qSharedFace, qRows); //helper function
      let currComp = [];

      //search through the graph to form connected components
      for(key in qComponents) {
            currComp = qComponents[key]; //initialize component

            //check if component values can be found in other keys
            for(i=0; i < currComp.length; i++) {
                  let value = currComp[i];

                  //check each key for the presence of the value
                  for(searchKey in qComponents) {
                        let searchArr = qComponents[searchKey];

                        //merge arrays if same value is found in two different keys
                        if((searchArr.includes(value)) && (searchKey != key)) {
                              searchArr.forEach(item => {
                                    if(!currComp.includes(item)) {
                                          currComp.push(item)
                                    }
                              })
                              delete qComponents[searchKey]; //remove searched key
                        }
                  }
            }
            delete qComponents[key]; //remove old key

            //replace key with a new formed component
            if(currComp.length) {
                  qComponents[key] = math.sort(currComp, math.compareNatural);
            }
      }

      //console.log(Object.values(qComponents)); //print output
      return Object.values(qComponents); //return connected components
}
```

Figure 3.5. The following figure outlines the algorithm which calculates connected components given the q-value, shared-face matrix, and array of row names as input.

```
const math = require("mathjs");

//Input:
let qValue3 = 3;
let qValue2 = 2;
let qValue1 = 1;
let qValue0 = 0;
let qSharedFace = math.matrix([[1,0,-1],[0,2,1],[-1,1,3]]);
let qRows = ["C1","C2","C3"];

//Output:
getQcomponents(qValue3, qSharedFace, qRows); //=> [["C3"]]
getQcomponents(qValue2, qSharedFace, qRows); //=> [["C2"],["C3"]]
getQcomponents(qValue1, qSharedFace, qRows); //=> [["C1"],["C2","C3"]]
getQcomponents(qValue0, qSharedFace, qRows); //=> [["C1","C2","C3"]]
```

Figure 3.6. This figure provides a code snippet of test cases and their outputs from the given shared-face matrix to find connected components for each q-value.

Chapter 4 Application

This chapter describes application of the proposed methodology to a real-world problem for water quality and water health monitoring. The effort was made in order to acquire some insight into the structural connectivity of the water flow network. The network is composed of various study sites each monitoring the trends in water quality and its chemical composition. For this reason, the study explores the use of the structural language of Q-analysis and its potential integration with DGGS as an approach for handling large quantities and various types of spatial information, as well as modelling of geographic features. The analysis aims to examine various types of connectivity between water sites, as well as to recognize patterns of potential water contamination, its sources and spread directionality at various levels of the DGGS hierarchy. At the same time, it is important to stress that it is not suggested to use the results of this application for the purposes of water quality mitigation, treatment procedures or environmental enforcement practices in order to improve the water health. The goal of this study is to provide a mean for the practical application of Q-analysis within DGGS framework, its potential advantages and benefits to work in this area of expertise for well trained specialists. Ideally, the methodology is suitable with any spatial datasets or spatial features globally as long as they are properly articulated and presented within DGGS hierarchy.

The chapter begins with description of data sources, study area and software used within the scope of this application in Section 4.1. Then, it continues with an introduction to the water health application and necessary preparations in Section 4.2. Section 4.3 in concerned with exploring backcloth connectivity structure, as well as performing and interpreting the Q-analysis output. Section 4.4 demonstrates the use of Q-analysis in a DGGS context by defining necessary hierarchical components and their relations.

4.1 Preliminary Information

The specific data sources that have been incorporated into the analysis, as well as the various auxiliary components, software and tools are described in the following subsections.

4.1.1 Data and Study Area

The primary dataset used in the following analysis is the Cumulative Impacts Monitoring of Aquatic Ecosystem Health of Yellowknife Bay, Great Slave Lake provided by the Mackenzie DataStream platform and the Government of the Northwest Territories, Canada (Government of the Northwest Territories, 2019). This dataset was collected over the period of 2014-2016 in order to gain a better picture of the current environmental conditions in Yellowknife Bay and the ways for their control, including water quality, sediment and aquatic food chain. The project was successful in providing new insights for assessment of the Yellowknife Bay ecosystem. This analysis was generally performed for the year of 2014 because it was more complete and had a higher number of records compared with other years.

Additionally, the CanVec series of the topographic data of Canada (Natural Resources Canada, 2019) and Canadian Digital Elevation Model (CDEM) (Natural Resources Canada, 2015) were used in order to assist in determining the direction of the water flow. For such purpose, CanVec hydro features (e.g., rivers, water bodies) in the vector format at the scale of 1:50,000, as well as the CDEM with a horizontal accuracy of 10-50 metres and vertical accuracy of 1-10 metres, were processed in the Yellowknife Bay, Great Slave Lake area, NWT, Canada and integrated in the analysis.

4.1.2 Software

H3 is a geospatial software library used primarily for the hierarchical partitioning and spatial indexing of geographic features on the Earth's surface. The library was developed by Uber Technologies, Inc., and released for public use under the open source license in 2018 (Brodsky, 2018). One of the distinct features of H3 is integration of the hexagon-based grid partitions with aperture 7. In this research the software was mainly used for the purposes of vector data type conversions into the hexagonal grids at various spatial resolutions, such as points, lines and polygons. These spatial features were also indexed via the provided API functionality for conversion of geographic coordinates and generation of hexagon grids layers. The JavaScript version of H3 is the particular binding which was used in this work. The version of H3 package used is 3.0.1.

Sigma.js is a highly customizable open source graph library developed in JavaScript, which provides APIs for integration of network datasets in web applications (Jacomy et al., 2017). The library was primarily used to display connectivity structure of the water flow network. The version of Sigma.js used is 1.2.1.

Node.js is the open source run time environment and platform used to execute JavaScript code (Node.js Foundation, 2019). JavaScript is the only programming language applied in this work, which was primarily used for working with H3 library for feature conversions, processing raw water quality data, passing the processed data to the graphing APIs of Sigma.js, as well as implementing and computing of Q-analysis methodology. The version of Node.js used is 8.11.3.

The rest of visualization and map making process was completed via ArcMap (Esri Canada, 2019) version 10.3.1 and QGIS (QGIS project, 2019) version 2.8.2 software, which were used concurrently. Both of these software packages are fairly popular and built to work with various forms of Geographic Information Systems (GIS) and data, thus include various tools that are assembled together in order to customize, edit, process and visualize spatial information.

4.2 Defining a Problem

To begin, it is necessary first to establish two well-defined sets, which will be formally used for generation of a binary relation matrix and further Q-analysis. Given the data for the cumulative monitoring of the aquatic ecosystem health (Government of the Northwest Territories, 2019), it is possible to define such well-defined sets. In particular, sets C (hexagon cells) – a set of hexagon-based locations represented by DGGS cells, and W (water sites) – a set of water sample sites that exist within the boundaries of DGGS cells, were defined (Figure 4.1 cf. Figure A.1); such that:

C = {c01, c02, c03, c04, c05, c06, c07, c08, c09, c10, c11, c12, c13, c14, c15}

W = {W1, W2b, W2s, W3b, W3s, W4, W6, W7, W8, W9, W10b, W10s, W11, W12, W13, W14_1, W14b, W14_2, W17b, W17s}

The initial resolution of H3 cells (H3 resolution 8) (Table B.1) was chosen in such a way that at least one site location exists per cell area, and for the cell areas to overlap with at least one waterbody in order to determine or approximate its water flow direction. One should also note how versatile it is to define cell units at different spatial resolutions, if required, and aggregate water sample sites to these spatial units accordingly.

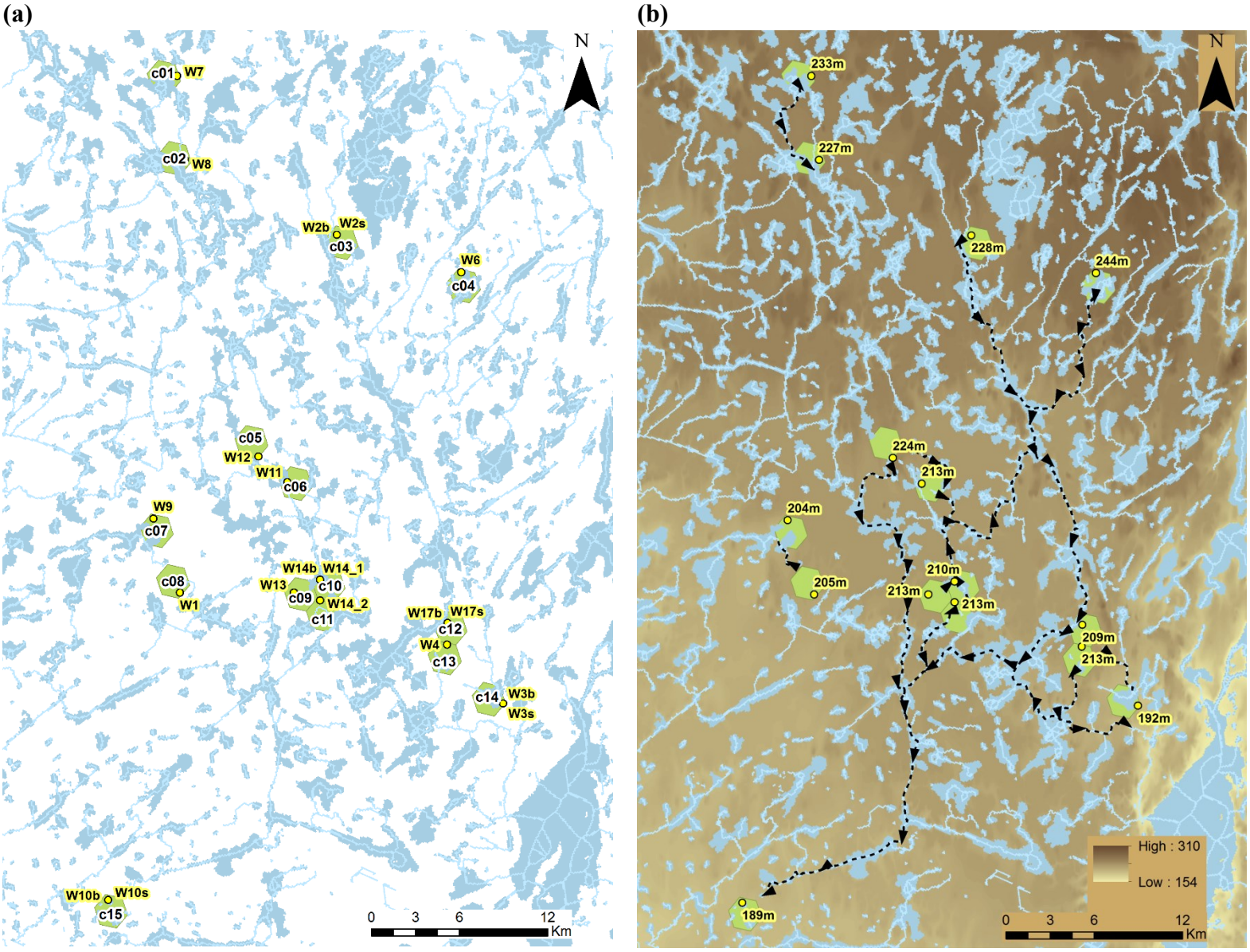


Figure 4.1. (a) Illustration of water sample sites (yellow labels) within DGGS hexagon cells (white labels). (b) Flow direction and elevation information as the primary parameters for determining connectivity structure between the spatial units. The resolution of H3 cells is 8. Data sources: (Government of the Northwest Territories, 2019; Natural Resources Canada, 2015, 2019).

Based on the nature of the problem, it is reasonable to assume that observations of the water monitoring sites have tendency to share similar water characteristics amongst related areas in the region as long as there is established connectivity between these spatial hexagon units. This connectivity is, in fact, a water network system. In other words, each hexagon unit is defined by the set of water sample sites which are directly connected to these hexagons through the water network system (Figure 4.1). Thus, using the direction of the water flow, as well as the digital elevation model of the particular geographic area, it is possible to derive the ($\lambda \subseteq C \times W$) relation. The relation indicates that some hexagon cell C_a has a potential to show similar observation characteristics of some water site W_b under the following connectivity rules and assumptions:

1) If some water site W_a is located within the boundary of its parent cell C_a then by definition (C_a, W_a) are related.
2) A water site W_a is said to be related to some cell C_b if there is direct water flow connection from the W_a parent cell C_a to the C_b, such that no other cell areas are in the way of the water flow.
3) A water site W_a is said to be related to the cell C_b if both cells C_a and C_b intersect a water body (e.g., lake), and elevation of C_a is higher than C_b.

Visually, such connectivity can be represented on a map (Figure 4.1), which, for example, shows that spatial area (c02) is likely to retain similar water characteristics determined by both (W8) and (W7) water sites. However, area (c01) is only characterized by (W7) site, since elevation and flow direction does not permit (W8) to easily travel into (c01). Similarly, while water site (W6) is visually traceable through the water network to the area (c14), their connectivity cannot be established due to area (c12) being on the way. Thus, (W6) is directly related to the (c12) instead.

In cases when spatial units are neighbours (e.g., without apparent water flow information between them), the connectivity and flow direction is then determined based on the elevation data and sharing of a water body. For example, water sites (W13) and (W14_2) will flow into (c10) due to them having the higher elevation (e.g., 213m versus 210m) and sharing of a water body (Figure 4.2), but not vice versa.

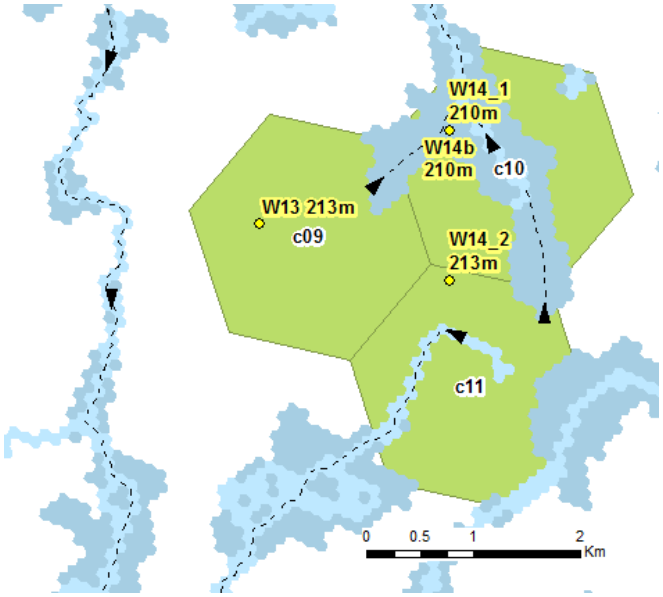


Figure 4.2. Assuming the water flow direction using the elevation data for the hexagon areas that share same body of water.

Additional connectivity assumptions were made in the areas where hydrological flow information is incomplete. In particular, with reasonable support of auxiliary information that indicates flow direction, such as presence of the connected lake system or elevation change, the flow direction was estimated as shown (Figure 4.3).

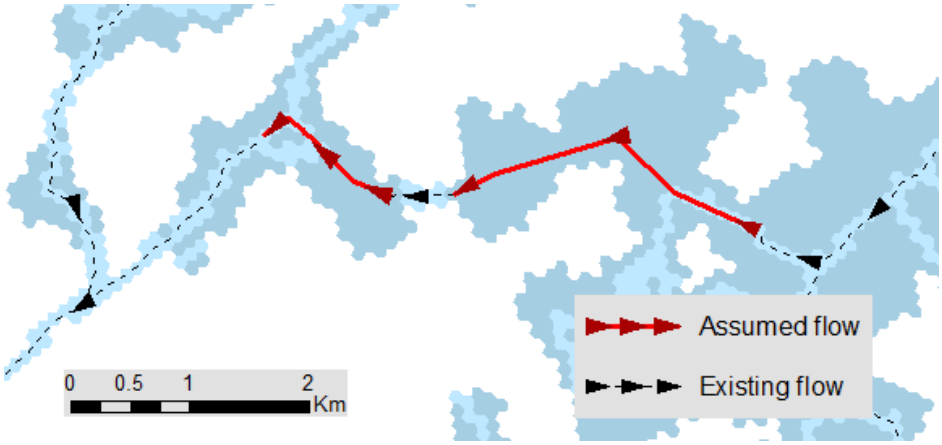


Figure 4.3. Example of the water flow assumption considering the existing directions of the water flow.

While connectivity methodology illustrated here was derived using the real-world elevation and water flow information, it is still suggested that qualified water and hydrology specialists use

more reliable in-situ data, as well as additional information (e.g., water catchment areas, ground water flow) specifically collected for this kind of analyses. Nonetheless, considering the illustrative purpose of this application the current data will suffice well to demonstrate the usefulness of such methodology for similar GIS applications.

4.3 Applied Q-analysis

This section outlines the significance and richness of the Q-analysis concept, as the language invented to study multidimensional connectivity and structural characteristics of data. The concept is based on the explicit definition of cover sets, and relation between their elements (e.g., water flow). Such relations expressed in a binary form allow for unconstrained interaction between the sets allowing them to have more general basis on which an element can be considered related to another element, which is opposite to the functional thinking (see Section 2.5.2). As stated by Gould, the idea of cover sets and its corresponding Q-analysis creates grounds where data can "speak for themselves" revealing its important pattern and structure (Gould, 1980, p. 174, 1981). The structure in this sense is often referred to as backcloth, the concept already familiar to the reader (see Section 3.3.1). Together, direct and conjugate simplicial complexes are what form the backcloth structure, which at the same time provides grounds for data to exist and be transmitted through the backcloth (Atkin, 1972, pp. 163–165; Gaspar & Gould, 1981, p. 190; Johnson & Wanmali, 1981, p. 273). The data in this sense is known as traffic (see Section 3.3.2). Traffic can be substituted by any observation set and directly attached to the geometric structure of the backcloth (Atkin, 1980, pp. 383–384). Both backcloth and traffic terms are explained in more details within the scope of this application.

4.3.1 Exploring Q-connectivity

The connectivity structure presented earlier can be captured by a binary matrix (Table 4.1), where λ relation indicates that each hexagon area has specific water characteristics defined by the water sites of that area. The relation also outlines that observation outcomes from other sites may be carried over through the physical features of the land as well, such as elevation and water

flow network in this case. In other words, if some water site is related to a hexagon area based on the established connectivity rules, their relation is marked as 1 in the corresponding table cell, otherwise it is left as blank. Note that formally absence of a relation is marked by 0; however, in (Table 4.1) zeros are omitted for visual clarity.

Table 4.1. A binary relation matrix which defines λ relation between hexagon areas and their water sites, where set {c01,…c15} represent hexagon cells and set {W1,…W17s} represents water sites. The value of 1 in the table implies that both hexagon cell and water site are related, whereas blank cell indicates the absence of the relation.

λ	W1	W2b	W2s	W3b	W3s	W4	W6	W7	W8	W9	W10b	W10s	W11	W12	W13	W14_1	W14b	W14_2	W17b	W17s
c01								1												
c02								1	1											
c03		1	1																	
c04							1													
c05														1						
c06													1	1						
c07	1									1										
c08	1																			
c09															1					
c10															1	1	1	1		
c11																		1		
c12		1	1				1						1			1	1		1	1
c13						1														
c14				1	1	1													1	1
c15											1	1		1				1	1	1

Each hexagon area can be further represented as a geometric object known as simplex. At the same time, simplex is represented by the list of vertices, where each vertex is a site location that has a potential to contribute to the water characteristics of a particular spatial unit (i.e., hexagon). For example, simplex σ_5(c15) can be represented as the subset σ_5(c15) = <W10b, W10s, W12, W14_2, W17b, W17s> of all water site locations in the W set, where each site location is connected with each other. In this case, the σ_5(c15) simplex is related to 6 water site locations, which means its dimension is (q = 5), hence its σ_5 subscript symbolization. In other words, q-dimension is always one unit less than the total number of vertices that describe some spatial area. Now, consider another simplex σ_1(c06), which has only two vertices σ_1(c06) = <W11, W12>. It happens that both σ_1(c06) and σ_5(c15) simplices have one vertex in common <W12>, which joins them together in a group forming a connected multidimensional space. Considering

each simplex in the binary relation ($\lambda \subseteq C \times W$) it is possible to construct a complete simplicial complex $K_C(W; \lambda)$. As geometric representation, the simplicial complex reflects the connectivity structure with each element $W_i \in W$ being a vertex (Figure 4.4).

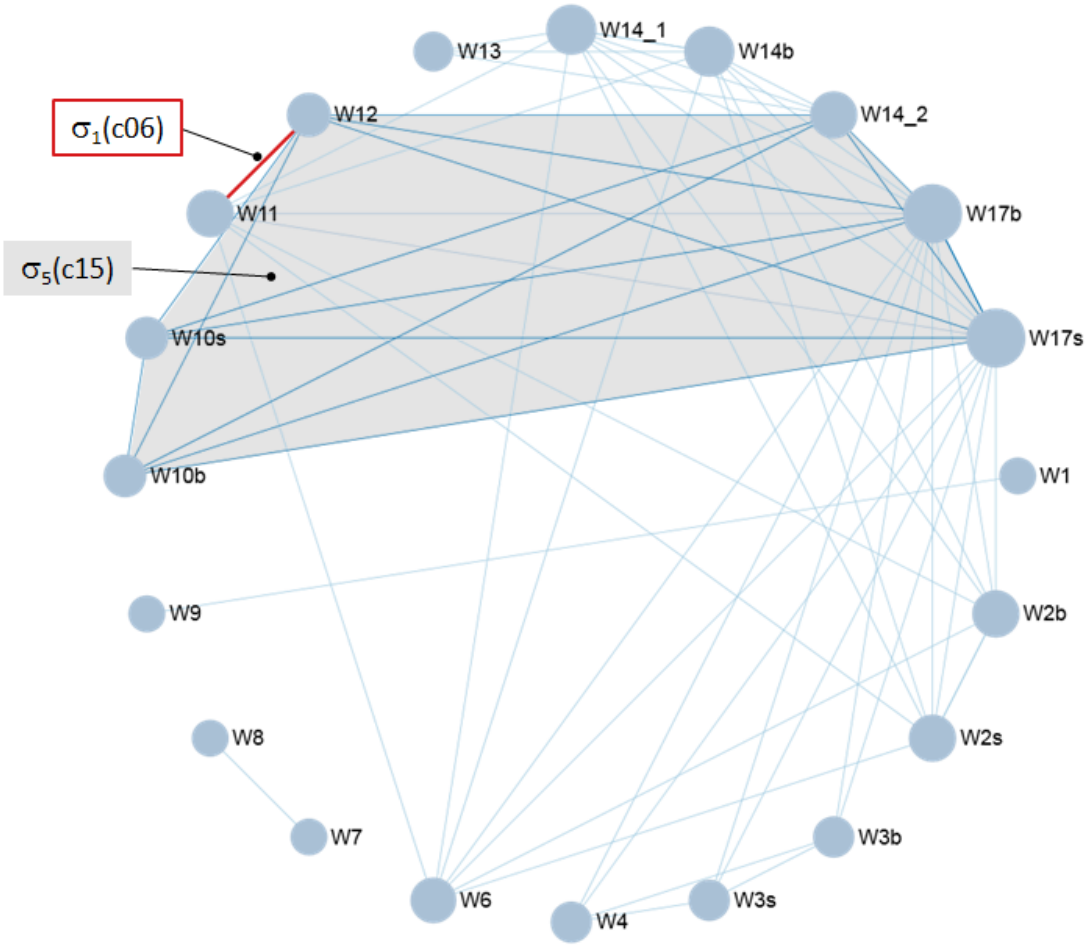


Figure 4.4. Geometric representation of $K_C(W; \lambda)$ simplicial complex, with labelled σ_5(c15) and σ_1(c06) simplices. The relative size of each vertex is proportional to their degrees (i.e., number of connected edges).

It is also possible to observe from the table (Table 4.1) and also from the graph (Figure 4.4) that certain vertices are shared across multiple simplices (i.e., hexagon areas) and are of various dimensionalities forming one well defined and connected structure which links all areas together (i.e., the simplicial complex representation of the relation λ). This connectivity is known as q-connectivity, and used to explore direct and indirect connections between simplices. At the same time, it is also important to differentiate the fact that some simplices can be q-connected and do

not have any vertices in common. If two simplices share vertices directly, such property is called q-nearness. Thus, q-nearness implies q-connectivity, but not vice versa (Beaumont & Gatrell, 1982, p. 16). The terms q-connectivity and q-nearness are two important concepts in Q-analysis used to determine and differentiate connectivity within the connected chains (see Section 3.4). For example, simplex σ_1(c03) is q-connected to σ_5(c15) through the hypervolume σ_7(c12) creating a chain of connected simplices. The statement can be confirmed by the physical water network (Figure 4.1 cf. Table 4.1), because σ_1(c03) flows first into σ_7(c12), and then σ_7(c12) continues to σ_5(c15). Thus:

- σ_1(c03) is 1-near to σ_7(c12) via <W2b, W2s>
- σ_7(c12) is 1-near to σ_5(c15) via <W17b, W17s>
- σ_1(c03) is 1-connected to σ_5(c15) via <W17b, W17s>

The properties of q-nearness and q-connectivity in $K_C(W; \lambda)$ can be further explored via the shared-face matrix representation, which was adapted from Atkin (1974, p. 186) (Table 4.2).

Table 4.2. Shared-face matrix of the $K_C(W; \lambda)$ simplicial complex.

	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13	c14	c15
c01	0	0	-	-	-	-	-	-	-	-	-	-	-	-	-
c02		1	-	-	-	-	-	-	-	-	-	-	-	-	-
c03			1	-	-	-	-	-	-	-	-	1	-	-	-
c04				0	-	-	-	-	-	-	-	0	-	-	-
c05					0	0	-	-	-	-	-	-	-	-	0
c06						1	-	-	-	-	-	0	-	-	0
c07							1	0	-	-	-	-	-	-	-
c08								0	-	-	-	-	-	-	-
c09									0	0	-	-	-	-	-
c10										3	0	1	-	-	0
c11											0	-	-	-	0
c12												7	-	1	1
c13													0	0	-
c14														4	1
c15															5

Note that due to the table being symmetric around the diagonal line, only one half of it is shown for visual clarity; however, formally (i, j) pair of (row, column) value is equal to the (column,

row) value at (j, i), where (i, j) ∈ C. For example, value at (c03, c12) is equal to the value at (c12, c03), yet the latter one is omitted as mentioned. The table illustrates relation between the simplices (hexagon areas) where corresponding numbers reflect q-connectivity (i.e., they share (q+1) water sites). The dash symbol (-) is equivalent to (-1), which means that two simplices are disconnected (i.e., they share zero water sites). Note that diagonal values indicate q-dimension of a simplex. Instantly, it is possible to observe that areas with high q-dimension (e.g., c12, c15, c14, c10) play a significant role in the water network system, as will be shown shortly. With this in mind it is now appropriate to perform the formal Q-analysis on the $K_C(W; \lambda)$ simplicial complex to explore q-connectivity at different dimensional levels.

4.3.2 Direct Q-analysis

Q-analysis is mainly concerned with explaining the structural features of relationships between variables by identifying chains of connected simplices. In this study, these are hexagon areas that form a connected group at each q-dimension starting from the highest down to a 0-dimension. As mentioned previously, the dimension is calculated based on the number of vertices (here water sites) that define each hexagon, and is used to group simplices into connected components (Table 4.3). Q-analysis ranks each simplex according to its q-dimension or q-value, and tracks at which q-level a simplex enters the structure as well as the point when it becomes connected with other simplices. Note, the measure of the q-dimension is cumulative, which means that once a simplex enters the structure it will remain in the structure. For example, σ_7(c12) enters the structure at q = 7, since its dimension is 7; however, it is also present at q = 6, 5,…0, because its q-dimension is higher than all subsequent q-dimensions in the table (Table 4.3).

The complementary geometric representation (Figure 4.5), on the other hand, captures the connectivity for all vertices within the direct simplicial complex $K_C(W; \lambda)$. At first, the produced results might be confusing; however, it will soon be clear how much contextual information it is possible to retrieve from these findings. A detailed procedure for performing Q-analysis and finding connected components have been implemented in JavaScript (see Section 3.4.2) according to the formal definition of the Q-analysis theory (Atkin, 1972, 1974, 1980; Beaumont

& Gatrell, 1982; Gould, 1980; Johnson, 1981a). Therefore, reader is encouraged to refer to these original sources for additional information and theoretical background behind the concept.

Table 4.3. Q-analysis output of the $K_C(W; \lambda)$ simplicial complex.

q-value	Connected Components of $K_C(W; \lambda)$
7	{c12}
6	{c12}
5	{c12}, {c15}
4	{c12}, {c14}, {c15}
3	{c10}, {c12}, {c14}, {c15}
2	{c10}, {c12}, {c14}, {c15}
1	{c02}, {c03, c10, c12, c14, c15}, {c06}, {c07}
0	{c01, c02}, {c03, c04, c05, c06, c09, c10, c11, c12, c13, c14, c15}, {c07, c08}

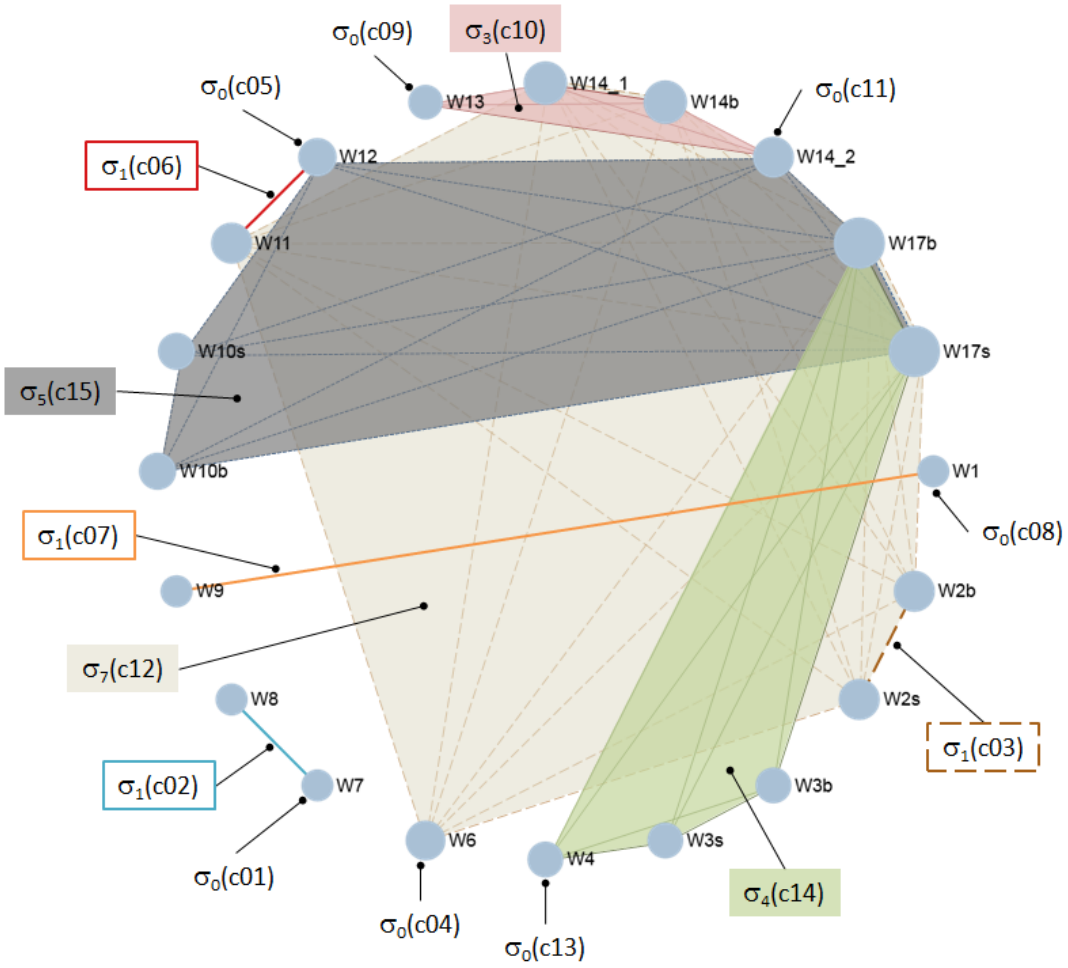


Figure 4.5. The figure represents connected components of the $K_C(W, \lambda)$ simplicial complex.

Given the Q-analysis output of the $K_C(W; \lambda)$ complex it is possible to observe that component {c12} consist of a single simplex and has the highest q-dimension in the complex, q = 7. This brings a lot of attention to this area, as it indicates that simplex σ_7(c12) is related and defined by the largest number of vertices (i.e., q + 1 = 8) or water sites. According to the physical context of the application, this means that this area is the most sensitive to the dynamic changes in water and one of the first areas that can potentially reflect water quality in the region, since water flow at this location can be traced from multiple water site locations. At q = 6, a similar trend is observed. It is shown that component {c12} still consist of a single σ_7(c12) simplex, and is the only one that can exist in this dimension (i.e., have (q + 1 = 7) or more vertices). The next σ_5(c15) simplex enters the structure at q = 5, following by σ_4(c14) simplex at q = 4. At this point, it is possible to observe that all three of these simplices share two vertices <W17b, W17s> which form a simplicial complex; however, they all appear to be disconnected at q = 4, forming 3 separate components. This make sense because in order for them to be connected into a single component they must share at least (q + 1 = 5) vertices at q = 4, while sharing only 2 vertices <W17b, W17s> in total. Nevertheless, a clear pattern is starting to emerge forming a connectivity structure of the water network, as well as the significance of individual areas and their water sites for the water monitoring purposes. At this point, it is important to clarify that, in this context, sharing of vertex or water site implies sharing of water quality parameters including pollution sources if these were to occur.

Furthermore, at q = 1 it becomes apparent that component {c03, c10, c12, c14, c15} along with its corresponding simplices form a chain of connections, grouping them all into one connected structure (Figure 4.5). This means that all of them share at least (q + 1 = 2) vertices through a chain of connected simplices, such that:

- σ_1(c03) is 1-near to σ_7(c12) via <W2b, W2s>
- σ_3(c10) is 1-near to σ_7(c12) via <W14_1, W14b>
- σ_7(c12) is 1-near to σ_4(c14) via <W17b, W17s>
- σ_7(c12) is 1-near to σ_5(c15) via <W17b, W17s>
- σ_1(c03) is 1-connected to σ_4(c14) via <W17b, W17s>
- σ_1(c03) is 1-connected to σ_4(c15) via <W17b, W17s>

- σ_3(c10) is 1-connected to σ_5(c14) via <W17b, W17s>
- σ_3(c10) is 1-connected to σ_5(c15) via <W17b, W17s>
- σ_4(c14) is 1-near to σ_5(c15) via <W17b, W17s>

At this time, it is possible to infer several statements regarding the observed connectivity structure at q = 1. Firstly, it is evident that areas (c03 and c10) are indirectly connected or q-connected to (c14 and c15) areas through the existing water network. This can be potentially confirmed and identified by at least two water sites at each location, in the case of any changes in the water quality at (c03 and c10). The same can be anticipated for the areas which are connected directly or are q-near. Secondly, at q = 1 there are also three additional {c02}, {c06}, {c07} components that appear to be disconnected. This means that all of them have at least (q + 1 = 2) vertices, yet they are not shared with the rest of the structure. Lastly, q-nearness between c14 and c15 is not the same as the q-nearness of the other 1-near areas. Here, q-nearness shows that sources are shared by inheritance without direct water flow connectivity (i.e., the water flow is opposite between the sites), whereas in other cases the sources are shared by direct physical transmission of water through the network. In other words, c14 and c15 are related in terms of sharing similar water characteristics of W17b and W17s water sites inherited from area c12, while not having direct water flow connectivity.

Topologically, this connectivity can be reflected via the inverted graph structure (Figure 4.6), also known as conjugate simplicial complex (see Section 2.5.3) the full advantage of which is revealed in the next section of the thesis. Note, the relation between c14 and c15 is shown as a dashed line, to differentiate that their connectivity does not imply connectivity by a direct water flow connection, but by inheritance of water characteristics from c12. Additionally, there is another similar connectivity between c10 and c15; however, is it not reflected until q = 0 level, since σ_3(c10) is 0-near with σ_5(c15) as shown in the shared-face matrix (Table 4.2). The analysis of this type is commonly referred to as local structure analysis, and mostly focused on identifying similar behaviours and connectivity structure within individual connected components (Atkin, 1972, p. 155; Beaumont & Gatrell, 1982, p. 17).

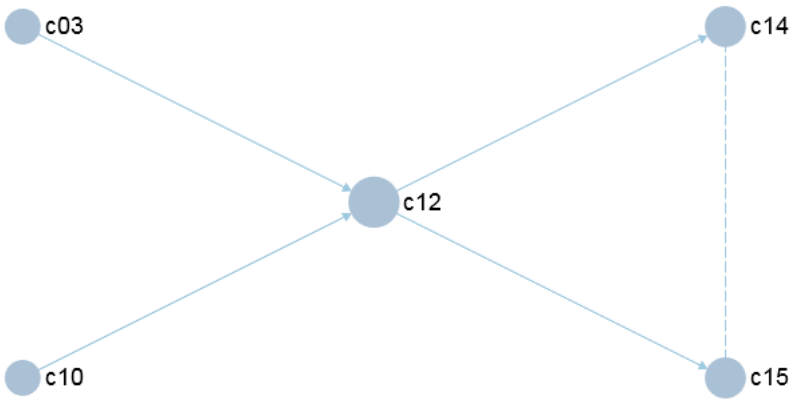


Figure 4.6. The figure shows connectivity at q = 1 for the {c03, c10, c12, c14, c15} connected component. The direction of a line here indicates the direction of the water flow, whereas dashed line implies connectivity between areas by inheriting similar water characteristics from c12.

As q-dimension drops down to q = 0, it is now possible to identify that all remaining simplices have entered the structure and observe the occurrence of three natural chain clusters {c01, c02}, {c03, c04, c05, c06, c09, c10, c11, c12, c13, c14, c15} and {c07, c08}. Such finding suggests that communication between these disconnected components is impossible, since based on the established water flow connectivity these areas cannot have any vertices in common. By inference this also implies that these components cannot share pollution sources and have similar water characteristics if one of the components were to be polluted. At the same time, it is important to recognize that while disconnected components can share a common source of upstream pollutions, these, however, cannot propagate as local sources between them. In other words, same source of pollution can affect more than one disconnected component.

4.3.3 Conjugate Q-analysis

In order to get a better insight on the connectivity structure, consider also inverting the relationship of the original binary matrix (Table 4.1) and its corresponding $K_C(W; \lambda)$ simplicial complex to the inverse λ^{-1} relation (Table 4.4). In this case, set C can be treated as the vertex set instead, and used to generate the previously mentioned conjugate simplicial complex $K_W(C; \lambda^{-1})$. The inverse relation can be easily acquired by transposing the original binary matrix, as in to read rows as columns and vice versa (Table 4.4).

Table 4.4. The transposed binary matrix used to generate the $K_W(C; \lambda^{-1})$ conjugate complex. The value of 1 indicates that water site characteristics affect the corresponding area.

λ^{-1}	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13	c14	c15
W1							1	1							
W2b			1									1			
W2s			1									1			
W3b														1	
W3s														1	
W4													1	1	
W6				1								1			
W7	1	1													
W8		1													
W9							1								
W10b															1
W10s															1
W11						1						1			
W12					1	1									1
W13									1	1					
W14_1										1		1			
W14b										1		1			
W14_2										1	1				1
W17b												1		1	1
W17s												1		1	1

The above table reveals water sites that have a direct influence on hexagon areas. For example, W1 water site affects areas c07 and c08, whereas W9 affects only area c07. In the similar fashion, Q-analysis can be also applied onto the $K_W(C; \lambda^{-1})$ conjugate complex to identify chain of connected simplices of water sites. The significance of which is important, as it grants the opportunity to study the structure from different viewpoints. In particular, the $K_C(W; \lambda)$ complex reveals a spatial point of view or cell-oriented perspective with the connectivity chains necessary for interaction between hexagon spaces, whereas complex $K_W(C; \lambda^{-1})$ offers a data point of view or site-oriented perspectives and analogous chains (Atkin, 1972, pp. 157–158, 1974, pp. 32–33). Accordingly, the shared-face matrix of the $K_W(C; \lambda^{-1})$ simplicial complex (Table 4.5), its Q-analysis (Table 4.6) and geometric representation (Figure 4.7) can be generated in a similar manner to observe the λ^{-1} relation. The dimensionality of the shared-face matrix and Q-analysis represents the connectivity with other water sites in the area.

Table 4.5. Shared-face matrix of the $K_W(C; \lambda^{-1})$ simplicial complex.

	W1	W2b	W2s	W3b	W3s	W4	W6	W7	W8	W9	W10b	W10s	W11	W12	W13	W14_1	W14b	W14_2	W17b	W17s
W1	1	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-
W2b		1	1	-	-	-	0	-	-	-	-	-	0	-	-	0	0	-	0	0
W2s			1	-	-	-	0	-	-	-	-	-	0	-	-	0	0	-	0	0
W3b				0	0	0	-	-	-	-	-	-	-	-	-	-	-	-	0	0
W3s					0	0	-	-	-	-	-	-	-	-	-	-	-	-	0	0
W4						1	-	-	-	-	-	-	-	-	-	-	-	-	0	0
W6							1	-	-	-	-	-	0	-	-	0	0	-	0	0
W7								1	0	-	-	-	-	-	-	-	-	-	-	-
W8									0	-	-	-	-	-	-	-	-	-	-	-
W9										0	-	-	-	-	-	-	-	-	-	-
W10b											0	0	-	0	-	-	-	0	0	0
W10s												0	-	0	-	-	-	0	0	0
W11													1	0	-	0	0	-	0	0
W12														2	-	-	-	0	0	0
W13															1	0	0	0	-	-
W14_1																1	1	0	0	0
W14b																	1	0	0	0
W14_2																		2	0	0
W17b																			2	2
W17s																				2

Table 4.6. Q-analysis output of the $K_W(C; \lambda^{-1})$ simplicial complex.

q-value	Connected Components of $K_W(C; \lambda^{-1})$
2	{W12}, {W14_2}, {W17b, W17s}
1	{W1}, {W2b, W2s}, {W4}, {W6}, {W7}, {W11}, {W12}, {W13}, {W14_1, W14b}, {W14_2}, {W17b, W17s}
0	{W1, W9}, {W7, W8}, {W2b, W2s, W3b, W3s, W4, W6, W10b, W10s, W11, W12, W13, W14_1, W14_2, W14b, W17b, W17s}

The output of the shared-face matrix shows a lot of zeros (Table 4.5), which implies that these water sites share as much as one hexagon cell location between each corresponding site. The following can be confirmed in the Q-analysis table (Table 4.6). Q-analysis indicates that the highest dimension (i.e., q = 2) has three disconnected components {W12}, {W14_2} and {W17b, W17s}. Component {W17b, W17s} is the only one which has 2 connected water sites; however, this connection is not surprising since both W17b and W17s belong to a single hexagon location c12. Thus, they are already related spatially.

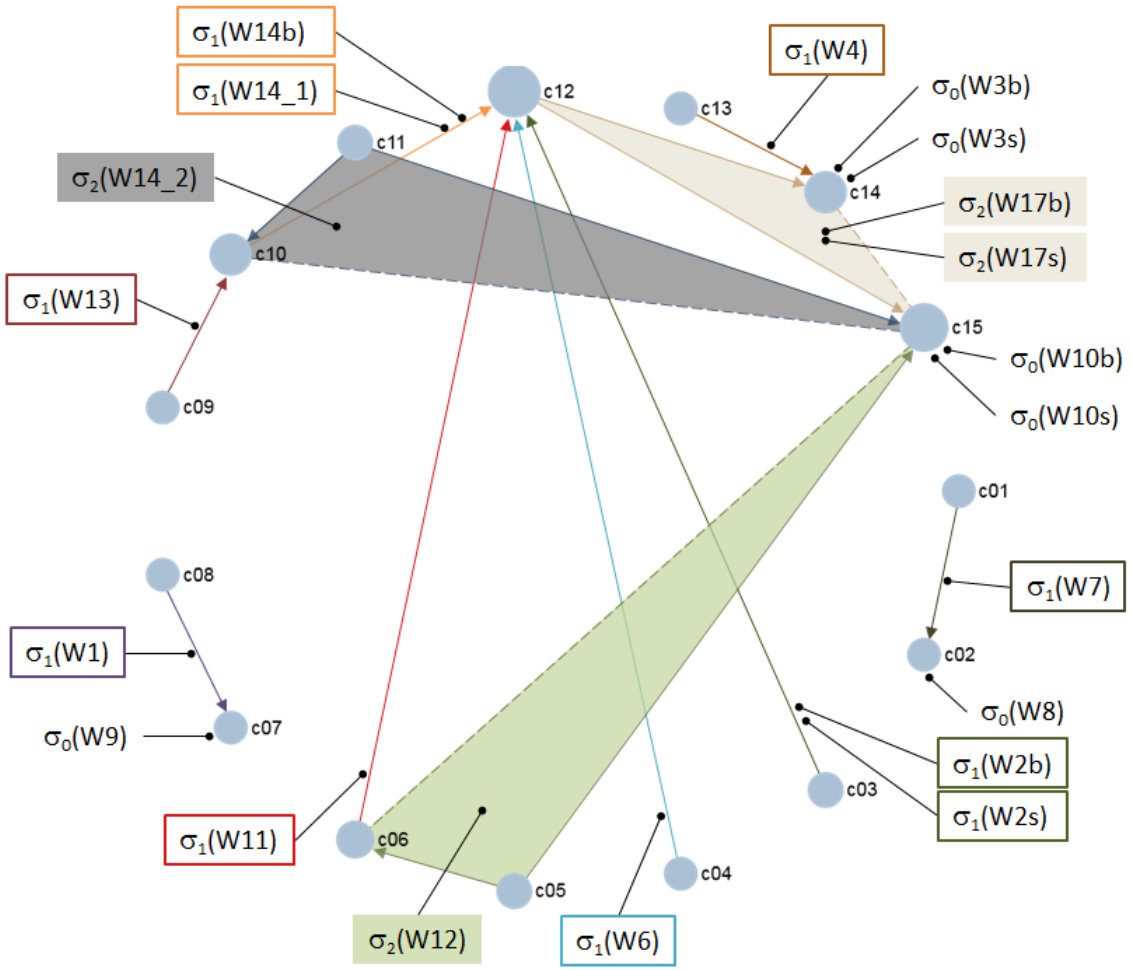


Figure 4.7. Geometric representation of the conjugate $K_W(C; \lambda^{-1})$ simplicial complex. The arrows indicate the direction of the water flow, whereas dashed lines imply connectivity by inheritance.

One interesting feature that can be observed here is that each component at q = 2 seem to connect its destinations areas which are not directly connected (i.e., absence of the direct water flow). This can be further confirmed graphically by the corresponding connectivity (i.e., dashed line) (Figure 4.7). For example, site simplex σ_2(W12) originates in the c05 area vertex, which at the same time flows into vertices c06 and c15, making their direct simplices q-near. Note the inverse λ^{-1} relation between the hexagon areas and water sites in this case, such that:

- σ_2(W12) (site in c05) is 0-near to σ_1(W11) (site in c06) via <c06>
- σ_2(W12) (site in c05) is 0-near to σ_1(W10b) (site in c15) via <c15>
- σ_2(W12) (site in c05) is 0-near to σ_1(W10s) (site in c15) via <c15>

Thus, causing σ_1(c06) = <W11, W12> and σ_5(c15) = <W10b, W10s, W12, W14_2, W17b, W17s> direct simplices connected via simplex <W12> (Figure 4.5), such that:

- σ_1(c06) is 0-near to σ_5(c15) via <W12>

Similarly, simplex σ_2(W14_2) originates in the c11 area and flows in two directions, such as c10 and c15 vertices. This makes σ_3(c10) = <W13, W14_1, W14b, W14_2> and σ_5(c15) = <W10b, W10s, W12, W14_2, W17b, W17s> connected via simplex <W14_2> (Figure 4.5), such that:

- σ_3(c10) is 0-near to σ_5(c15) via < W14_2>

Finally, both simplices σ_2(W17b) and σ_2(W17s) originate in c12 and flow into c14 and c15. Thus, making σ_4(c14) = <W3b, W3s, W4, W17b, W17s> and σ_5(c15) = <W10b, W10s, W12, W14_2, W17b, W17s> connected via simplex <W17b, W17s> (Figure 4.5), such that:

- σ_4(c14) is 1-near to σ_5(c15) via <W17b, W17s>

This suggests that areas c05, c11 and c12 along with their corresponding water sites are the sources with the highest spread variability. In particular, they flow in two different directions, whereas other areas seem to flow in one direction only or have no further water flow within the study area boundaries. What is interesting here is that these parameters of the water network were not revealed through the direct λ relation, but the inverse λ^{-1} relation. This underlines the importance and practical significance of Q-analysis from both spatial and data perspectives.

At q = 1 the analysis identifies 11 separate components none of which are 1-connected apart from areas that contain more than one water site (e.g., c03, c10, and c12). The 1-connectivity, in this case, implies sharing of 2 same hexagon areas. This by itself communicates that different water sites affect different hexagon areas, which makes the water flow highly fragmented. Fragmentation, in this case, can be viewed as decentralized water network, where water flow mainly travels in one direction only. At the same time, components at q = 2, can be also viewed as decentralized water network; however, they are the most centralized in the $K_W(C; \lambda^{-1})$ complex, since the corresponding water sites have more destinations (i.e., two destinations each). Fragmentation in the water context can be viewed as a positive thing, since if one site were to be polluted it will spread in less number of different directions. Thus, preventing a significant

environmental impact, as well as giving more time for the government authorities to respond and mitigate the issue. Additionally, fragmentation is also considered as part of the global structure analysis and mostly used to give insights about the structure of the whole complex at each q-level (Atkin, 1972, p. 155; Beaumont & Gatrell, 1982, p. 17).

Lastly, at q = 0 it is possible to reconfirm the same results found also on the $K_C(W; \lambda)$ complex, as Q-analysis on $K_W(C; \lambda^{-1})$ shows 3 separated components and makes 3 natural chain clusters. These components are equivalent to the components of the $K_C(W; \lambda)$ complex, yet with hexagon cells acting as vertices due to the λ^{-1} relation for the data viewpoint perspective.

4.4 Q-analysis and DGGS

While the example of the water health system is appropriate to illustrate the concept of Q-analysis, one must not forget that one of the biggest advantages of Q-analysis is its natural integration with DGGS. It has potential to work with various datasets and data types for the purposes of getting insight of the structural and connectivity properties. Thus, the reader should not constrain their thinking into one particular application presented here. The key contribution of this research is to build upon the earlier developed Q-analysis theory in order to pave the way for the enhanced analytical capabilities within the DGGS data architecture domain. The water quality and water health application here serves as an illustration of how it can be achieved in practice. The application also demonstrates how backcloth and traffic sets can be constructed with various dataset choices as long as these choices are justifiable contextually.

It is fascinating to learn that back in 1970s Atkin has implicitly recognized the need for DGGS (Atkin, 1974, p. 117), as he attempted to cover a town area with a set of discrete areal spaces called lozenges (S. A. Roberts, personal communication, October 9, 2018). In his work Atkin has stressed that lozenges size must complement the size of geographic features. In particular, the lozenges must be big enough to cover a feature but also small enough to not interfere with other features, as well as they must be subject to aggregation. Thus, producing lozenge covers at different hierarchical levels was necessary to meet the requirements of the study.

Today, DGGS offers a natural solution to the Atkin's issue, providing a mathematical structure where lozenges can be treated as DGGS cells (e.g., hexagons, squares, triangles) along with their aperture and hierarchical properties. The cells can be further associated with some descriptive spatial information to produce a meaningful interpretation of the phenomena via integrated Q-analysis, which is capable of handling data at various levels of hierarchical aggregation (i.e., spatial resolution). Thus, it seems promising to experiment with the idea in a broader context and applications of various origin, as well as different DGGS types.

4.4.1 Backcloth and Traffic

To reiterate, a structure of the water quality and health application discussed in this chapter is defined by the two well-defined cover sets C (hexagon cells) and W (water sites) along with a series of binary relations between each element of these sets. Both sets define what is known a backcloth formed by the ($\lambda \subseteq C \times W$) and ($\lambda^{-1} \subseteq W \times C$) relations. The backcloth connectivity structure has a direct influence on traffic, the existence of which is able to support or constrain the traffic. This idea was earlier proposed by Atkin, where he states that traffic requires certain backcloth connectivity in order for traffic to exist on it (Atkin, 1972, p. 163, 1980, p. 384).

Traffic, in this context, was chosen to be observed concentrations of water quality parameters, such as aluminium, turbidity, strontium and barium, collected at each water station in the study region. The water parameters here were mainly chosen based on the Canadian guidelines criteria for drinking water quality (Health Canada, 2019), in order to single out ones that are either pose high risk for water contamination or exist in significant concentrations in water. The idea proposed by Atkin is well supported in this context as well, as the existence of water network and physical water features provide grounds for measuring and monitoring water quality in designated locations (i.e., water sites). These locations form a connectivity structure of the backcloth (i.e., vertices), which in fact determine whether traffic can be attached to their geometric structure. In other words, designated water site locations and existence of the physical water network, provide grounds for the water parameters traffic to be collected (i.e., traffic is supported), whereas absence of those physical features implies that such data will not exists (i.e., traffic is constrained). This is consistent with the Einsteinian view on backcloth and traffic

geometries (see Section 2.6), such that adding or removing physical water sites (i.e., backcloth changes) have a direct impact on how water quality data is collected (i.e., traffic changes).

Similarly to the backcloth, traffic can also form a connected structure by being directly related to the earlier defined set C (hexagon cells), such that each hexagon areal simplex can carry some water parameters traffic. In this case, a new traffic set T can be defined, such that:

$$T = \{Al, Tur, Sr, Ba\},$$

where Al = aluminum, Tur = turbidity, Sr = strontium and Ba = barium. Correspondingly, a new relation ($\rho \subseteq C \times (W \times T)$) is formed as a mapping function of traffic vertices onto the backcloth vertices, along with its direct $K_C(W \times T; \rho)$ and conjugate $K_{WxT}(C; \rho^{-1})$ simplicial complexes.

The theoretical background for such mapping of traffic onto the backcloth was suggested by Johnson (1983b, pp. 484–486) based on his research on the structural complexity of television policy (Gould & Johnson, 1980). Based on his work it is suggested to arrange the relation between hexagon cells, water sites and water parameters as a two-dimensional binary matrix, such that both water site and water parameter elements appear in the same hexagon cell (see Section 3.3.3). What might not be obvious at this point is how exactly each cover set relates to one another. To clarify this, and also reach a stage at which the backcloth-traffic mapping can be performed, there are several developments that are necessary to consider, one of which is the concept of slicing (see Sections 2.5.2, 3.2.1).

Consider the raw traffic data (Table 4.7) acquired from each water station and related to the corresponding hexagon areas. The relation in this case is not in a binary format, but a decimal number indicating the concentration measure for a given water parameter and related hexagon cell. The data in this table was processed by taking the average value when multiple observations were recorded for a single water site. Similarly, if more than one water site falls within the border of a hexagon area, then their values were averaged likewise for each water parameter accordingly. Alternatively, it is also possible to have various approaches for the same data processing requirement. For example, one might consider filtering out samples that don't have certain number of observations recorded or additional vertices added to the traffic structure. However, these were not considered in order to avoid unnecessary complication of the analysis

and focus primarily on the main objectives of this research. Nevertheless, possibilities for the further application extension are available.

Table 4.7. This table provides processed cumulative impacts monitoring data for the aquatic ecosystem health in 2014, where each hexagon ID is associated with corresponding average concentrations of water parameters. All units are indicated in brackets, the N/A cells signify absence of data record for the given hexagon cell ID. Data source: (Government of the Northwest Territories, 2019).

cell ID	Aluminum (ug/L)	Turbidity (NTU)	Strontium (ug/L)	Barium (ug/L)
c01	16.450	2.450	71.525	19.700
c02	20.300	2.800	74.950	21.200
c03	3.500	N/A	64.450	17.400
c04	56.475	10.150	74.850	21.775
c05	20.000	2.050	104.250	31.400
c06	13.825	1.550	106.750	32.925
c07	15.975	2.100	80.825	23.050
c08	6.400	N/A	2170.000	55.800
c09	14.500	2.600	114.750	36.400
c10	7.0125	N/A	128.375	41.325
c11	N/A	0.850	N/A	N/A
c12	2.200	N/A	125.000	40.500
c13	13.525	1.150	65.200	17.950
c14	3.600	N/A	68.000	18.500
c15	7.500	N/A	88.750	25.900

The concept of slicing is mainly concerned with identifying some threshold parameter θ, such that if an observation value in the data matrix exceeds the threshold than the relation is assigned a value of 1, otherwise it is 0 or left out as blank. Given the water health data (Table 4.7), it is possible to obtain additional traffic relation ($\mu \subseteq C \times T$) as a binary matrix (Table 4.8), such that each element of $\mu_{(C,T)} > \theta$, where:

- $\theta_{Tur} = 1$ NTU (Nephelometric Turbidity Units)
- $\theta_{Al} = \theta_{Sr} = \theta_{Ba} = 60^{th}$ percentile

The process of defining the θ parameter requires careful consideration as different slicing parameters will generate different structures accordingly.

Table 4.8. The binary matrix of the relation μ, which represents the connectivity between hexagon areas C and water parameters T, where set C = {c01,…c15} and set T = {Al, Tur, Sr, Ba}. The value of 1 indicates that the value in the corresponding cell is greater than the threshold parameter θ.

μ	Al	Tur	Sr	Ba
c01	1	1		
c02	1	1		
c03				
c04	1	1		
c05	1	1	1	1
c06		1	1	1
c07	1	1		
c08			1	1
c09	1	1	1	1
c10			1	1
c11				
c12			1	1
c13		1		
c14				
c15				

In this application each θ parameter was chosen based on the data spread, as well as the guidelines for the quality of Canadian drinking water (Health Canada, 2019). In particular, the turbidity parameter was chosen to be $\theta_{Tur} = 1$ NTU, because value of 1 NTU or less is recommended "to ensure effectiveness of disinfection and for good operation of the distribution system" (Health Canada, 2019). Therefore, all values above 1 are considered a threat to the health of the water distribution system in the area. Similarly, aluminum, strontium and barium parameters were also compared to the national standards for the drinking water; however, in this case no significant and immediate threats were identified. To clarify, the concentrations of aluminum, strontium and barium were within the maximum acceptable concentration limits as outlined by the national guidelines criteria (Health Canada, 2019). Nonetheless, an interesting data insight is still possible to observe by slicing in the values in the higher percentile, such as $\theta_{Al} = \theta_{Sr} = \theta_{Ba} = 60^{th}$ percentile. Ideally, in cases when datasets have larger sample sizes the 80-90^{th} percentile might be more appropriate to use. However, in this case 60^{th} percentile was a reasonable solution since it permits to preserve diverse connectivity of the traffic structure. By

increasing the θ parameter higher will single out only the extreme values of individual water impact parameters, and much connectivity structure will be lost.

It is important to mention that in this application the absence of the concentration records (i.e., N/A values) for Al, Tur, Sr and Ba in some hexagon cells (Table 4.7) were excluded from the analysis and treated as non-existent. Hence, those water parameter vertices cannot be found in the corresponding areal simplices. While such approach avoids additional changes to the raw water health data, it can be also appropriate to treat the unknown values as if they exceeded the θ parameters (i.e., as binary values of 1) following a precautionary principle. In other words, one can assume that all N/A records are greater than the threshold parameters in order to consider the worst-case scenario and follow up with the suitable preventative measures, if it turns out to be the case.

At this time, everything is ready for both ($\lambda \subseteq C \times W$) and ($\mu \subseteq C \times T$) relations to be grouped into the ($\rho \subseteq C \times (W \times T)$) relation in order to generate the combined binary matrix with its traffic mapped onto the backcloth. The output gets classified by each traffic parameter into four separate groups. A reader, however, should not be misled into thinking that they are separated in any way, as they remain in one relation ρ. In general, the process requires iteration through each element in λ and μ relations (Table 4.1, Table 4.8) to calculate the Cartesian product between sets W (water sites) and T (water parameters traffic). A pair, values of which are both equal to 1 gets assigned a new binary value 1, otherwise it is 0 and left out as blank. For example, area c13 has shown some problems with exceeding measurement of turbidity, given that this area only gets readings from the single water site W4, hence its equivalent row and column location of (c13, (W4 x Tur)) gets assigned a value of 1. The process further continues with the usual Q-analysis procedure, which includes calculation of the shared-face matrices and connected components for both $K_C(W \times T; \rho)$ and $K_{WxT}(C; \rho^{-1})$ simplicial complexes, the outcomes of which are discussed in greater detail in the results chapter (see Section 5.3).

4.4.2 Hierarchical Q-analysis

Up until this point, the analysis so far has been performed on a single fixed level of generalization, hierarchical level N for instance. As stated earlier (see Section 3.3), the significant advantage of Q-analysis methodology is its ability to distinguish cover sets at various levels of generality (Atkin, 1980, pp. 387–388; Gould & Johnson, 1980; Johnson, 1983a, 1983b, 1986; Johnson & Wanmali, 1981). Such property is of particular interest for the purposes of the water health application, since as the analysis gets more comprehensive it is not long until one identifies and requires other various hierarchical arrangements of the earlier defined cover sets C, W and T, as well as their λ, μ and ρ relations.

Given the hierarchical nature of DGGS the implications for the use of Q-analysis at various spatial resolutions is obvious, since one might require DGGS cells to be at higher or lower resolution depending on application needs. Even more so, the sets themselves must be allowed to exist at various hierarchical levels, where sets at hierarchical level N get covered by the sets at N+1 level. This means that relationship between sets at different hierarchical levels must be carefully specified, such that set elements at level N may be permitted to relate to various sets at level N+1. Such property is essential for the connectivity to be properly reflected in a binary relation structure, instead of forcing connectivity into unnatural partitions instead of cover sets (see Section 2.5.1, Figure 2.7) (Beaumont & Gatrell, 1982, p. 5; Gaspar & Gould, 1981, p. 190; Gould, 1980, p. 174; Gould & Johnson, 1980, p. 181).

If the attention is turned back to the water health application, it is not difficult to apply the same principles here as well. Primarily, the goal is to cover the study area with a set of discrete spaces known here as hexagon cells in order to capture another set of water site locations within their boundaries by covering the areas of their direct allocation. Thus, having made the set C (hexagon cells) as the cover set at some hierarchical level N, where N is equivalent to the resolution of the hexagon cells, such as H3 resolution 8 (Table B.1). What can be informative now is to apply the same procedure at the higher level of generalization, in order to study the changes in the backcloth connectivity structure. This requires the hexagon cover set to increase its cell size boundary and the resolution to adapt to the lower H3 resolution 7 (Table B.1). The backcloth hierarchy, in this case, amends to the N+1 level accordingly (Figure 4.8 cf. Figure A.1). For this matter, the hexagon aperture property is of great use here, since each aperture 7 hexagon cell at

level N is a child of a parent cell at level N+1. At the same time, the aperture is not restricted here, other kinds can be used as well, such as aperture 3 or 4 (see Section 3.3). This is consistent with the concept that elements at level N may be related to more than one element at level N+1, since it covers instead of strictly partitioning hexagon elements.

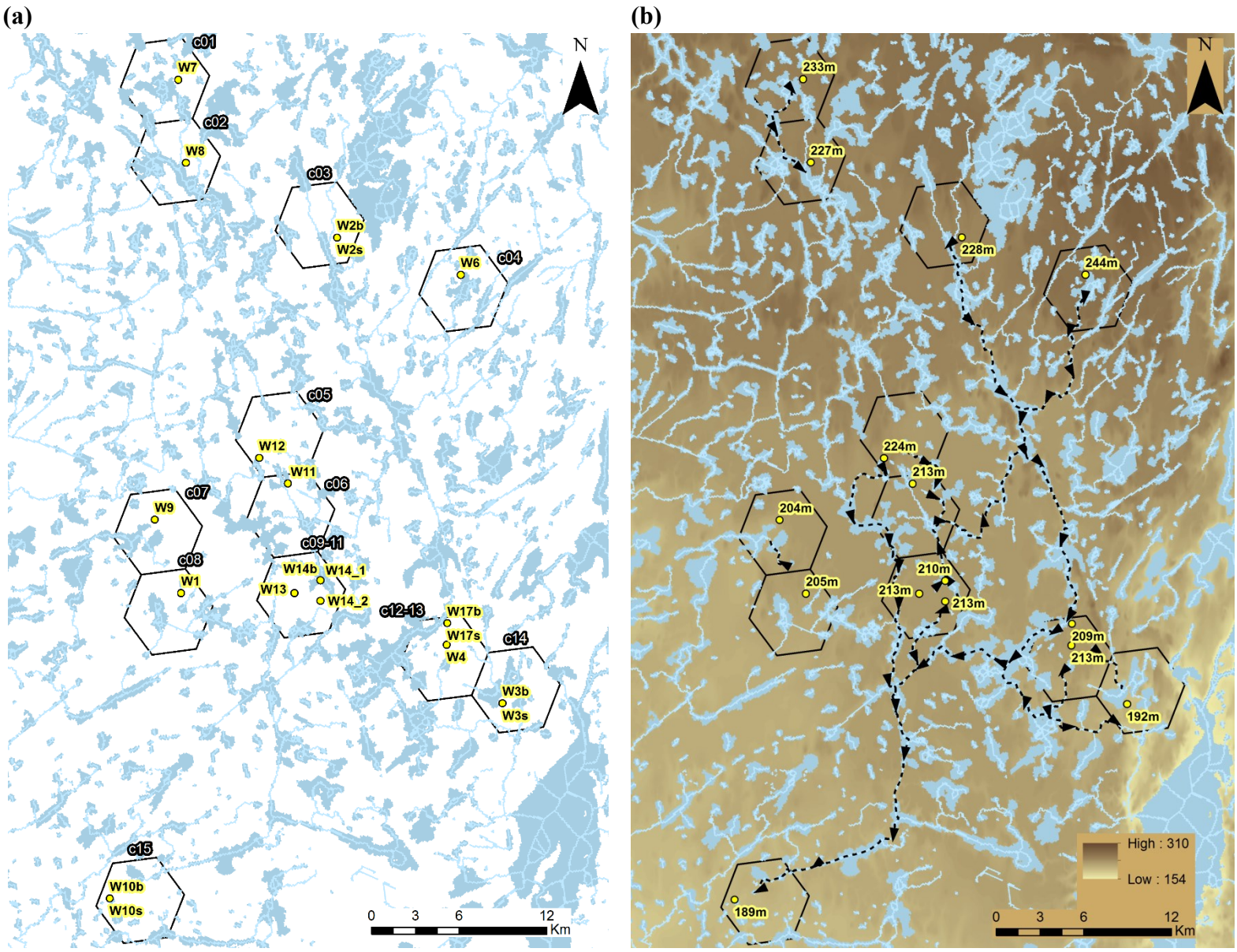


Figure 4.8. (a) The figure illustrates location of water sample sites (yellow labels) within DGGS hexagon cells at the broader generalization level N+1. (b) Flow direction and elevation information in the study area, as the primary parameters for determining connectivity structure between hexagons and water sites. The resolution of H3 cells is 7. Data sources: (Government of the Northwest Territories, 2019; Natural Resources Canada, 2015, 2019).

Similarly, traffic can be also hierarchically structured (see Section 3.3.2) (Gould & Johnson, 1980; Johnson, 1983b), if the interest is to study water parameters traffic at different levels of abstraction or aggregation. In this sense, a set of water descriptive parameters T (water parameters traffic) is created on another hierarchical level M, where M indicates concentration of specific water elements, such as T = {Al, Tur, Sr, Ba}. At levels M+1 and M-1 these parameters can be arranged into broader or finer categories via the molecular composition of each element respectively. In this context, it is suitable for the traffic parameters to be aggregated into the broader M+1 level groups each indicating the type, severity and general treatment guidelines for the measured water parameters. For example, set T at hierarchical level M can be further classified into treatment related, microbiological and inorganic chemical parameter groups at the M+1 level (Health Canada, 2019). The idea of a cover set is also well supported in the water traffic hierarchy, since it is possible for a water parameter to be related to more than one element at the higher level. Uranium, for example, though not used in this application but at the higher generalization level it is both inorganic chemical and radiological parameter (Health Canada, 2019). Given the rationale behind the backcloth and traffic hierarchies their sets at N, N+1, M and M+1 levels can be summarized in the following tables (Table 4.9, Table 4.10).

Table 4.9. Hierarchical hexagon cells at N and N+1 levels, and their corresponding changes to the cover set at N+1.

N	N+1
c01	c01
c02	c02
c03	c03
c04	c04
c05	c05
c06	c06
c07	c07
c08	c08
c09	c09-11
c10	
c11	
c12	c12-13
c13	
c14	c14
c15	c15

Table 4.10. Hierarchical water parameter elements at M and M+1 levels, and their corresponding changes to the cover set at M+1. Data source: (Health Canada, 2019).

M	M+1
Aluminum	Treatment related parameters
Turbidity	Microbiological parameters
Strontium	Inorganic chemical parameters
Barium	

Note that while being completely independent it is possible to combine different backcloth and traffic hierarchies (Gould & Johnson, 1980, p. 183), as long as such connection is possible to interpret in some contextual sense via the earlier established ($\rho \subseteq C \times (W \times T)$) relation. Accordingly, in the presented water application it is appropriate to combine backcloth and traffic hierarchies in four pairs, such as (N, M), (N, M+1), (N+1, M) and (N+1, M+1), since all of them are both contextually and algebraically valid. However, to avoid repetitive and similar outcome statements, only the (N, M) and (N+1, M+1) hierarchical pairs were interpreted in this study.

The pair (N, M) has already been studied in great detail (see Section 4.3), now consider the pair (N+1, M+1) as it introduces significant changes in the hierarchical structure of the backcloth and traffic. One must remember, however, that changes in the hierarchical configuration of cover sets will cause changes to the resultant connectivity matrix and to the Q-analysis output. To recap, given the earlier defined sets C, W and T, it is now possible to derive their cover sets at N+1 and M+1 levels, such that:

C^{N+1} = {c01, c02, c03, c04, c05, c06, c07, c08, c09-11, c12-13, c14, c15}

W = {W1, W2b, W2s, W3b, W3s, W4, W6, W7, W8, W9, W10b, W10s, W11, W12, W13, W14_1, W14b, W14_2, W17b, W17s}

T^{M+1} = {Treatment, Microbiological, Inorganic}

Note, how elements of the set W do not have hierarchical superscript assigned and remained unchanged. This is because set W is the non-hierarchical subject matter set of physical features (i.e., set of vector points representing water site locations). The use of the non-hierarchical set also guarantees consistency for merging backcloth and traffic, as well as hierarchical analyses (i.e., studying same subject matter for various spatial (set C) and attribute (set T) aggregations) (see Section 3.3). Consequently, three corresponding relations along with their direct and conjugate simplicial complexes are derived as shown:

$$\lambda^{N+1} \subseteq C^{N+1} \times W \rightarrow K_C{}^{N+1}(W, \lambda^{N+1}) \rightarrow K_W(C^{N+1}, \lambda^{-(N+1)})$$

$$\mu^{(N+1, M+1)} \subseteq C^{N+1} \times T^{M+1} \rightarrow K_C{}^{N+1}(T^{M+1}, \mu^{(N+1, M+1)}) \rightarrow K_T{}^{M+1}(C^{N+1}, \mu^{-(N+1, M+1)})$$

$$\rho^{(N+1, M+1)} \subseteq C^{N+1} \times (W \times T^{M+1}) \rightarrow K_C{}^{N+1}(W \times T^{M+1}, \rho^{(N+1, M+1)}) \rightarrow K_{WxT}{}^{M+1}(C^{N+1}, \rho^{-(N+1, M+1)})$$

Conceptually, relationship ρ can be captured by the following diagram (Figure 4.9). As discussed previously, the diagram here also illustrates the ability for the particular application to exist at the lower N-1 hierarchical level. The implications of the N-1 level are not covered in this research; nevertheless, such options are available to explore if data permits. Very likely, that hierarchical arrangement might not be always appropriate for other descriptive datasets, as some data may not generalize into higher or lower hierarchical levels, but exist only at the fixed hierarchy.

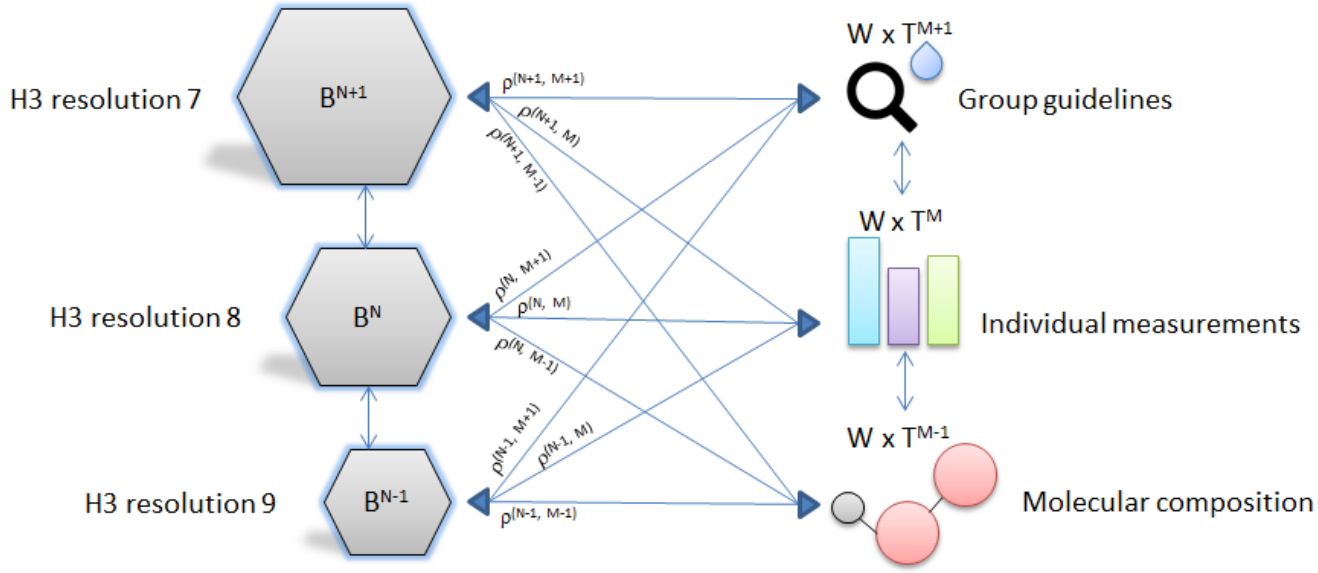


Figure 4.9. Graphical illustration of the hierarchical concept in the water health application, and formation of the corresponding ρ relation between the hierarchical sets. The image also demonstrates how backcloth and traffic can be integrated within DGGS.

It is also interesting to observe the difference and comparison between the hierarchical relation structures for N, N+1, M and M+1 levels, reflected in the resulting binary connectivity matrices for the λ^{N}, λ^{N+1}, λ^{-N}, $\lambda^{-(N+1)}$ (Table C.1, Table C.2, Table C.3, Table C.4), $\mu^{(N, M)}$, $\mu^{(N+1, M+1)}$, $\mu^{-(N, M)}$, $\mu^{-(N+1, M+1)}$ (Table D.1, Table D.2, Table D.3, Table D.4) and $\rho^{(N, M)}$, $\rho^{(N+1, M+1)}$, $\rho^{-(N, M)}$, $\rho^{-(N+1, M+1)}$ (Table E.1, Table E.2, Table E.3, Table E.4) relations. Their implications are further revealed through Q-analysis, and discussed in the results chapter (see Section 5.3).

Chapter 5 Results

In the previous chapters of this research much attention was put in formalizing a practical basis for how Q-analysis can fit into the overall DGGS perspective, and some of the advantages for doing so. In this chapter, the focus is put on reporting DGGS assessment, theoretical developments and Q-analysis output interpretation. Specifically, in Section 5.1 the chapter summarizes main findings for the operational capability of DGGS. Section 5.2 outlines theoretical advancements which were necessary in order for Q-analysis to be conformant with DGGS architecture. Finally, Section 5.3 provides interpretation of practical application and output of applied Q-analysis.

5.1 DGGS Assessment

The assessment and evaluation of available DGGS implementations have shown that there is great potential in utilizing such systems for the large-scale production purposes with the provided essential functional capabilities for data handling. At the same time, the study suggests further exploration and development of operational DGGS in order to promote enhanced capabilities for data analysis and interpretation, as well as evolution of the OGC abstract specification.

5.1.1 Functionality

Given the developed methodology (see Section 3.1.1) the study shows that it is possible to use selected DGGS solutions in a real-world application, such as an intersection analysis for example (Figure 5.1), as long as core functionality to convert geographic features into the grid cells, and the ability to extract their geometries are provided. Close exploration of these software indicate that all four libraries provide a set of methods for indexing data across hierarchies, which rely strictly on the coordinate information of the raw data. Once indexed, a geographic feature (e.g., point, line or polygon) is no longer defined by its latitude-longitude coordinates, but by a cell index which is embedded in a DGGS hierarchy.

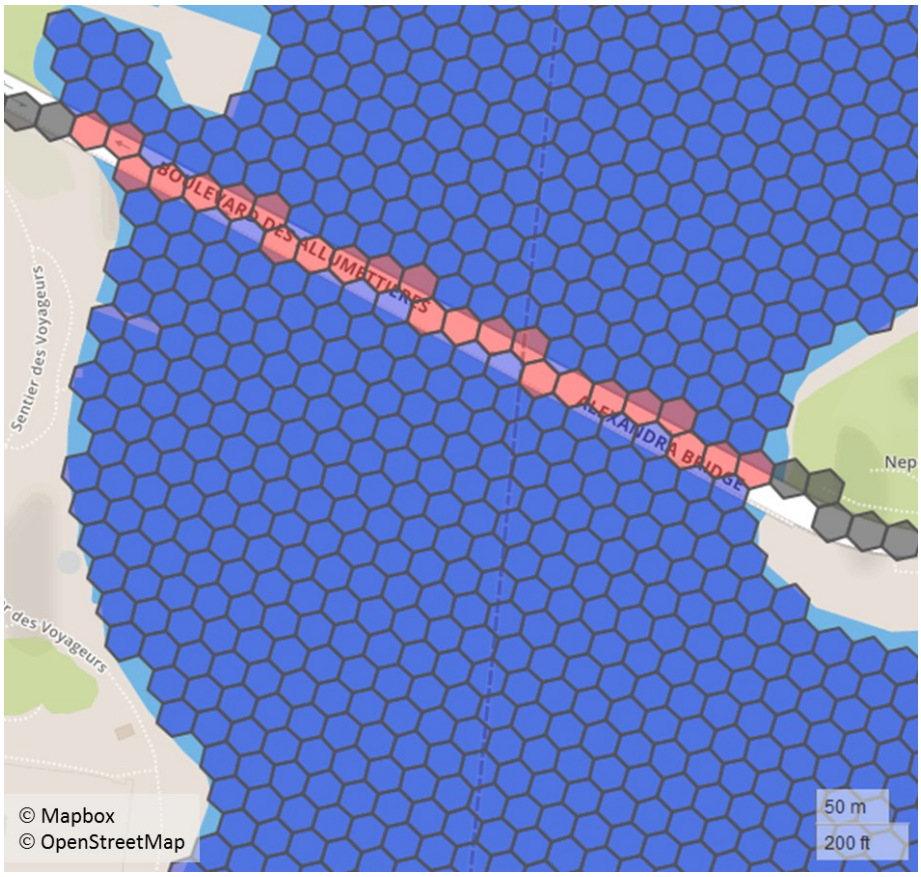


Figure 5.1. This figure demonstrates the intersection analysis as the complete procedure for integrating, processing and visualizing road network (in grey) and water (in blue) features in a DGGS framework. The results show the common intersection (in red) area for both features. This particular example was generated via the H3 library (Uber Technologies Inc., 2019).

One important observation which is necessary to consider here during the indexing stage is the potential loss of information. Given the index or cell ID, a grid system is capable of determining its resolution and centroid location for future processing or visualization, but not its original coordinate information. This means if the initial cell resolution is chosen to be too coarse, the location of a feature (e.g., point) will be replaced by the location of its cell centroid reducing potentially valuable precision (Figure 5.2). However, this same limitation can be viewed as an opportunity to consistently capture the uncertainty of the spatial location of the original data.

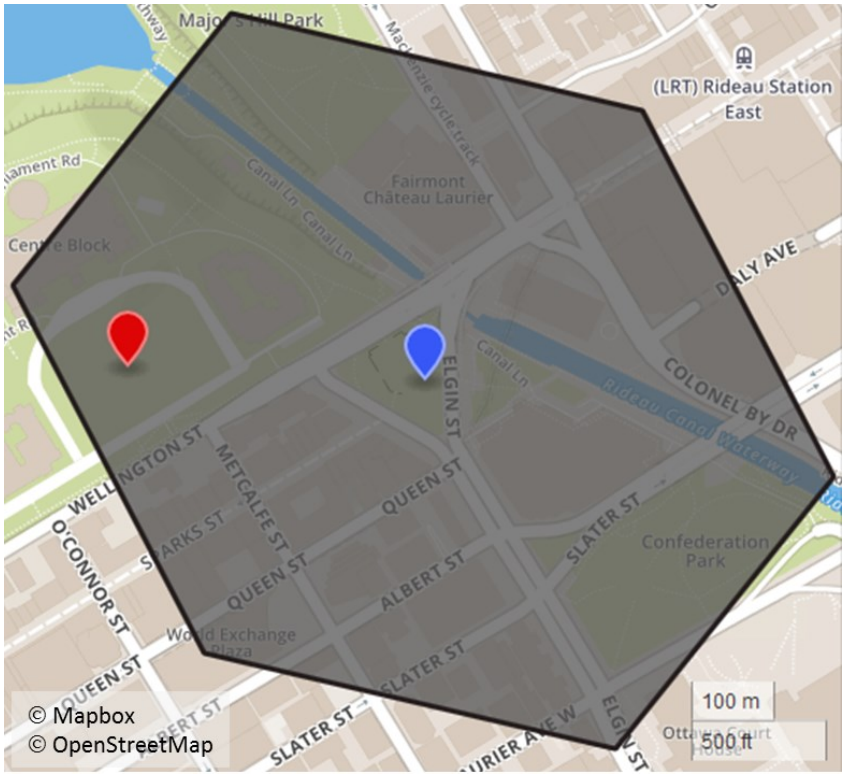


Figure 5.2. This figure illustrates a point example before and after its conversion to a DGGS cell. It indicates how a point (in red) earlier identified by its own geographic coordinates is now identified by its converted cell: its ID within the hierarchy and geographic location of the centroid (in blue). This illustration was created via the dggridR library (Barnes, 2016).

Theoretically, similar implementations can be extended to all DGGS libraries as long as the aforementioned functionality is provided. Certainly, the analytical capabilities of DGGS do not stop here and might in the future be enhanced in order to make use of the full potential of their hierarchical nature. Ideally, additional methods including distance, contiguity and spatial weights must be considered for other purposes, such as terrain analysis, travel cost methods or pattern distribution. This requires and leads to developing methods for handling more complex grid structures (Figure 5.3). In order to improve the efficiency of geometry representation and display, certain libraries have provided a way for compressing grid geometry into multivariable resolutions (Figure 5.3).

Amongst the presented DGGS implementations considered here only H3 and S2 libraries currently offer such functionality. Additional rules for handling hierarchical structure are necessary for successful data integration and spatial analysis procedures. Fundamentally,

algorithms must take advantage of the parent-child relationship to define rules on how user-integrated spatial data will be aggregated or distributed when these grids structures change dynamically. Thus, further investigation of this specific question is encouraged especially for non-congruent hexagon apertures of 3 and 4, as well as mixed apertures. More attention should be also paid to the aperture 7 hexagons data containment, and their overlapping and underlapping issues, due to the truncated hexagon precision at each consecutive hierarchy level (Brodsky, 2018). At the same time, it is worth remembering that Uber's own applications of aperture 7 hexagons caused no substantial obstacles for dynamic monitoring of their rides and prices on a city level scale (Brodsky, 2018).

(a) **(b)**

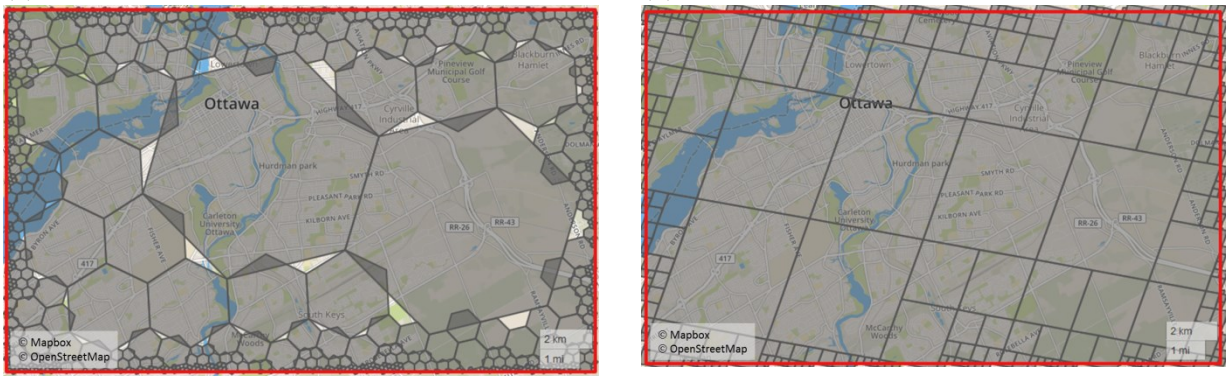


Figure 5.3. This figure demonstrates functionality of converting a polygon feature into a compressed grid structure via (a) H3 hexagon-based (Uber Technologies Inc., 2019) and (b) S2 square-based (S2Geometry, n.d.) hierarchies. These grid structures are of particular interest since they require additional methods for assigning and extracting data to and from them.

Another important consideration for practical DGGS application is technical support and corresponding user and development documentations, as they play an important role in use of these systems. In this sense, perhaps it was the most challenging to work with the OpenEAGGR library as its outdated build packages, technical documentations, usage methods and limited user base support prevented the use of its functionality to full potential. In relation to the other libraries the provided technical support was found to be sufficient in order to complete the set objectives. This summarizes an important discussion point that long term support of an active development community behind the software should be considered prior to adoption in real-world applications.

5.1.2 Performance

In considering library performance under increasing data volumes for individual aperture-shape pair combinations, the study discovered unexpected results. As shown (Figure 5.4), the assessment was applied on each library, shape and aperture separately in order to determine the ability to work with datasets of different size.

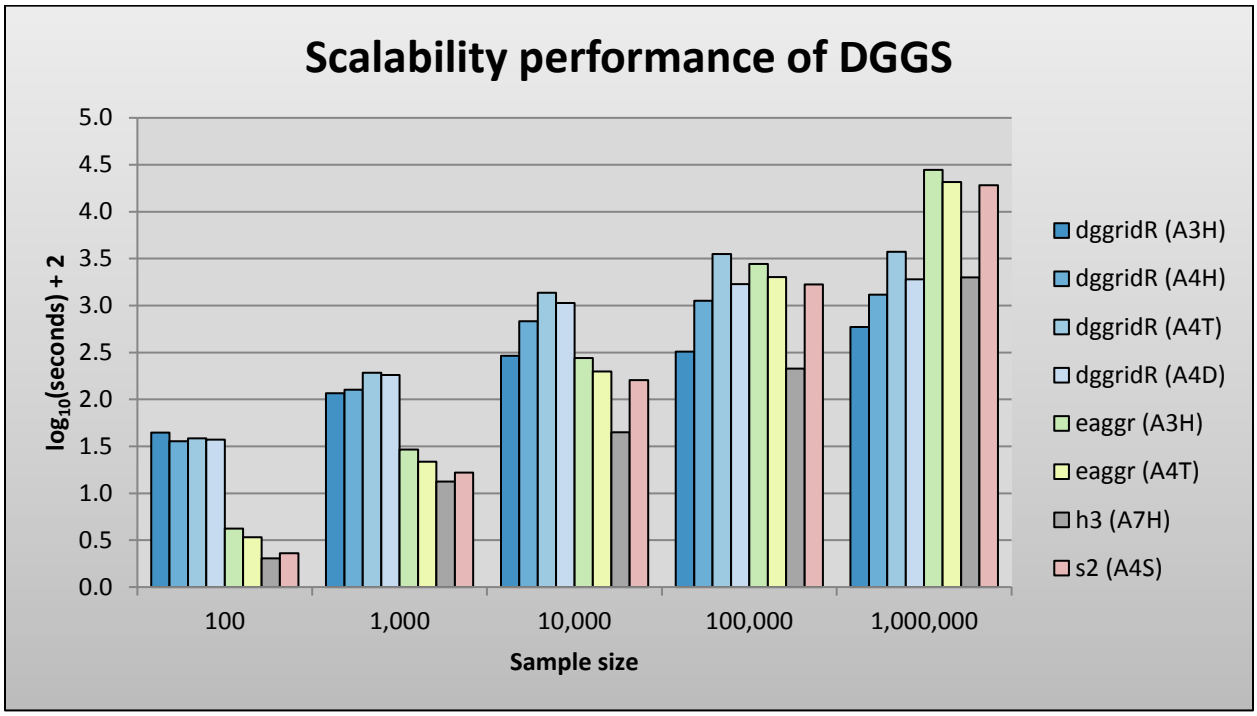


Figure 5.4. The graph indicates performance of individual DGGS aperture-shape pair combinations based on their availability and DGGS implementation considerations. The x-axis indicates the number of sample points. The y-axis represents the $\log_{10}$ function of time in seconds shifted 2 units up. The abbreviations in the legend represent the following: A – aperture, H – hexagon, T – triangle, D – diamond, S – square. For example, A3H stands for aperture 3 hexagon.

Results indicate that dggridR library is able to process large datasets faster than other libraries. While it is recognized that the timing component might be influenced by multiple factors, what is interesting is that dggridR library seems to perform substantially slower with smaller size datasets compared with OpenEAGGR, H3, S2, and vice versa for the larger datasets. According to this particular test and use of larger datasets, the dggridR analysis showed that aperture 3 hexagons were the most efficient amongst other dggridR groups (A4H, A4T, A4D)

outperforming even triangles. This particular finding is surprising to observe since triangles are generally thought to have faster performance compared to other shapes (e.g., hexagons, squares) (Peterson, 2017). On the contrary, triangles do tend to perform better than hexagons with OpenEAGGR, likely due to differences between the software implementations.

Regarding the remaining OpenEAGGR, H3 and S2 libraries, their performance showed consistency: their relative performance remained the same with H3 library being consistently faster. However, direct comparison between them or any other libraries was not possible. Overall, the performance test shows that all the above implementations are suitable and operational for working with large datasets within a reasonable time frame. It is important to mention that due to the large range between the measured values, the results were scaled using $\log_{10}(\text{seconds}) + 2$ (Figure 5.4).

5.1.3 OGC Compliance

The final part of this section presents a comparative analysis of DGGS implementations against the OGC abstract specification, which requires a DGGS data model to follow the core requirements necessary for defining a reference system across a hierarchy as well as functionalities that allow cell management and integration with external applications (Table 5.1 cf. Table 2.1). In summary, results show that it is not possible to confirm fulfilment of all OGC requirements. H3 library is the closest one to the completion of the abstract specification. However, since it is impossible to confirm criterion 7, a statement confirming the full compliance of the H3 library cannot be made. In fact, criterion 7 cannot be confirmed for any of the libraries since technical documentation is not provided to demonstrate its fulfilment. Regarding other unfulfilled criteria, dggridR does not seem to support hierarchical navigation across a hierarchy (criterion 15), nor present integration with external applications (criteria 17, 18). Similarly, OpenEAGGR falls short in meeting the area preservation requirement (criteria 3 and 5) (Figure 5.5), whereas S2 lacks the equal area property (criteria 8 and 9).

Table 5.1. This table outlines the final assessment of the chosen DGGS libraries and evaluates them according to the OGC abstract specification for the DGGS data model. The information in this table was partially supported by the following sources for dggridR (Barnes, 2016; Sahr, 2018), H3 (Brodsky, 2018; Uber Technologies Inc., 2019), OpenEAGGR (Bush, 2017; Riskaware Ltd., 2019) and S2 (Kreiss, 2016; S2Geometry, n.d.; Veach, 2017; Zeroghan, 2019). To be read in conjunction with Table 2.1.

Criteria	dggridR	H3	OpenEAGGR	S2
1	Partially satisfied, criteria 7, 15, 17, 18 could not be asserted	Partially satisfied, criterion 7 could not be asserted	Partially satisfied, criteria 3, 5, 7 could not be asserted	Partially satisfied, criteria 7, 8, 9 could not be asserted
2	All able to cover the complete surface of the Earth			
3	Overlapping or underlapping cells were not observed	Overlapping or underlapping cells were not observed	Ideally this requirement is fulfilled, however practical application indicates the opposite	Overlapping or underlapping cells were not observed
4	All capable of generating grids at various resolutions			
5	Anomalies were not observed across hierarchy	Anomalies were not observed across hierarchy	The requirement is not met due to the case in criterion 3	Anomalies were not observed across hierarchy
6	Hexagon, triangle and diamond shapes are simple polygons	Hexagon is a simple polygon	Hexagon and triangle shapes are simple polygons	Square is a simple polygon
7	Unable to determine for all, excluded from the evaluation			
8	Claimed to generate grids of equal area. The ratio between pentagon and hexagon cells is 5/6	Each subsequent cell has 1/7 of the area from its parent cell. Equal area property of pentagons is not explicitly stated	Equal area cells are generated at each resolution level. Equal area property of pentagons is not explicitly stated	Does not claim equal area property, however the average cell size is guaranteed to be within a factor of 1.5
9	Icosahedron	Icosahedron	Icosahedron	Integrates hexahedron/cube polyhedron of unequal size

Table 5.1. Continued.

Criteria	dggridR	H3	OpenEAGGR	S2
10	Hexagons with apertures 3, 4 and mixed apertures 4-3; triangles with aperture 4; and diamonds with aperture 4	Uses hexagons with aperture 7 grid partitioning	Hexagons with aperture 3 and triangles with aperture 4	Uses squares with aperture 4 grid partitioning
11	Hierarchy based indexing method	Hierarchy based indexing method	Uses offset coordinate indexing for hexagons and hierarchical indexing for triangles	Uses Hilbert space-filling curve
12	All use unique indexes for each of their cells			
13	Location of a cell is assigned to its centroid	Location of a cell is assigned to its centroid	Location of a cell is assigned to its centroid	Cell center specifies its position and subdivision level
14	All provide various quantization methods that allows for conversion, storage and retrieval of various spatial features within DGGS framework.			
15	Hierarchical and neighbourhood methods are missing	Equipped with method for hierarchical and neighbourhood navigation	Full support of hierarchical navigation, and only partial support of neighbourhood navigation	Equipped with method for hierarchical and neighbourhood navigation
16	Supports method for intersection between cells and shapefile	Supports methods for basic spatial operations	Supports large variety of spatial analysis methods	Supports methods for basic spatial operations
17	Examples of data queries are not known	Integration with geojson2H3 library and kepler.gl application permits data query	Integration with PostgreSQL, PostGIS and Elasticsearch software permit data query	Integration with Sidewalk Labs and Pokémon Go meets this requirement

Table 5.1. Continued.

Criteria	dggridR	H3	OpenEAGGR	S2
18	Examples of data broadcast are not known	Integration with geojson2H3 library and kepler.gl application permits data broadcast	Integration with PostgreSQL, PostGIS and Elasticsearch software permit data broadcast	Integration with Sidewalk Labs and Pokémon Go meets this requirement

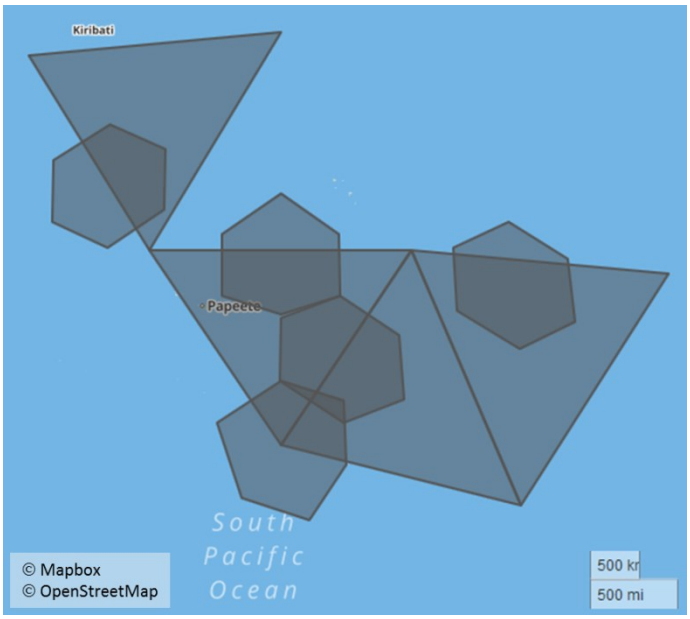


Figure 5.5. This figure illustrates that hexagon generation via the OpenEAGGR library (Riskaware Ltd., 2019) is unable to meet the requirements for positional uniqueness and area preservation because of the overlapping hexagon cells. The same conclusion cannot be made regarding triangles as shown.

As a discussion point, the study reflects on these criteria and suggests their further exploitation in order to encourage their refinement in future releases. In particular, criterion 1 is the requirement that includes two major subgroups: reference frame elements and functional algorithm elements. This makes it difficult to evaluate such an all-inclusive criterion independently, since its fulfilment also guarantees the fulfilment of the remaining criteria. Therefore, it is suggested to remove this criterion as an individual requirement and use it as a validation check for the entire DGGS in the end of the comparative process. The wording of criterion 7 was hard to follow,

perhaps it can be reworded to make it easier for developers to demonstrate they have met the criteria. Criterion 9 appears to be a product of criterion 3, since if the positional uniqueness is preserved for the initial tessellation of a based polyhedron (criterion 3), this would also satisfy the fulfilment for the initial tessellation and equal area property (criterion 9). Therefore, perhaps these criteria can be merged. Similarly, criteria 11 and 12 seem to enforce similar requirement regarding spatial indexing, unique identifier and cell addressing. It is also suggested to combine these two in future releases, since the use of an indexing method also implies the use of unique identifiers. Criteria 17 and 18 are also associated in the way that they are complimentary to each other, since the existence of methods to receive, interpret and process external data queries (criterion 17) requires methods to return the results back to external application (criterion 18) which made the initial query request. Thus, existence of one requires the existence of the other. It is advised to merge these requirements into one as a communication criterion.

In general, it is suggested to improve the current OGC abstract specification by clarifying some ambiguity in its definition and adding more contextual information and methods on how to implement the criteria. This would allow for a more constructive approach to maintain the criteria by software developers and provide users the ability to cross validate them. Some of these issues are occurring due to the fact that despite aiming to create a clear standard, a more comprehensive and broad approach was chosen instead to accommodate the need for the linkages with other data standards and DGGS, given the early stage of development of DGGS conceptual model and its implementation standard (OGC, 2019).

5.2 Q-analysis for DGGS Data Infrastructure

The appearance of DGGS has led to the development of new scalable solutions for data integration, analysis and visualization, in order to provide enhanced analytical capability and understanding of the global data complexity. Therefore, the need for the development of additional operations is demanded, in order to provide methods capable interpreting data in a form compatible with the core DGGS infrastructure (OGC, 2017). Apart from this, recent studies have also recognized the necessity for new and improved data driven methodologies, in order to

overcome challenges associated with complex data systems and formalization of geographic knowledge in understandable and explicit manner (Miller & Goodchild, 2015).

Recognizing these requirements, this research makes one of its main contributions to the DGGS paradigm by proposing the use of Q-analysis set theory as a tool capable of preserving and concisely describing complex system interactions. The suggested integrated methodology attempts to provide theoretical argumentation as well as a case study for the concurrent use of Q-analysis within future DGGS implementations and developments. In particular, much effort was dedicated into formalizing theoretical concepts of cover sets and resulted simplicial complexes with DGGS aperture and hierarchical properties. These developments extend the functional capability of DGGS for data management and interpretation with explicit and rich algebraic techniques proficient to analyse interactions within complex systems and datasets. In fact, the hierarchical properties of DGGS and subject matter features are considered as multilevel backcloth systems, whereas spatial data associated with DGGS cells as multilevel traffic. Thus, Q-analysis is made conformant with DGGS architecture for the study and analysis of the multilevel backcloth-traffic systems within the data rich environment. The practical execution of the proposed theoretical argumentation has shown that Q-analysis is in fact suitable and advantageous to the growth of DGGS, as well as fitting for the spatial analysis.

Additionally, the study also outlines some of the attractive features of the Q-analysis methodology for exploring a system under continuous changes. A system in this case can be understood as a set of interactions or connectivity between its components or attributes, which are significant in understanding its behaviour. The study confirms the benefits of using simplicial complexes and their geometric representation as a way to define system behaviour. Their abstract visualization provides additional knowledge about system complexity and its interactions, not achievable by mere mapping of information assigned to DGGS cells on a physical space. Lastly, the study also puts emphasis on the value of Q-analysis for using relational approach, which is claimed to be data friendly and retains much of the original information via explicit definition of cover sets and their unconstrained interactions.

5.3 Interpreting Hierarchical Q-analysis

A particular advantage of formalizing Q-analysis methodology within a DGGS data infrastructure is the ability to have a consistent configuration of space across complex network systems suitable for comparison at various hierarchical levels. In other words, a clear hierarchical arrangement of sets can be established and their corresponding structures compared conveniently, since DGGS cells at higher level of generalization cover those at the lower level of generalization. This property is considered important as it provides grounds for comparison and useful insight of a system under changes to its connectivity structure caused by the changes in spatial configuration and descriptive domains (i.e., spatial resolution). In this case, the Q-analysis output of the water quality and water health application is interpreted at different hierarchical levels, including comparison of the structural characteristics for backcloth, traffic and backcloth-traffic mapping. It is interesting to learn how Q-analysis accurately reflects physical connectivity of the water system in the study area, and preserves much of its interaction and transmission capability when provided with well-defined sets and water flow connectivity. In addition, the analysis reveals the role and significance of individual areas in the multidimensional connectivity structure, including grounds for the spread of pollution and its possible likelihood.

5.3.1 Backcloth Connectivity

The major results of this section are outputs of the hierarchical Q-analysis for direct (Table 5.2) and conjugate backcloth complexes (Table 5.3), as well as their geometric realizations (Figure 5.6) for both N and N+1 hierarchical levels. The analysis is mainly based on the backcloth connectivity structure between spatial units (i.e., hexagons) and physical features (i.e., water sites) defined for spatial aggregation levels N (Table C.1, Table C.3) and N+1 (Table C.2, Table C.4).

Considering the output of the direct Q-analysis for the backcloth connectivity structure at level N the study has identified component {c12} to have the highest dimensionality at q = 7 (Table 5.2). This indicates that simplex σ_7(c12) and its corresponding area c12 is a key location in the complex which connects the largest number of water sites (i.e., 8) through the water network in

the study region. This finding is important since it means that water sites in this location (i.e., W17b and W17s) have a significant role in water quality monitoring, removal or malfunction of which can lead to the structure being disconnected and make the pollution tracing more difficult in the water system. Similarly, components {c15}, {c14} and {c10} enter the structure at q = 5, q = 4 and q = 3 respectively, and can be also considered as high dimensional since these locations with their corresponding water sites connect the entire network system into one multidimensional structure. In the context of water quality application these areas are of particular importance, since they can be used to monitor changes of water health, as they are the most connected within the water quality monitoring system. Thus, much water flow is passing through these areas. Specifically, it seems that simplices σ_7(c12), σ_5(c15), σ_4(c14) are connected via vertices <W17b, W17s>, whereas simplices σ_7(c12) and σ_3(c10) via vertices <W14_1, W14b> (Figure 5.6a). This implies that water quality measurements collected by these water sites (i.e., vertices) have a potential to propagate down the stream including cases when water happen to be contaminated. In particular, these areas form a connectivity chain, such that c10 flows into c12, and c12 flows into c15 and c14 for the hierarchical level N (Figure A.1). The connectivity chains become even more apparent at the lower q-dimensions, such as at q = 1 and q = 0 (Table 5.2). Here, Q-analysis identifies distinct connected components which include chain of simplices connected via water sites, capable of transmitting similar water quality information (i.e., traffic) through the water network. Primarily, the entire structure consists of three connected components {c01, c02}, {c07, c08} and {c03, c04, c05, c06, c09, c10, c11, c12, c13, c14, c15} which suggest that their corresponding areas are disconnected and do not share the same water site locations. Thus, it is safe to assume that these areas might be substantially different in terms of water quality characteristics, including cases of water pollution.

Comparing the output of the direct Q-analysis for the backcloth connectivity at N+1 hierarchical level (Table 5.2), the study noticed that connectivity structure changed to be more connected. The hierarchy and changes in connectivity are explained by the increased areal coverage of each spatial unit (i.e., hexagon) while the scale of water sites remained unchanged (Figure A.1). It is clear that connectivity shifts towards simplices, such as σ_8(c15), σ_6(c12-13), σ_5(c06) and σ_4(c14) since they are identified as separate components at q = 8, q = 6, q = 5 and q = 4 respectively (Table 5.2). This is different compared to the Q-analysis output at level N, since at N+1 level

component {c15} was identified as the one with the highest dimension at q = 8 instead. This is due to the increased spatial aggregation of hexagons which allowed more isolated water sites to have access to different water streams. The maximum q-value (i.e., 8) has also increased accordingly signifying that more water sites have direct access to the σ_8(c15) simplex. For example, simplices σ_0(c09), σ_3(c10) and σ_0(c13) at level N can now have a direct water flow connectivity to simplex σ_8(c15), as composite of σ_4(c09-11) and σ_6(c12-13) simplices at N+1 levels respectively (Figure 5.6b cf. Figure A.1). At the same time, simplex σ_0(c05) at level N+1 has become connected to the simplex σ_8(c15) via the intermediate simplex σ_5(c06) instead, which was connected directly at level N. This suggests that water quality characteristics of simplex σ_8(c15) are described by a different set of water sites at level N+1. While recognizing that such aggregation from level N to N+1 might lead to suppression of important spatial information, the analysis of this sort can be also viewed as an uncertainty analysis which can be considered in the water health application to account for additional water flow connectivity. In this context, the analysis at different hierarchical levels is also a trade-off between information precision and accuracy. To elaborate, the output at level N appears to be more precise but lacks information accuracy. Whereas the analysis at level N+1 is less precise, due to aggregation of spatial information, but more accurate since it is possible to have more confidence in the collected data (i.e., more water sites describe one spatial unit). In general, the structure of the direct simplicial complex at N+1 level (Figure 5.6b) appears to be more connected and less fragmented which implies that more water sites (i.e., vertices) are being shared amongst the hexagons (i.e., simplices). For example, a first chain of connected simplices is started to appear much earlier in the structure, such as component {c06, c09-11} at q = 4 (Table 5.2) to indicate that each simplex share 5 water sites with one another.

The conjugate Q-analysis at level N reveals additional insight into the water network system from the site-oriented point of view (Table 5.3), as opposed to the previously discussed cell-oriented perspective of the direct Q-analysis at level N. Here, the analysis report relatively low q-values, compared with the direct Q-analysis output. This is explained by the fact that most water sites can be traced from their initial location to next immediately connected area in the network. Considering that majority of sites flow into one direction only, the resulting connectivity structure turns out to be less connected. Nevertheless, much contextual information is still

possible to retrieve from these analyses. Mainly, the study identifies components {W12}, {W14_2} and {W17b, W17s} at q = 2 (Table 5.3), which show that their corresponding simplices can, in fact, flow in two different directions. This is clearly visible on the geometric representation of the complex (Figure 5.6c), as well as the study area map (Figure A.1). This implies that destination areas are related by sharing similar water characteristics, while not having direct water flow connectivity between each other. Such property can be very useful in cases when spread variability (i.e., number of connected edges) must be determined or predicted. In other words, area destinations are capable of inheriting similar water properties from their shared sources. In general, due to the large number of distinct connected components at q = 1 (Table 5.3) the complex can be considered as fragmented, which means that water flows mostly in one direction only. For example, it flows towards the earlier identified high dimensional areas, such as c12, c15, c14 and c10 (Figure A.1). It is also worth noting that both direct (Table 5.2) and conjugate Q-analysis (Table 5.3) at level N show consistent results, since both identify three connected components at the lowest dimension q = 0. The direct Q-analysis identifies hexagon cells, whereas conjugate Q-analysis – water sites that are located within hexagon boundaries.

As for the conjugate Q-analysis at N+1 level (Table 5.3), the procedure continues to confirm results determined from the direct Q-analysis, as well as to reveal new additional details. Similarly to the direct Q-analysis at N+1 level, the hierarchy of the conjugate Q-analysis at N+1 level is also characterized by the increased areal coverage of the spatial units (i.e., hexagons) while the scale of water sites remained unchanged (Figure A.1). The connectivity of the water network shows that components {W11}, {W13, W14_1, W14_2, W14b} and {W4, W17b, W17s} at q = 2 and their corresponding simplices are able to flow in two different directions. The water flow connectivity here is well reflected by the geometric representation of the conjugate simplicial complex (Figure 5.6d). From the graph, it is also possible to identify vertices c12-13, c09-11 and c06 to have the highest spread variability, which is slightly different from the analysis at level N. In particular, at N+1 level it shows that simplex σ_2(W11) can now flow in two different directions, whereas simplex σ_1(W11) at level N in only one (Figure 5.6c cf. Figure 5.6d). This phenomenon is explained by the increased uncertainty of spatial units at N+1 level. A particularly interesting behaviour was identified in the {W11, W13, W14_1, W14_2, W14b} connected component at q = 1. Here, the analysis detects a loop, or circular water flow connectivity between vertices c06 and c09-11. This same behaviour is also reflected in a direct

Q-analysis at level N+1 as {c06, c09-11} component at q = 4 (Table 5.2). In other words, c06 flows into c09-11, and vice versa (Figure A.1). Such connectivity was not possible to identify at level N. Perhaps analysis on N+1 level infers additional properties of the physical water flow connectivity not visible or not identifiable at level N, due to the stricter and more constrained spatial configuration. Of course, such connectivity must be confirmed by the additional in-depth hydrological analysis of the water flow behaviour in the study region. At the lowest dimension q = 0, the results seem to be consistent at both N and N+1 levels, since three separate connected components composed of same simplices were also identified (Table 5.3). Results confirm the fact that no direct water flow is possible between the areas within identified connected components. Even after increasing the areal uncertainty, results show that these areas are indeed far away from each other, and thus less likely to share similar water quality characteristics.

Table 5.2. Q-analysis output of $K_C^N(W, \lambda^N)$ and $K_C^{N+1}(W, \lambda^{N+1})$ direct simplicial complexes. The curly brackets signify a connected component and its simplices. The N/A value indicates absence of the corresponding q-value in the complex.

q-value	$K_C^N(W, \lambda^N)$	$K_C^{N+1}(W, \lambda^{N+1})$
8	N/A	{c15}
7	{c12}	{c15}
6	{c12}	{c12-13}, {c15}
5	{c12}, {c15}	{c06}, {c12-13}, {c15}
4	{c12}, {c14}, {c15}	{c06, c09-11}, {c12-13}, {c14}, {c15}
3	{c10}, {c12}, {c14}, {c15}	{c06, c09-11, c15}, {c12-13}, {c14}
2	{c10}, {c12}, {c14}, {c15}	{c06, c09-11, c12-13, c14, c15}
1	{c02}, {c06}, {c07}, {c03, c10, c12, c14, c15}	{c02}, {c07}, {c03, c06, c09-11, c12-13, c14, c15}
0	{c01, c02}, {c07, c08}, {c03, c04, c05, c06, c09, c10, c11, c12, c13, c14, c15}	{c01, c02}, {c07, c08}, {c03, c04, c05, c06, c09-11, c12-13, c14, c15}

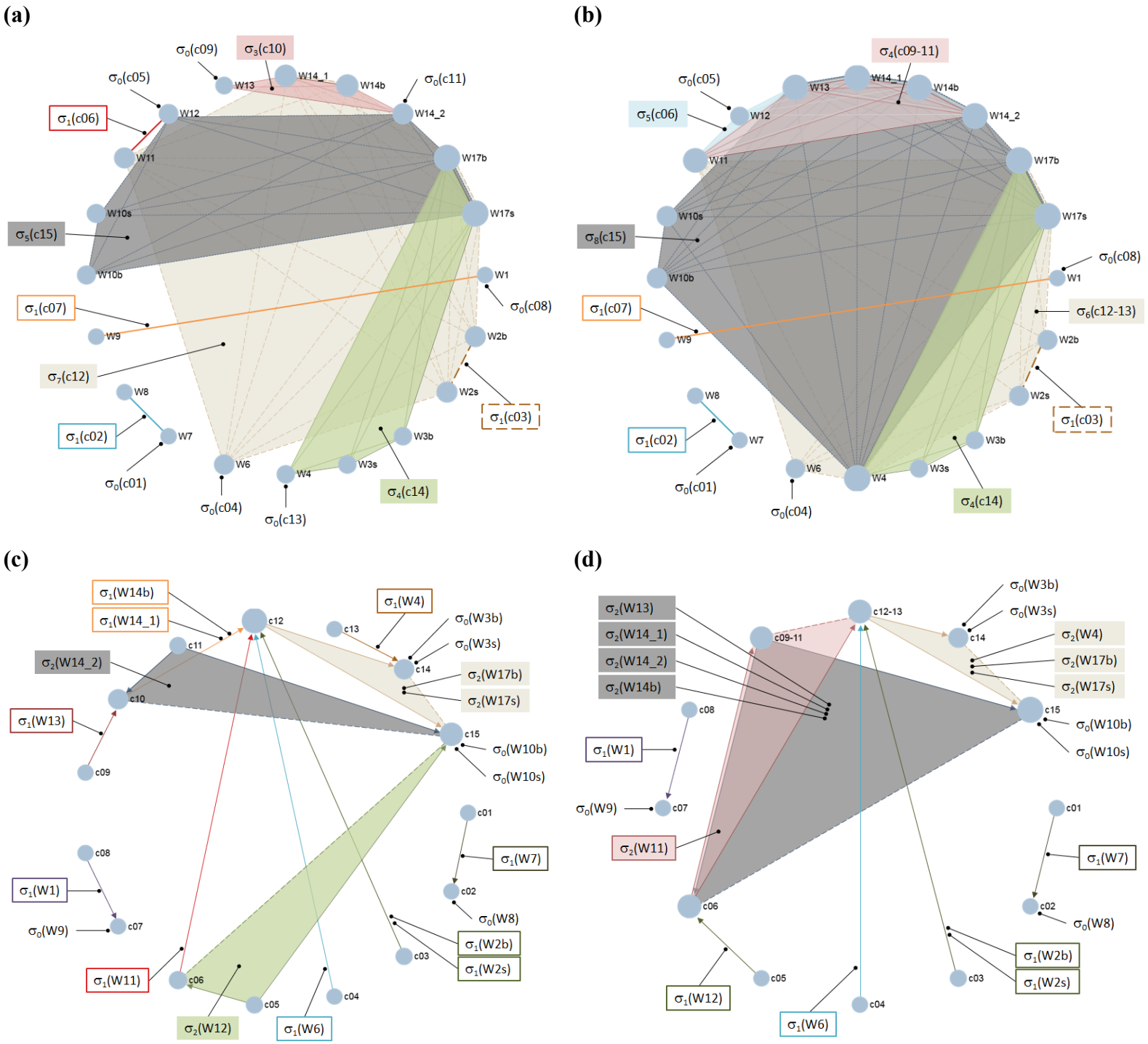


Figure 5.6. Geometric representation of (a) $K_C^N(W, \lambda^N)$ (b) $K_C^{N+1}(W, \lambda^{N+1})$ (c) $K_W(C^N, \lambda^{-N})$ (d) $K_W(C^{N+1}, \lambda^{-(N+1)})$ simplicial complexes.

Table 5.3. Q-analysis output of $K_W(C^N, \lambda^{-N})$ and $K_W(C^{N+1}, \lambda^{-(N+1)})$ conjugate simplicial complexes. The curly brackets signify a connected component and its simplices.

q-value	$K_W(C^N, \lambda^{-N})$	$K_W(C^{N+1}, \lambda^{-(N+1)})$
2	{W12}, {W14_2}, {W17b, W17s}	{W11}, {W13, W14_1, W14_2, W14b}, {W4, W17b, W17s}
1	{W1}, {W2b, W2s}, {W4}, {W6}, {W7}, {W11}, {W12}, {W13}, {W14_1, W14b}, {W14_2}, {W17b, W17s}	{W1}, {W2b, W2s}, {W6}, {W7}, {W11, W13, W14_1, W14_2, W14b}, {W12}, {W4, W17b, W17s}
0	{W1, W9}, {W7, W8}, {W2b, W2s, W3b, W3s, W4, W6, W10b, W10s, W11, W12, W13, W14_1, W14_2, W14b, W17b, W17s}	{W1, W9}, {W7, W8}, {W2b, W2s, W3b, W3s, W4, W6, W10b, W10s, W11, W12, W13, W14_1, W14_2, W14b, W17b, W17s}

5.3.2 Traffic Connectivity

Traffic is also a fundamental concept in Q-analysis, which is a term that addresses various descriptive features (i.e., attribute information) related to the geometric structure of the backcloth. The results of this section are mainly focused on interpreting water quality parameters (i.e., traffic), their q-connectivity and hierarchical aggregation (Table D.1, Table D.2, Table D.3, Table D.4) within the scope of water health application and DGGS.

By itself traffic interpretation might not be very informative as the relation between spatial units and water parameters mainly outline areas where θ parameters exceed their limits. Nonetheless, it is still possible to get some insight of the collected data when its connectivity is investigated. This is clearly visible from the output of the direct Q-analysis for both (N, M) and (N+1, M+1) hierarchical levels (Table 5.4). Here, the areas with the highest number of exceeding θ parameters appear at high q-values, whereas the less polluted areas enter the structure subsequently. The q-value indicates (q + 1) parameters above their θ thresholds. Note, the connected components in the direct Q-analysis (Table 5.4) do not imply water flow connectivity between their spatial units, but simply state that all corresponding simplices have equal or greater than (q + 1) parameters above their θ limits.

More comprehensive interpretation can be achieved from the output of the conjugate Q-analysis (Table 5.5). Here, Q-analysis identifies the most widespread water parameters or characteristics which affect water quality in the study area. Primarily, the analysis has determined component {Tur} (i.e., turbidity) to be the most widespread at q = 7. In fact, q-value signifies that (q + 1) areas have exceeded the maximum allowable measure of turbidity in the study region. At the lower q-values (e.g., from q = 5 to q = 3), it also seems that two distinct connected chains {Al, Tur} and {Ba, Sr} have been formed at the hierarchical level (N, M). The fact that they tend to appear as separate components suggest that they are affecting a slightly different set of areas, which can also mean that they have originated from two different sources of pollution.

Interestingly enough, at the conjugate hierarchical level (N+1, M+1) component {Ba, Sr} appears as {inorganic}, which indicates that both Ba (barium) and Sr (strontium) are inorganic chemicals. Thus, perhaps their q-connectivity says something about their similar chemical properties, which allows them to be combined at the next hierarchical level. By inference, this would also suggest that there is some similarity between Al (aluminum) and Tur (turbidity), despite the fact that Al is a treatment related parameter and Tur is a microbiological parameter at (N+1, M+1) level. Regardless, this statement can be true since by definition turbidity measurement can contain particles of both inorganic and organic elements (Health Canada, 2019). Thus, it is very likely that spike in Tur measurement can lead to the spike in Al concentration. Either way, additional analysis of chemical and physical water properties would be necessary to perform in order to confirm these statements.

Table 5.4. Q-analysis output of $K_C^N(T^M, \mu^{(N, M)})$ and $K_C^{N+1}(T^{M+1}, \mu^{(N+1, M+1)})$ direct simplicial complexes. The curly brackets signify a connected component and its simplices. The N/A value indicates absence of the corresponding q-value in the complex.

q-value	$K_C^N(T^M, \mu^{(N, M)})$	$K_C^{N+1}(T^{M+1}, \mu^{(N+1, M+1)})$
3	{c05, c09}	N/A
2	{c05, c06, c09}	{c05, c09-11}
1	{c01, c02, c04, c05, c06, c07, c08, c09, c10, c12}	{c01, c02, c04, c05, c06, c07, c09-11, c12-13}
0	{c01, c02, c04, c05, c06, c07, c08, c09, c10, c12, c13}	{c01, c02, c04, c05, c06, c07, c08, c09-11, c12-13}

Table 5.5. Q-analysis output of $K_T^M(C^N, \mu^{-(N, M)})$ and $K_T^{M+1}(C^{N+1}, \mu^{-(N+1, M+1)})$ conjugate simplicial complexes. The curly brackets signify a connected component and its simplices.

q-value	$K_T^M(C^N, \mu^{-(N, M)})$	$K_T^{M+1}(C^{N+1}, \mu^{-(N+1, M+1)})$
7	{Tur}	{microbiological}
6	{Tur}	{microbiological}
5	{Al, Tur}, {Ba, Sr}	{microbiological, treatment}
4	{Al, Tur}, {Ba, Sr}	{microbiological, treatment}, {inorganic}
3	{Al, Tur}, {Ba, Sr}	{microbiological, treatment, inorganic}
2	{Al, Ba, Sr, Tur}	{microbiological, treatment, inorganic}
1	{Al, Ba, Sr, Tur}	{microbiological, treatment, inorganic}
0	{Al, Ba, Sr, Tur}	{microbiological, treatment, inorganic}

5.3.3 Backcloth-Traffic Mapping

As the final product of this project the study attempts to perform the mapping of the hierarchical traffic onto the hierarchical backcloth (Table E.1, Table E.2, Table E.3, Table E.4) in order to get a sense of water movement and changes of water quality characteristics (i.e., traffic) through the water network system (i.e., backcloth). The results are reported as the output of the direct (Table 5.6) and conjugate (Table 5.7) Q-analysis for both (N, M) and (N+1, M+1) hierarchical levels, as well as geometric representation of the conjugate simplicial complexes (Figure 5.7). At the same time, parallels between both hierarchical levels are made for comparison and interpretation of the output. In a general sense the process of mapping reveals areas where individual water parameters exceed their defined θ thresholds.

Given the output of the direct Q-analysis at hierarchical level (N, M) the study shows that {c12} is the single highest dimensional component found at q = 12 (Table 5.6). This finding confirms the fact that area c12 is the key location in the study region, since the water in its approximate vicinity has the highest potential to be negatively impacted by various sources of pollution coming from the other upstream areas. In other words, σ_{12}(c12) is still the most connected simplex where q-value indicates (q + 1) number of water quality components (i.e., traffic) of potential hazard. Accordingly, simplices σ_7(c10) and σ_7(c15) also appear to be exposed to a negative environmental impact as they enter the structure after σ_{12}(c12) at q = 7. At levels q = 6, 5 and 4, one interesting observation can be made. It seems that component {c06} is more

connected than {c14}, which is opposite in the regular backcloth structure at the hierarchical level N (Table 5.6 cf. Table 5.2). This says that simplex σ_4(c14) faces less environmental threat (i.e., less traffic reaches it) compared to the σ_6(c06) after considering water quality data (i.e., traffic), even though σ_4(c14) is more connected in the backcloth. In a general sense, it also seems that at high q-values components are quite disconnected. This means that simplices do not share many vertices and most traces of pollution do not make it to the next destination area (i.e., hexagon). Such behaviour is clearly visible in geometric realization of the combined conjugate simplicial complex at level (N, M) (Figure 5.7a), discussed shortly. On the other hand, it is possible to observe three similar distinct connected components starting from q = 3 down to q = 0 (Table 5.6). What is interesting, however, is that areas c03 and c11 are completely excluded from the structure. This includes even the lowest q-level q = 0 where simplices can share as little as one vertex, which means that no pollution has been detected in these areas. While it might not be a direct interest of the application, the analysis provides empirical evidence that these areas have no observed pollution (i.e., no traffic variables have exceeded defined θ parameters).

Regarding the direct Q-analysis at the next hierarchical level (N+1, M+1) it seems that dimensionality of this structure has increased to q = 17, where component {c15} is identified (Table 5.6). This indicates that simplex σ_{17}(c15) can be polluted by the (q + 1) number of water parameters accumulated from different water sites. Due to changes in spatial resolution it is also possible to observe interesting changes in the traffic flow. In fact, it seems that both σ_{16}(c06) and σ_{13}(c09-11) simplices have become a new connectivity hub where much water traffic can be concentrated. In fact, both of them form a connected component at q = 13, which means that they have a potential of sharing 14 vertices. In addition, at q = 9 the analysis identifies two components {c06, c09-11, c15} and {c12-13}, which says something about movement of traffic through the backcloth. Specifically, it seems that much water pollution (i.e., traffic) have a strong tendency to move via connected chain of simplices (i.e., hexagons) in the {c06, c09-11, c15} component, with {c12-13} not being included in it and becoming more of an independent source. This is different from the analysis at (N, M) level, implications of which determined that area c12 is the key connectivity point of the structure (Table 5.6). The fact that simplex σ_9(c12-13) enters the structure significantly later indicates that its connectivity has changed, which is true since area c10 can no longer access area c12 directly as it was before at (N, M) level (Figure

A.1). Overall, changes in spatial aggregation of hexagon units have strong impact on the flow directionality of water traffic by reducing the complexity of network connectivity. The process also demonstrates that other connectivity factors can be added to the structure, such as component {c06, c09-11, c15} to reveal additional spatial behaviour and variability at the cost of the generalized information (e.g., aggregating areas c09, c10 and c11). Thus, analysis at various hierarchical levels can be quite advantageous in cases when it is necessary to account for certain level of accuracy and mitigate data gaps of the desired precision. Besides, it can be also beneficial when more general and global system connectivity patterns are analysed in conjunction with generalized statistical information not available at high precision, such as census data.

As for the conjugate Q-analysis for the hierarchical level (N, M) it is of particular importance, as it helps to tracks changes of water quality characteristics through different chains of connected water sites (i.e., simplices). Specifically, the analysis identifies water sites W12 as the source of Al, Ba, Sr, Tur, and sites W17b, W17s as the sources of Ba and Sr. These water sites are found at q = 2, which means they are capable of negatively affecting three areas including their places of origin (Table 5.7 cf. Figure 5.7a). At the same time, at q = 1 similar tendency is observed with each component being able to flow in at least one direction, including earlier mentioned W12, W17b and W17s sites. Note, each simplex within these connected components belong to single spatial area (i.e., hexagon). The interpretation of the output at q = 0 is the most important, as it depicts movement of the individual parameters within the complex. For example, it is clear that traces of Al and Tur detected by site W7 in c01, can be also detected by W8 in c02, suggesting that pollution has travelled from c01 into c02 (Figure 5.7a). In turn, it seems that Ba and Sr detected by W1 in c08 have not reached a directly connected area c07, since its water site W9 shows exceeding concentrations of Al and Tur instead. This might indicate that c07 has either a separate source of pollution or the actual water flow connectivity between these areas is not strong enough to carry these pollutants over. Lastly, the largest component {W4_Tur, W6_Al, W6_Tur, W11_Ba, W11_Sr, W11_Tur, W12_Al, W12_Ba, W12_Sr, W12_Tur, W13_Al, W13_Ba, W13_Sr, W13_Tur, W14_1_Ba, W14_1_Sr, W14b_Ba, W14b_Sr, W17b_Ba, W17b_Sr, W17s_Ba, W17s_Sr} might be the hardest to interpret. However, with the use of the study area map (Figure A.1) and geometric representation of the conjugate simplicial complex (Figure 5.7a) it is possible to identify the upstream location of individual sites and traversal of

water parameters through the network system. In particular, from the graph it seems that sites W12 (in c05) and W13 (in c09) are the two initial upstream locations where exceeding concentrations of Al, Ba, Sr and Tur are detected. From W12 the water flows to W11 (in c06) carrying over Ba, Sr and Tur, and then continues to W17b and W17s (in c12) where only Ba and Sr have been detected above the allowable limits. Similarly, from W13 the water flows to W14b and W14_1 (in c10) where only Ba and Sr are transferred, and then continues to W17b and W17s (in c12) where Ba and Sr are detected also. Given these observations, it is possible to conclude that Ba and Sr have a stronger impact on water quality and health in the study region since these elements are capable of traveling further within the system, compared to Al and Tur. This is also confirmed by the fact that concentrations of Al or Tur did not make it from W4 (in c13) to W3b and W3s (in c14), from W12 (in c05) to W10b and W10s (in c15), and W6 (in c04) to W17b and W17s (in c12) (Figure 5.7a cf. Figure A.1). The study does not exclude other factors which could have contributed to these outcomes, such as travel distance, water treatment procedures or even hydrological processed (e.g., sedimentation).Considering that these factors were not the focus of the application it is not possible to determine a degree of their importance. At the same time, note that water sites W2b and W2s (in c03) and W14_2 (in c11) are not found in the conjugate structure (Table 5.7 cf. Figure A.1). This has been also identified in the direct backcloth structure mentioned earlier, due to the fact that these water sites and their hexagon areas have met the standards for the water quality measurements. On the other hand, water sites such as W3b and W3s (in c14) and W10b and W10s (in c15) are also not included in the structure, even though their hexagon areas (i.e., c14 and c15) appear as high dimensional simplices in the direct backcloth structure (Table 5.7 cf. Table 5.6). It means that in reality traces of water pollution were not detected by these water sites, even though they are supposed to be highly susceptible to the negative impact via the water network.

Finally, the conjugate Q-analysis for the hierarchical level (N+1, M+1) provides additional view of the network system complexity where generalization is the product not only of the spatial aggregation of hexagon cells, but also contextually for the water quality parameters. The implications of traffic generalization might not be appropriate in all cases; nevertheless, they are of great significance when the focus is to identify all elements that belong to a specific group. For instance, microbiological parameters (e.g., turbidity) have the highest priority when it comes to the treatment guidelines of the water quality (Health Canada, 2019). Thus, it can be of great

interest to focus on the whole group rather than individual parameters. As a result, the connectivity analysis (Table 5.7) and the corresponding graphical representation (Figure 5.7b) identify water sites W11 (in c06); W4, W17b, W17s (in c12-13); and W13, W14b, W14_1, W14_2 (in c09-11) at q = 2 instead. This implies that these sites flow in two different directions, whereas new sites at q = 1 only in one direction. At q = 0 the study provides more insight into the observed water flow circularity between areas c06 and c09-11. It is possible to contemplate that such circular movement of water can have a direct impact on how the water parameters spread across the network. This can be a potential reason why no contaminations were found in areas c14 and c15 at either of the hierarchical levels by the conjugate Q-analysis, since water elements may be primarily deposited to the neighboring areas only and not spread further down the stream. Certainly, this statement must be also confirmed with a detailed hydrological analysis. While such interpretation is substantially different from the conjugate analysis at the (N, M) level, it also provides a different interpretation scenario which may be helpful in identifying sources of pollution and choosing appropriate mitigation strategies.

In summary, these interpretations communicate an important property of Q-analysis, such that direct analysis reveals a degree to which different areas are susceptible to the pollution, whereas conjugate analysis assists in monitoring actual changes in the water quality system and movement of individual elements. It is also clear that Q-analysis without traffic is merely an analysis of general system connectivity, whereas backcloth-traffic mapping leads to understanding of changes in water quality parameters (i.e., traffic) in the network system (i.e., backcloth). In other words, it is possible to observe the extent to which changes of traffic values in one area can promote changes in other areas via chain of connected simplices. In fact, this statement is also considered as one of the objectives for interpreting traffic on the backcloth (Johnson, 1990, p. 284).

Table 5.6. Q-analysis output of $K_C^N(W \times T^M, \rho^{(N, M)})$ and $K_C^{N+1}(W \times T^{M+1}, \rho^{(N+1, M+1)})$ direct simplicial complexes. The curly brackets signify a connected component and its simplices. The N/A value indicates absence of the corresponding q-value in the complex.

q-value	$K_C^N(W \times T^M, \rho^{(N, M)})$	$K_C^{N+1}(W \times T^{M+1}, \rho^{(N+1, M+1)})$
17	N/A	{c15}
16	N/A	{c06}, {c15}
15	N/A	{c06}, {c15}
14	N/A	{c06}, {c15}
13	N/A	{c06, c09-11}, {c15}
12	{c12}	{c06, c09-11}, {c15}
11	{c12}	{c06, c09-11, c15}
10	{c12}	{c06, c09-11, c15}
9	{c12}	{c06, c09-11, c15}, {c12-13}
8	{c12}	{c06, c09-11, c15}, {c12-13}
7	{c10}, {c12}, {c15}	{c06, c09-11, c15}, {c12-13}
6	{c06}, {c10}, {c12}, {c15}	{c06, c09-11, c15}, {c12-13}
5	{c06}, {c10}, {c12}, {c15}	{c06, c09-11, c12-13, c14, c15}
4	{c06}, {c10}, {c12}, {c14}, {c15}	{c06, c09-11, c12-13, c14, c15}
3	{c02}, {c07}, {c05, c06, c09, c10, c12, c14, c15}	{c02}, {c06, c09-11, c12-13, c14, c15}
2	{c02}, {c07}, {c05, c06, c09, c10, c12, c14, c15}	{c02}, {c07}, {c05, c06, c09-11, c12-13, c14, c15}
1	{c01, c02}, {c07, c08}, {c04, c05, c06, c09, c10, c12, c14, c15}	{c01, c02}, {c07}, {c04, c05, c06, c09-11, c12-13, c14, c15}
0	{c01, c02}, {c07, c08}, {c04, c05, c06, c09, c10, c12, c13, c14, c15}	{c01, c02}, {c07, c08}, {c04, c05, c06, c09-11, c12-13, c14, c15}

Table 5.7. Q-analysis output of $K_{WxT}^{M}(C^{N}, \rho^{-(N, M)})$ and $K_{WxT}^{M+1}(C^{N+1}, \rho^{-(N+1, M+1)})$ conjugate simplicial complexes. The curly brackets signify a connected component and its simplices.

q-value	$K_{WxT}^{M}(C^{N}, \rho^{-(N, M)})$	$K_{WxT}^{M+1}(C^{N+1}, \rho^{-(N+1, M+1)})$
2	{W12_Al, W12_Ba, W12_Sr, W12_Tur}, {W17b_Ba, W17b_Sr, W17s_Ba, W17s_Sr}	{W11_I, W11_M}, {W4_I, W4_M, W17b_I, W17b_M, W17s_I, W17s_M}, {W13_I, W13_M, W13_T, W14_1_I, W14_1_M, W14_1_T, W14_2_I, W14_2_M, W14_2_T, W14b_I, W14b_M, W14b_T}
1	{W6_Al, W6_Tur}, {W7_Al, W7_Tur}, {W12_Al, W12_Ba, W12_Sr, W12_Tur}, {W13_Al, W13_Ba, W13_Sr, W13_Tur}, {W4_Tur}, {W11_Ba, W11_Sr, W11_Tur}, {W1_Ba, W1_Sr}, {W14_1_Ba, W14_1_Sr, W14b_Ba, W14b_Sr}, {W17b_Ba, W17b_Sr, W17s_Ba, W17s_Sr}	{W6_M, W6_T}, {W7_M, W7_T}, {W12_I, W12_M, W12_T}, {W4_I, W4_M, W17b_I, W17b_M, W17s_I, W17s_M}, {W1_I}, {W11_I, W11_M, W13_I, W13_M, W13_T, W14_1_I, W14_1_M, W14_1_T, W14_2_I, W14_2_M, W14_2_T, W14b_I, W14b_M, W14b_T}
0	{W7_Al, W7_Tur, W8_Al, W8_Tur}, {W1_Ba, W1_Sr, W9_Al, W9_Tur}, {W4_Tur, W6_Al, W6_Tur, W11_Ba, W11_Sr, W11_Tur, W12_Al, W12_Ba, W12_Sr, W12_Tur, W13_Al, W13_Ba, W13_Sr, W13_Tur, W14_1_Ba, W14_1_Sr, W14b_Ba, W14b_Sr, W17b_Ba, W17b_Sr, W17s_Ba, W17s_Sr}	{W7_M, W7_T, W8_M, W8_T}, {W1_I, W9_M, W9_T}, {W4_I, W4_M, W6_M, W6_T, W11_I, W11_M, W12_I, W12_M, W12_T, W13_I, W13_M, W13_T, W14_1_I, W14_1_M, W14_1_T, W14_2_I, W14_2_M, W14_2_T, W14b_I, W14b_M, W14b_T, W17b_I, W17b_M, W17s_I, W17s_M}

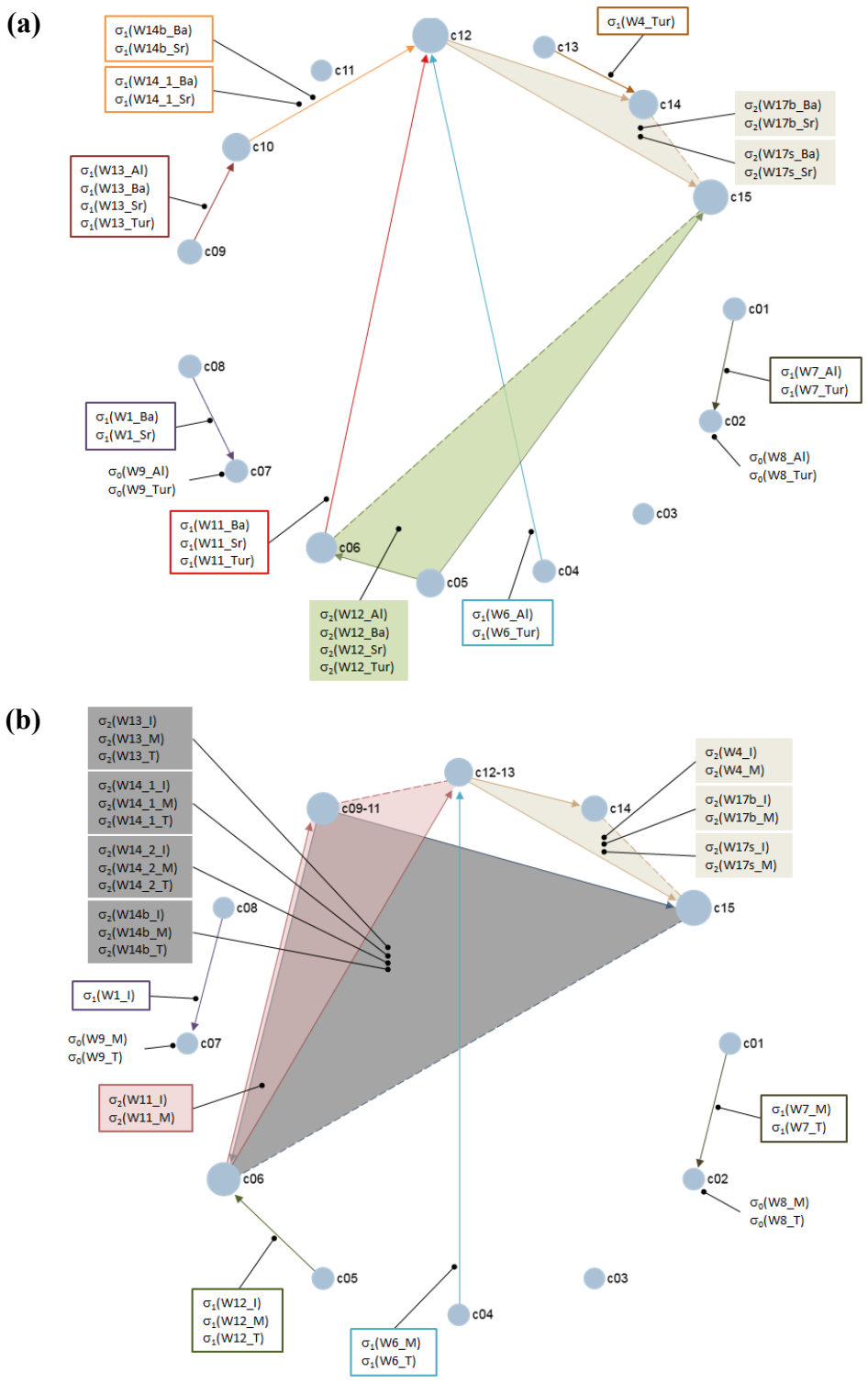


Figure 5.7. Geometric representation of (a) $K_{W\times T}{}^{M}(C^{N}, \rho^{-(N, M)})$ and (b) $K_{W\times T}{}^{M+1}(C^{N+1}, \rho^{-(N+1, M+1)})$ conjugate simplicial complexes.

Chapter 6 Conclusion

As the final part of this thesis, this chapter summarizes main contributions made by this research and suggests some necessary steps for the successful growth and development of this research. In particular, in Section 6.1 the study reports major contributions to the field, which includes assessment of DGGS implementations, enhancement of the DGGS analytical capability and a case study of the water quality and health. Section 6.2 concludes with the future perspectives and thoughts on how this research can be improved further.

6.1 Study Outcomes

The emergence of DGGS opened up new horizons and opportunities for comprehensive geospatial analysis that is suitable for integration with multiple data sources and data types, and is consistent in both use and representation. The framework was designed with the purpose of providing a uniform grid based system capable of discretizing spatial information globally and performing its analytical interpretation through a hierarchy of equal area cells. Accordingly, each corresponding cell references a specific area on the surface of the Earth, as well as can contain multiple data values and be uniquely identified within the hierarchy system.

A detailed assessment of available DGGS solutions and their evaluation for operational capability have shown to have great potential and versatility for the purposes of data handling and performing basic spatial operations, such as intersection analysis, neighbourhood search or data indexing, including possibilities for visualization. At the same time, the study recognizes that not all libraries are at the same level of development and much work is still required in order to reach a point when such systems are possible to use without the need to understand library-specific implementation details or technical characteristics. In other words, the study concludes that much technical preparation is necessary in order to perform a required operation, not to mention advanced GIS analyses such as terrain generation, pattern recognition or network modelling.

In addition to the operational basis, the study also developed a scalability analysis, designed specifically to benchmark the performance of selected DGGS applications under simulated

accumulation of data volumes. This assessment is of particular importance, as it provides a quantitative measure of library's ability to cope with large-scale datasets for the real-world application and production purposes. The results show that all reviewed DGGS implementations are capable of handling large data volumes in a timely manner. Overall, the dggridR library was found to be the most efficient at processing large-scale datasets, while recognizing the variation in performance due to different factors. One particularly interesting observation made for the dggridR was to discover that the use of triangles seemed to be outperformed by other shapes, whereas triangles are known to be generally faster. The same, however, cannot be said about the OpenEAGGR library, since its triangles do seem to show better performance.

For the concluding part of the DGGS assessment the study carries out a detailed comparative and compatibility review of the chosen DGGS implementations against current OGC abstract specification for DGGS (OGC, 2017). The review of the OGC abstract specification affirms that it is not possible to confirm successful fulfilment of all OGC criteria by any of the evaluated DGGS software. The H3 library appears to be the most complete lacking only criterion 7, which was also not possible to confirm for any of the other libraries considered here. The OpenEAGGR and S2, libraries are lacking criteria 3, 5, 7 and 7, 8, 9 respectively, as well as the dggridR library having four unfulfilled criteria 7, 15, 17 and 18. For the most part, the missing requirements are the consequences of lacking explicit documentation and continuing development of the current software. The analysis suggests making some improvements to the OGC abstract specification in order to clarify contextual ambiguity and provide more explicit guidelines for conformance.

While the detailed assessment and evaluation of DGGS software show great potential and advantages for GIS data management and analysis, the lack of advance analytical capability motivated a further search for more explicit data friendly techniques complimentary with the data rich environment and holistic DGGS framework. In this regard, the study explored the use of Q-analysis, a technique known for its ability to facilitate interpretation perspective of complex data systems using concepts of algebraic topology and a relational approach to data analysis (Atkin, 1974). Much attention was dedicated to the exploration of this particular methodology in order to take the full advantage of the emerging distributed hierarchical data systems (i.e., DGGS). Primarily, the goal was to meet the demands for formalizing geographical knowledge and understanding of data complexity in the science of complex systems (Miller & Goodchild,

2015), when other conventional methods fall short in capturing interaction between system components explicitly (Johnson, 2014, p. 8). As stated by Johnson, the term complexity is multifaceted and can address various issues, including but not limited to the properties of globality, network connectivity, multilevel and discrete dynamics (Johnson, 2014, p. 6). According to these properties and their definitions, the study has drawn parallels between complexity and core DGGS principles, which encouraged further exploration of DGGS data structures under umbrella of complex systems science. In other words, Q-analysis was found to be a suitable technique for exploring structural characteristics and complexity of data rich DGGS environment, due to its use of algebraic topology and simplicial complexes to accurately model and observe system behaviour and interactions under continuous changes.

As a matter of fact, the study contributes to the research in DGGS by extending the analytical capability of Q-analysis for the use within a hierarchical data system. Specifically, the developed methodology takes advantage of hierarchical cover sets and simplicial complexes to build a connectivity structure known as backcloth between well-defined sets of data, with one of them being a set composed of DGGS cells at some applicable hierarchical level. Furthermore, the formed backcloth can also carry associative information about the sets, and form additional connectivity structure called traffic. Hence, Q-analysis was made compatible with the complexity of DGGS data infrastructure via the multilevel backcloth-traffic systems. The research concludes that proposed methodology was found sufficient and suitable for exploring structural characteristics and insight of datasets at different hierarchical levels for both spatial and aspatial perspectives.

As the proof of concept, the study considered the water health and water quality application for the purposes of exploring practical implications and complexity measures of the proposed data analysis methodology. In the process, it has been shown that Q-analysis is useful in addressing physical connectivity and interaction of the water network system, including the spatial significance of individual water sites. It has been also recognized that changes between hierarchical levels (i.e., spatial resolution of DGGS cells) have a significant impact on the connectivity structure of the water network, which in turn introduces new system behaviour and insight (e.g., circular water flow connectivity). While the advantages are apparent, it is also suggested to perform and compare the outcomes at the lower and more detailed hierarchical

levels in order to validate the results and assist with their interpretation. In the end, the study concludes that careful interpretation of the connectivity chains between connected components and simplices reveal much detail regarding system behaviour, including degree of pollution spread and its corresponding chain affects.

Although the case study introduced here was limited primarily to the water health application and a specific dataset for the cumulative monitoring of the aquatic ecosystem (Government of the Northwest Territories, 2019), the developed technique is highly adaptable and likely to work with various data within the broad DGGS framework; as long as rules for deriving well-defined cover sets and contextual rationale are preserved. In fact, it is encouraged to integrate additional datasets in order to gain better understanding and systematic behaviour of the study subject. In summary, the implications of this research is to provide grounds for "letting the data speak for themselves" (Gould, 1981), and steer towards new and advance data driven solutions (Miller & Goodchild, 2015) in order to gain understanding of complexity (Johnson, 2014) in the data rich DGGS environment (OGC, 2017), as well as to promote further DGGS data infrastructure incentives and developments.

6.2 Future Work

Considering the outcomes provided by this research and the centrality of DGGS in the execution of this work, improvement to the OGC abstract specification for DGGS is an important and necessary place to start for the further advancement of this research. Being at its early stage of development, the OGC DGGS abstract specification must reach a point when it can be easily validated and enforced by end users.

Regarding the use of Q-analysis methodology the study also suggest further exploration of the additional quantitative measures available within the available theoretic developments, such as the concept of eccentricity. From the Atkin's point of view the measure of eccentricity shows how well an individual simplex is integrated within a simplicial complex (Atkin, 1974, pp. 33–35), which can be a useful property to explore for better understanding of the system components and their roles.

Another consideration for advancing the current methodology is the use of a time component more explicitly within DGGS framework and the corresponding space-time analysis. Such analysis was articulated through the concept of dynamics and graded patterns of traffic on backcloth (Atkin, 1974, pp. 126–137), as well as q-transmission mechanisms (Johnson, 1982).

Throughout this study DGGS with hexagon aperture 7 was mainly used for the application and modeling purposes. Nevertheless, as mentioned earlier within the study scope (see Section 3.3.1), possibilities for using other apertures, such as 3 or 4, or cell shapes are also supported by the Q-analysis methodology. Thus, it is suggested to explore these opportunities in more detail and determine if there are any other advantages or disadvantages for using one particular aperture or shape over the others.

Special consideration should also be given to mixed resolution structures (Figure 5.3), when DGGS cells do not have a fixed hierarchical level and their spatial resolution vary. Behaviour of these mixed structures within the Q-analysis framework is not completely understood, and they are of great interest and relevance to the future advancement of DGGS. Not to mention the fact that such mixed resolution structures are not provided by all DGGS implementations.

On the application side of things, some improvements to the methodology can be made by integrating additional data sources and contextual information related to the water health analysis. Specifically, these could be water catchment areas or ground water flow data, which should raise confidence and put on the presented results on firmer scientific ground. Additional information, such as population, urbanization factor or land cover data, can be also included in the analysis in order to gain insight and understanding into surface water pollution.

Lastly, this research has mostly focused on the regional scale analysis and a water health application, as a case study and a proof of concept for the integrated use of Q-analysis and DGGS. In the future, it is also suggested to perform an analysis on the global scale, to demonstrate its value for the truly global systems. In this sense, the future studies might consider exploring a more sophisticated global connectivity system where communication and data transmission between nodes or vertices is permitted in reverse directions too. This is different from the water heath application explored here, since water generally cannot flow in reverse (i.e., opposite direction). Thus, for example, a study could be conducted on the global air traffic

network system within the DGGS space framework, with all airport locations treated as the vertex set to form a backcloth. The spread of some sort of a disease or a virus (e.g., coronavirus) can be treated as traffic over this backcloth. In this regard, the spread of a disease can be monitored over the globe, which is a powerful and attractive capability. Given that other descriptive characteristics or information can be integrated in the process, the methodology can become of great value to determine the rationale behind the problem and assist in its mitigation.

References

Arza-García, M., Gil-Docampo, M., Ortiz-Sanz, J., & Martínez-Rodríguez, S. (2019). Virtual globes for UAV-based data integration: Sputnik GIS and Google Earth™ applications. *International Journal of Digital Earth*, *12*(5), 583–593. https://doi.org/10.1080/17538947.2018.1470205

Atkin, R. H. (1972). From cohomology in physics to q-connectivity in social science. *International Journal of Man-Machine Studies*, *4*(2), 139–167. https://doi.org/10.1016/S0020-7373(72)80029-4

Atkin, R. H. (1974). *Mathematical structure in human affairs*. Crane, Russak and Company, Inc.

Atkin, R. H. (1980). The Methodology of Q-Analysis: How to Study Corporations by Using Concepts of Connectivity. *Management Decision*, *18*(7), 380–390. https://doi.org/10.1108/eb001259

Atkin, R. H., & Casti, J. (1977). *Polyhedral Dynamics and the Geometry of Systems* [IIASA Research Report]. IIASA. http://pure.iiasa.ac.at/id/eprint/709/

Atkin, R. H., Johnson, J. H., & Mancini, V. (1971). An Analysis of Urban Structure Using Concepts of Algebraic Topology. *Urban Studies*, *8*(3), 221–242. https://doi.org/10.1080/00420987120080421

Bai, J., Zhao, X., & Chen, J. (2005). *Indexing of the discrete global grid using linear quadtree*. ISPRS Workshop on Service and Application of Spatial Data Infrastructure, Hangzhou, China, October 14–16.

Barnes, R. (2016). *DggridR: Discrete Global Grids for R*. https://github.com/r-barnes/dggridR

Barnes, R. (2019). Optimal orientations of discrete global grids and the Poles of Inaccessibility. *International Journal of Digital Earth*. https://doi.org/10.1080/17538947.2019.1576786

Barrett, C. L., Johnson, J. H., & Marathe, M. (2018). Big Data: High Performance Synthetic Information Environments: An integrating architecture in the age of pervasive data and computing. *Ubiquity*, *2018*, 1–11. https://doi.org/10.1145/3158342

Beaumont, J. R., & Gatrell, A. C. (1982). *An introduction to Q-analysis*. Geo Abstracts.

Ben, J., Li, Y., Zhou, C., Wang, R., & Du, L. (2018). Algebraic encoding scheme for aperture 3 hexagonal discrete global grid system. *Science China Earth Sciences*, *61*(2), 215–227. https://doi.org/10.1007/s11430-017-9111-y

Birch, C. P. D., Oom, S. P., & Beecham, J. A. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling*, *206*(3), 347–359. https://doi.org/10.1016/j.ecolmodel.2007.03.041

Bliemel, M. J., McCarthy, I. P., & Maine, E. M. A. (2014). An Integrated Approach to Studying Multiplexity in Entrepreneurial Networks. *Entrepreneurship Research Journal*, *4*(4). https://doi.org/10.1515/erj-2014-0007

Bondaruk, B., Roberts, S. A., & Robertson, C. (2019). Discrete Global Grid Systems: Operational Capability of the Current State of the Art. *Proceedings of the 7th Conference on Spatial Knowledge and Information Canada (SKI2019)*, *Volume 2323*. http://ceur-ws.org/Vol-2323/SKI-Canada-2019-7-6-1.pdf

Brodsky, I. (2018). *H3: Uber's Hexagonal Hierarchical Spatial Index*. Uber Engineering. https://eng.uber.com/h3/

Bush, I. (2017). *OpenEAGGR Software Design Document*. GitHub-Riskaware Ltd. https://github.com/riskaware-ltd/open-eaggr/blob/master/Documents/Software Design Document.pdf

Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S., & Schaub, T. (2016). *The GeoJSON Format*. Internet Engineering Task Force (IETF). https://tools.ietf.org/html/rfc7946

Casti, J. (1975a). *Polyhedral Dynamics—I: The relevance of Algebraic Topology to Human Affairs* [IIASA Working Paper]. IIASA. http://pure.iiasa.ac.at/id/eprint/398/

Casti, J. (1975b). *Polyhedral Dynamics—II: Geometrical Structure as a Basis for Decision Making in Complex Systems* [IIASA Research Memorandum]. IIASA. http://pure.iiasa.ac.at/id/eprint/482/

Casti, J., Kempf, J., Duckstein, L., & Fogel, M. (1979). Lake ecosystems: A polyhedral dynamics representation. *Ecological Modelling*, *7*(3), 223–237. https://doi.org/10.1016/0304-3800(79)90071-1

Comber, A., & Wulder, M. (2019). Considering spatiotemporal processes in big data analysis: Insights from remote sensing of land cover and land use. *Transactions in GIS*, *23*(5), 879–891. https://doi.org/10.1111/tgis.12559

Dowker, C. H. (1952). Homology Groups of Relations. *The Annals of Mathematics*, *56*(1), 84. https://doi.org/10.2307/1969768

Dutton, G. (1984). Part 4: Mathematical, Algorithmic and Data Structure Issues: Geodesic Modelling Of Planetary Relief. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *21*, 188–207. https://doi.org/10.3138/R613-191U-7255-082N

Dutton, G. (1999). Historical and conceptual background. In *A Hierarchical Coordinate System for Geoprocessing and Cartography* (Vol. 79, pp. 19–39). Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0011619

Esri Canada. (2019). *ArcGIS Platform*. https://esri.ca/en/products/arcgis-platform

Gaspar, J., & Gould, P. (1981). The Cova da Beira: An Applied Structural Analysis of Agriculture and Communication. In A. Pred (Ed.), *Space and Time in Geography* (pp. 183–214). CWK Gleerup Lund.

Gatrell, A. C. (1981). On the Structure of Urban Social Areas: Explorations Using Q-Analysis. *Transactions of the Institute of British Geographers*, *6*(2), 228–245. JSTOR. https://doi.org/10.2307/622137

Gibb, R. G. (2016). The rHEALPix Discrete Global Grid System. *IOP Conference Series: Earth and Environmental Science*, *34*, 012012. https://doi.org/10.1088/1755-1315/34/1/012012

Górski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., & Bartelman, M. (2005). HEALPix-Framework for High Resolution Discretization, and Fast Analysis of Data Distributed on the Sphere. *The Astrophysical Journal*, *622*(2), 759–771. https://doi.org/10.1086/427976

Górski, K. M., Wandelt, B. D., Hivon, E., Hansen, F. K., & Banday, A. J. (2018). *Introduction to HEALPix*. HEALPix-Data Analysis, Simulations and Visualization on the Sphere. https://healpix.sourceforge.io/html/intro.htm

Gould, P. (1980). Q-analysis, or a language of structure: An introduction for social scientists, geographers and planners. *International Journal of Man-Machine Studies*, *13*(2), 169–199. https://doi.org/10.1016/S0020-7373(80)80009-5

Gould, P. (1981). Letting the Data Speak for Themselves. *Annals of the Association of American Geographers*, *71*(2), 166–176. https://doi.org/10.1111/j.1467-8306.1981.tb01346.x

Gould, P., & Gatrell, A. (1979). A structural analysis of a game: The Liverpool v Manchester united cup final of 1977. *Social Networks*, *2*(3), 253–273. https://doi.org/10.1016/0378-8733(79)90017-0

Gould, P., & Johnson, J. H. (1980). National television policy: Monitoring structural complexity. *Futures*, *12*(3), 178–190. https://doi.org/10.1016/0016-3287(80)90021-X

Government of the Northwest Territories. (2019). *CIMP 161_Cumulative Impacts Monitoring of Aquatic Ecosystem Health of Yellowknife Bay, Great Slave Lake, NWT*. Mackenzie DataStream. https://doi.org/10.25976/48P4-ST77

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography*, *3*(3), 255–261. https://doi.org/10.1177/2043820613513121

Health Canada. (2019). *Guidelines for Canadian Drinking Water Quality—Summary Table*. Government of Canada. https://www.canada.ca/en/health-canada/services/environmental-workplace-health/reports-publications/water-quality/guidelines-canadian-drinking-water-quality-summary-table.html

Helmi, A. M., Farhan, M. S., & Nasr, M. M. (2018). A framework for integrating geospatial information systems and hybrid cloud computing. *Computers & Electrical Engineering*, *67*, 145–158. https://doi.org/10.1016/j.compeleceng.2018.03.027

Jacomy, A., Plique, G., & Guido, D. (2017). *Sigma.js*. http://sigmajs.org/

Jiang, B., & Omer, I. (2007). Spatial Topology and its Structural Analysis based on the Concept of Simplicial Complex. *Transactions in GIS*, *11*(6), 943–960. https://doi.org/10.1111/j.1467-9671.2007.01073.x

Johnson, J. H. (1978). A Q-analysis of television programmes. *International Journal of Man-Machine Studies*, *10*(4), 461–479. https://doi.org/10.1016/S0020-7373(78)80004-2

Johnson, J. H. (1981a). Some structures and notation of Q-analysis. *Environment and Planning B: Planning and Design*, *8*(1), 73–86. https://doi.org/10.1068/b080073

Johnson, J. H. (1981b). The Q-Analysis of Road Traffic Systems. *Environment and Planning B: Planning and Design*, *8*(2), 141–189. https://doi.org/10.1068/b080141

Johnson, J. H. (1982). Q-Transmission in simplicial complexes. *International Journal of Man-Machine Studies*, *16*(4), 351–377. https://doi.org/10.1016/S0020-7373(82)80046-1

Johnson, J. H. (1983a). Hierarchial set definition by Q-analysis, part I. The hierarchical backcloth. *International Journal of Man-Machine Studies*, *18*(4), 337–359. https://doi.org/10.1016/S0020-7373(83)80014-5

Johnson, J. H. (1983b). Hierarchical set definition by Q-analysis, part II. Traffic on the hierarchical backcloth. *International Journal of Man-Machine Studies*, *18*(5), 467–487. https://doi.org/10.1016/S0020-7373(83)80021-2

Johnson, J. H. (1986). Hierarchical Backcloth—Traffic Simulation. *Environment and Planning B: Planning and Design*, *13*(4), 415–436. https://doi.org/10.1068/b130415

Johnson, J. H. (1990). Interpretation and Hierarchical Set Definition in Q-Analysis. *Environment and Planning B: Planning and Design*. https://doi.org/10.1068/b170277

Johnson, J. H. (2014). *Hypernetworks in the science of complex systems* (Vol. 3). Imperial College Press.

Johnson, J. H., Denning, P., Delic, K. A., & Sousa-Rodrigues, D. (2018). Big Data: Big Data or Big Brother? That is the question now. *Ubiquity*, *2018*, 1–10. https://doi.org/10.1145/3158352

Johnson, J. H., & Wanmali, S. (1981). A Q-Analysis of Periodic Market Systems. *Geographical Analysis*, *13*(3), 262–275. https://doi.org/10.1111/j.1538-4632.1981.tb00734.x

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 1–10. https://doi.org/10.1177/2053951716631130

Kreiss, S. (2016). *S2 cells and space-filling curves: Keys to building better digital map tools for cities*. Medium. https://medium.com/sidewalk-talk/s2-cells-and-space-filling-curves-keys-to-building-better-digital-map-tools-for-cities-a312aa5e2f59

Leslie, T. W. (2001). Energetic geometries: The Dymaxion Map and the skin/structure fusion of Buckminster Fuller's geodesics. *Arq: Architectural Research Quarterly*, *5*(2), 161–170. https://doi.org/10.1017/S135913550100118X

Li, W., Song, M., Zhou, B., Cao, K., & Gao, S. (2015). Performance improvement techniques for geospatial web services in a cyberinfrastructure environment – A case study with a disaster management portal. *Computers, Environment and Urban Systems*, *54*, 314–325. https://doi.org/10.1016/j.compenvurbsys.2015.04.003

Li, X. (2013). Storage and addressing scheme for practical hexagonal image processing. *Journal of Electronic Imaging*, *22*(1), 010502. https://doi.org/10.1117/1.JEI.22.1.010502

Mahdavi-Amiri, A., Alderson, T., & Samavati, F. (2015). A Survey of Digital Earth. *Computers & Graphics*, *53*, 95–117. https://doi.org/10.1016/j.cag.2015.08.005

Mahdavi-Amiri, A., Alderson, T., & Samavati, F. (2016). *Data Management Possibilities for Aperture 3 Hexagonal Discrete Global Grid Systems* (Science Research & Publications) [Technical report]. University of Calgary. https://prism.ucalgary.ca/handle/1880/51524

Mahdavi-Amiri, A., Alderson, T., & Samavati, F. (2019). Geospatial Data Organization Methods with Emphasis on Aperture-3 Hexagonal Discrete Global Grid Systems. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *54*(1), 30–50. https://doi.org/10.3138/cart.54.1.2018-0010

Mahdavi-Amiri, A., Samavati, F., & Peterson, P. (2015). Categorization and Conversions for Indexing Methods of Discrete Global Grid Systems. *ISPRS International Journal of Geo-Information*, *4*(1), 320–336. https://doi.org/10.3390/ijgi4010320

Maletić, S., & Zhao, Y. (2017). Multilevel Integration Entropies: The Case of Reconstruction of Structural Quasi-Stability in Building Complex Datasets. *Entropy*, *19*(4), 172. https://doi.org/10.3390/e19040172

Middleton, L., & Sivaswamy, J. (2005). *Hexagonal image processing: A practical approach*. Springer.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, *80*(4), 449–461. https://doi.org/10.1007/s10708-014-9602-6

Nat, A. (2018). *Polyhedra Viewer*. Polyhedra Viewer. https://polyhedra.tessera.li

Natural Resources Canada. (2015). *Canadian Digital Elevation Model, 1945-2011, cdem_dem_085J_tif*. Government of Canada. https://open.canada.ca/data/en/dataset/7f245e4d-76c2-4caa-951a-45d1d2051333

Natural Resources Canada. (2019). *Topographic Data of Canada—CanVec Series, canvec_50K_NT_Hydro_shp*. Government of Canada. https://open.canada.ca/data/en/dataset/8ba2aa2a-7bb9-4448-b4d7-f164409fe056

Node.js Foundation. (2019). *Node.js*. https://nodejs.org/en/

OGC. (2017). *Topic 21: Discrete Global Grid Systems Abstract Specification*. http://docs.opengeospatial.org/as/15-104r5/15-104r5.html

OGC. (2019). *OGC DGGS Standards Working Group seeks public comment on new tasks of work added to its charter*. Opengeospatial.Org. https://www.opengeospatial.org/standards/requests/194

Olson, J. M. (2006). Map Projections and the Visual Detective: How to Tell if a Map is Equal-Area, Conformal, or Neither. *Journal of Geography*, *105*(1), 13–32. https://doi.org/10.1080/00221340608978655

Omer, I., & Goldblatt, R. (2017). Using space syntax and Q-analysis for investigating movement patterns in buildings: The case of shopping malls. *Environment and Planning B: Urban Analytics and City Science*, *44*(3), 504–530. https://doi.org/10.1177/0265813516647061

Perone, C. S. (2015). *Google's S2, geometry on the sphere, cells and Hilbert curve*. http://blog.christianperone.com/2015/08/googles-s2-geometry-on-the-sphere-cells-and-hilbert-curve/

Peterson, P. R. (2017). Discrete Global Grid Systems. In *International Encyclopedia of Geography: People, the Earth, Environment and Technology* (pp. 1–10). American Cancer Society. https://doi.org/10.1002/9781118786352.wbieg1050

Purss, M. B. J., Gibb, R., Samavati, F., Peterson, P., & Ben, J. (2016). The OGC® Discrete Global Grid System core standard: A framework for rapid geospatial integration. *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 3610–3613. https://doi.org/10.1109/IGARSS.2016.7729935

PYXIS innovation inc. (2017). *DIGITAL EARTH Insight, Decision-Support on a Virtual Globe*. http://www.pyxisinnovation.com/Products/Insight

QGIS project. (2019). *QGIS. A Free and Open Source Geographic Information System*. https://www.qgis.org/en/site/

Riskaware Ltd. (2019). *Release of OpenEAGGR*. Riskaware. https://www.riskaware.co.uk/release-of-openeaggr/

Roberts, S. A., Hall, G. B., & Calamai, P. H. (2001). Assessing polygon edge integrity. *Journal of Geographical Systems*, *3*(1), 87–105. https://doi.org/10.1007/PL00011469

S2Geometry. (n.d.). *S2 Geometry*. S2Geometry. Retrieved July 29, 2019, from http://s2geometry.io/

Sahr, K. (2008). Location coding on icosahedral aperture 3 hexagon discrete global grids. *Computers, Environment and Urban Systems*, *32*(3), 174–187. https://doi.org/10.1016/j.compenvurbsys.2007.11.005

Sahr, K. (2013). On the Optimal Representation of Vector Location using Fixed-Width Multi-Precision Quantizers. *ISPRS - International Archives of the Photogrammetry, Remote Sensing*

and Spatial Information Sciences, *XL-4/W2*, 1–8. https://doi.org/10.5194/isprsarchives-XL-4-W2-1-2013

Sahr, K. (2018). *DGGRID version 6.4: User documentation for discrete global grid software.* STCL-Southern Terra Cognita Laboratory. https://discreteglobalgrids.org/wp-content/uploads/2019/05/dggridManualV64.pdf

Sahr, K. (2019). Central Place Indexing: Hierarchical Linear Indexing Systems for Mixed-Aperture Hexagonal Discrete Global Grid Systems. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *54*(1), 16–29. https://doi.org/10.3138/cart.54.1.2018-0022

Sahr, K., White, D., & Kimerling, A. J. (2003). Geodesic Discrete Global Grid Systems. *Cartography and Geographic Information Science*, *30*(2), 121–134. https://doi.org/10.1559/152304003100011090

Snyder, J. P. (2006). An Equal-Area Map Projection For Polyhedral Globes. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *29*(1), 10–21. https://doi.org/10.3138/27H7-8K88-4882-1752

Song, L., Kimerling, A. J., & Sahr, K. (2002). Developing an Equal Area Global Grid by Small Circle Subdivision. In M. F. Goodchild & A. J. Kimerling (Eds.), *Discrete Global Grids: A Web Book*. https://escholarship.org/uc/item/9492q6sm#main

Sonis, M., & Hewings, G. (2000). *Introduction to Input-output Structural Q-Analysis*.

Sowkhya, B., Amaduzzi, S., & Raawal, D. (2018). Visualization and Analysis of Cellular & Twitter Data Using QGIS. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII-4-W8*, 199–209.

Spooner, R., & Batty, M. (1981). Networks of Urban Systems Analysts. *Environment and Planning B: Planning and Design*, *8*(4), 449–475. https://doi.org/10.1068/b080449

Titlow, J. P. (2013). *How Foursquare Is Building A "Humane" Map Framework To Rival Google's*. Fast Company. https://www.fastcompany.com/3007394/how-foursquare-building-humane-map-framework-rival-googles

Tong, X., Ben, J., Liu, Y., & Zhang, Y. (2013). Modeling and Expression of Vector Data in the Hexagonal Discrete Global Grid System. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XL-4/W2*, 15–25. https://doi.org/10.5194/isprsarchives-XL-4-W2-15-2013

Tong, X., Ben, J., Wang, Y., Zhang, Y., & Pei, T. (2013). Efficient encoding and spatial operation scheme for aperture 4 hexagonal discrete global grid system. *International Journal of Geographical Information Science*, *27*(5), 898–921. https://doi.org/10.1080/13658816.2012.725474

Uber Technologies Inc. (2019). *H3: A hexagonal hierarchical geospatial indexing system*. https://uber.github.io/h3/#/

Veach, E. (2017). Announcing the S2 Library: Geometry on the Sphere. *Google Open Source Blog*. https://opensource.googleblog.com/2017/12/announcing-s2-library-geometry-on-sphere.html

Wells, C. J. (2012). *The Platonic Solids*. TechnologyUK.Net. http://www.technologyuk.net/mathematics/geometry/platonic-solids.shtml

White, D. (2000). Global Grids from Recursive Diamond Subdivisions of the Surface of an Octahedron or Icosahedron. *Environmental Monitoring and Assessment*, *64*(1), 93–103. https://doi.org/10.1023/A:1006407023786

White, D., Kimerling, A. J., & Overton, S. W. (1992). Cartographic and Geometric Components of a Global Sampling Design for Environmental Monitoring. *Cartography and Geographic Information Systems*, *19*(1), 5–22. https://doi.org/10.1559/152304092783786636

White, D., Kimerling, A. J., Sahr, K., & Song, L. (1998). Comparing Area and Shape Distortion on Polyhedral-Based Recursive Partitions of the Sphere. *International Journal of Geographical Information Science*, *12*, 805–827. https://doi.org/10.1080/136588198241518

Zeroghan. (2019). A Comprehensive Guide to S2 Cells and Pokémon GO. *Pokemon GO Hub*. https://pokemongohub.net/post/article/comprehensive-guide-s2-cells-pokemon-go/

Appendix A Study Area Map

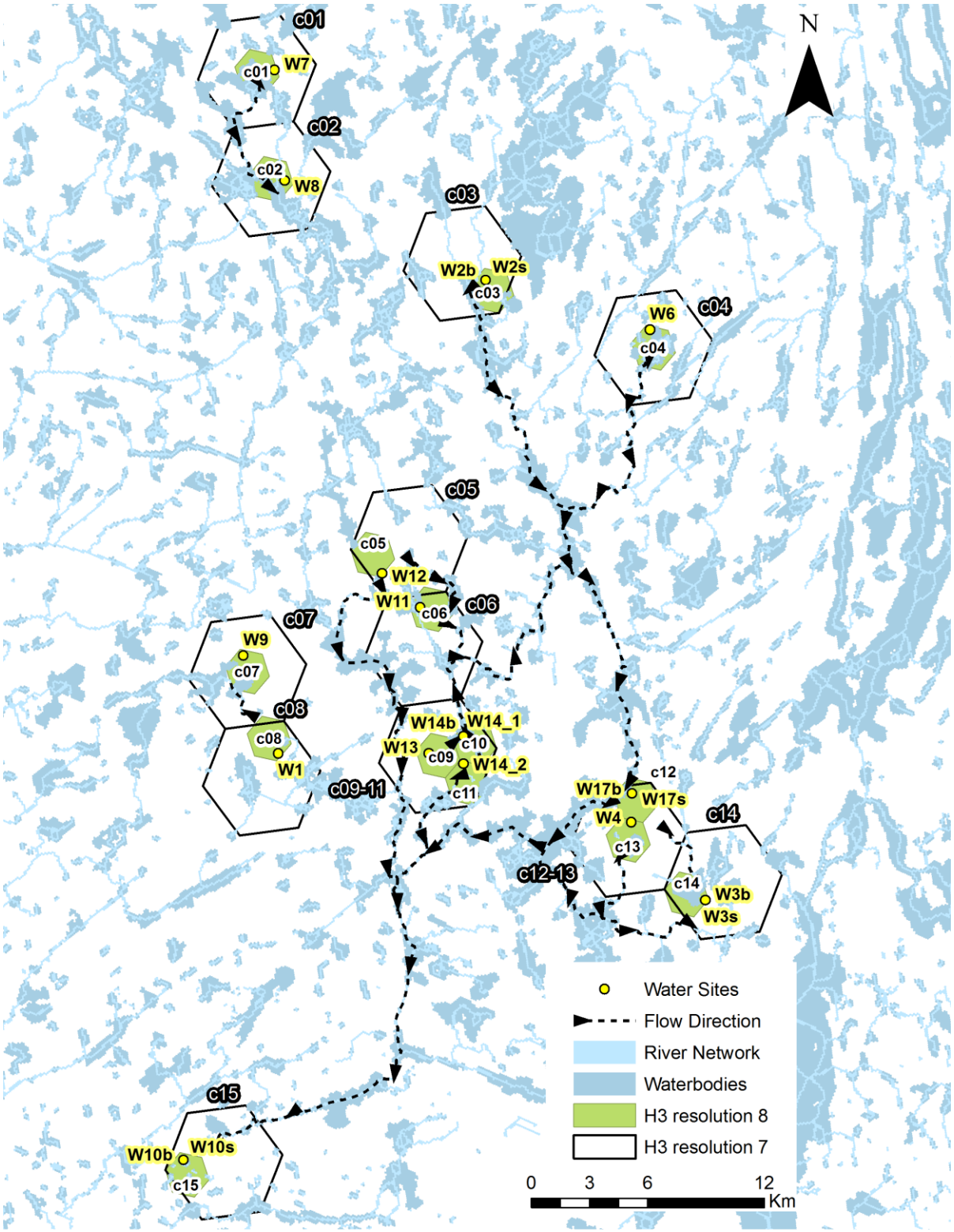


Figure A.1. The study area map of the Yellowknife Bay, Great Slave Lake, Canada, NWT. Data sources: (Government of the Northwest Territories, 2019; Natural Resources Canada, 2019).

Appendix B H3 Resolutions

Table B.1. Complete list of resolutions provided by the H3 library including area, edge length and index count properties. Data source: (Uber Technologies Inc., 2019).

H3 Resolution	Average Hexagon Area (km^2)	Average Hexagon Edge Length (km)	Number of unique indexes
0	4,250,546.8477000	1,107.712591000	122
1	607,220.9782429	418.676005500	842
2	86,745.8540347	158.244655800	5,882
3	12,392.2648621	59.810857940	41,162
4	1,770.3235517	22.606379400	288,122
5	252.9033645	8.544408276	2,016,842
6	36.1290521	3.229482772	14,117,882
7	5.1612932	1.220629759	98,825,162
8	0.7373276	0.461354684	691,776,122
9	0.1053325	0.174375668	4,842,432,842
10	0.0150475	0.065907807	33,897,029,882
11	0.0021496	0.024910561	237,279,209,162
12	0.0003071	0.009415526	1,660,954,464,122
13	0.0000439	0.003559893	11,626,681,248,842
14	0.0000063	0.001348575	81,386,768,741,882
15	0.0000009	0.000509713	569,707,381,193,162

Appendix C Hierarchical Backcloth Connectivity

Table C.1. A binary matrix of the direct λ^N relation, where 1 indicates connectivity between a cell area (row) and a water site (column).

λ^N	W1	W2b	W2s	W3b	W3s	W4	W6	W7	W8	W9	W10b	W10s	W11	W12	W13	W14_1	W14b	W14_2	W17b	W17s
c01								1												
c02								1	1											
c03		1	1																	
c04							1													
c05														1						
c06													1	1						
c07	1									1										
c08	1																			
c09															1					
c10															1	1	1	1		
c11																		1		
c12		1	1				1						1			1	1		1	1
c13						1														
c14				1	1	1													1	1
c15											1	1		1				1	1	1

Table C.2. A binary matrix of the direct λ^{N+1} relation, where 1 indicates connectivity between a cell area (row) and a water site (column).

λ^{N+1}	W1	W2b	W2s	W3b	W3s	W4	W6	W7	W8	W9	W10b	W10s	W11	W12	W13	W14_1	W14b	W14_2	W17b	W17s
c01								1												
c02								1	1											
c03		1	1																	
c04							1													
c05														1						
c06													1	1	1	1	1	1		
c07	1									1										
c08	1																			
c09-11													1		1	1	1	1		
c12-13		1	1			1	1						1						1	1
c14				1	1	1													1	1
c15						1					1	1			1	1	1	1	1	1

Table C.3. A binary matrix of the conjugate λ^{-N} relation, where 1 indicates connectivity between a water site (row) and a cell area (column).

λ^{-N}	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13	c14	c15
W1							1	1							
W2b			1									1			
W2s			1									1			
W3b														1	
W3s														1	
W4													1	1	
W6				1								1			
W7	1	1													
W8		1													
W9							1								
W10b															1
W10s															1
W11						1						1			
W12					1	1									1
W13									1	1					
W14_1										1		1			
W14b										1		1			
W14_2										1	1				1
W17b												1		1	1
W17s												1		1	1

Table C.4. A binary matrix of the conjugate $\lambda^{-(N+1)}$ relation, where 1 indicates connectivity between a water site (row) and a cell area (column).

$\lambda^{-(N+1)}$	c01	c02	c03	c04	c05	c06	c07	c08	c09-11	c12-13	c14	c15
W1							1	1				
W2b			1							1		
W2s			1							1		
W3b											1	
W3s											1	
W4										1	1	1
W6				1						1		
W7	1	1										
W8		1										
W9							1					
W10b												1
W10s												1
W11						1			1	1		
W12					1	1						
W13						1			1			1
W14_1						1			1			1
W14b						1			1			1
W14_2						1			1			1
W17b										1	1	1
W17s										1	1	1

Appendix D Hierarchical Traffic Connectivity

Table D.1. A binary matrix of the direct $\mu^{(N, M)}$ relation, where 1 indicates exceeding of the threshold value in the corresponding cell areas.

$\mu^{(N,M)}$	Al	Tur	Sr	Ba
c01	1	1		
c02	1	1		
c03				
c04	1	1		
c05	1	1	1	1
c06		1	1	1
c07	1	1		
c08			1	1
c09	1	1	1	1
c10			1	1
c11				
c12			1	1
c13		1		
c14				
c15				

Table D.2. A binary matrix of the direct $\mu^{(N+1, M+1)}$ relation, where 1 indicates exceeding of the threshold value in the corresponding cell areas.

$\mu^{(N+1,M+1)}$	Treatment	Microbiological	Inorganic
c01	1	1	
c02	1	1	
c03			
c04	1	1	
c05	1	1	1
c06		1	1
c07	1	1	
c08			1
c09-11	1	1	1
c12-13		1	1
c14			
c15			

Table D.3. A binary matrix of the conjugate $\mu^{-(N, M)}$ relation, where 1 indicates exceeding of the threshold value in the corresponding cell areas.

$\mu^{-(N,M)}$	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13	c14	c15
Al	1	1		1	1		1		1						
Tur	1	1		1	1	1	1		1				1		
Sr					1	1		1	1	1		1			
Ba					1	1		1	1	1		1			

Table D.4. A binary matrix of the conjugate $\mu^{-(N+1, M+1)}$ relation, where 1 indicates exceeding of the threshold value in the corresponding cell areas.

$\mu^{-(N+1,M+1)}$	c01	c02	c03	c04	c05	c06	c07	c08	c09-11	c12-13	c14	c15
Treatment	1	1		1	1		1		1			
Microbiological	1	1		1	1	1	1		1	1		
Inorganic					1	1		1	1	1		

Appendix E Backcloth-Traffic Mapping

Table E.1. A binary matrix of the direct $\rho^{(N, M)}$ relation, where 1 indicates exceeding concentrations of the water quality parameters in the corresponding water sites (columns) and cell areas (rows).

$\rho^{(N,M)}$	W1 x Ba	W1 x Sr	W4 x Tur	W6 x Al	W6 x Tur	W7 x Al	W7 x Tur	W8 x Al	W8 x Tur	W9 x Al	W9 x Tur	W11 x Ba	W11 x Sr	W11 x Tur	W12 x Al	W12 x Ba	…
c01						1	1										
c02						1	1	1	1								
c03																	
c04				1	1												
c05															1	1	
c06												1	1	1	1	1	
c07	1	1								1	1						
c08	1	1															
c09																	
c10																	
c11																	
c12				1	1							1	1	1			
c13			1														
c14			1														
c15															1	1	

Table E.1. Continued.

$\rho^{(N,M)}$	W12 x Sr	W12 x Tur	W13 x Al	W13 x Ba	W13 x Sr	W13 x Tur	W14_1 x Ba	W14_1 x Sr	W14b x Ba	W14b x Sr	W17b x Ba	W17b x Sr	W17s x Ba	W17s x Sr
c01														
c02														
c03														
c04														
c05	1	1												
c06	1	1												
c07														
c08														
c09			1	1	1	1								
c10			1	1	1	1	1	1	1	1				
c11														
c12							1	1	1	1	1	1	1	1
c13														
c14											1	1	1	1
c15	1	1									1	1	1	1

Table E.2. A binary matrix of the direct $\rho^{(N+1, M+1)}$ relation, where 1 indicates exceeding concentrations of the water quality parameters in the corresponding water sites (columns) and cell areas (rows).

$\rho^{(N+1,M+1)}$	W1 x I	W4 x I	W4 x M	W6 x M	W6 x T	W7 x M	W7 x T	W8 x M	W8 x T	W9 x M	W9 x T	W11 x I	W11 x M	W12 x I	W12 x M	W12 x T	…
c01						1	1										
c02						1	1	1	1								
c03																	
c04				1	1												
c05														1	1	1	
c06												1	1	1	1	1	
c07	1									1	1						
c08	1																
c09-11												1	1				
c12-13		1	1	1	1							1	1				
c14		1	1														
c15		1	1														

Table E.2. Continued.

$\rho^{(N+1,M+1)}$	W13 x I	W13 x M	W13 x T	W14_1 x I	W14_1 x M	W14_1 x T	w14_2 x I	w14_2 x M	w14_2 x T	W14b x I	W14b x M	W14b x T	W17b x I	W17b x M	W17s x I	W17s x M
c01																
c02																
c03																
c04																
c05																
c06	1	1	1	1	1	1	1	1	1	1	1	1				
c07																
c08																
c09-11	1	1	1	1	1	1	1	1	1	1	1	1				
c12-13													1	1	1	1
c14													1	1	1	1
c15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table E.3. A binary matrix of the conjugate $\rho^{-(N, M)}$ relation, where 1 indicates exceeding concentrations of the water quality parameters in the corresponding water sites (rows) and cell areas (columns).

$\rho^{-(N,M)}$	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13	c14	c15
W1 x Ba							1	1							
W1 x Sr							1	1							
W4 x Tur													1	1	
W6 x Al				1								1			
W6 x Tur				1								1			
W7 x Al	1	1													
W7 x Tur	1	1													
W8 x Al		1													
W8 x Tur		1													
W9 x Al							1								
W9 x Tur							1								
W11 x Ba						1						1			
W11 x Sr						1						1			
W11 x Tur						1						1			
W12 x Al					1	1									1
W12 x Ba					1	1									1
W12 x Sr					1	1									1
W12 x Tur					1	1									1
W13 x Al									1	1					
W13 x Ba									1	1					
W13 x Sr									1	1					
W13 x Tur									1	1					
W14_1 x Ba										1		1			
W14_1 x Sr										1		1			
W14b x Ba										1		1			
W14b x Sr										1		1			
W17b x Ba												1		1	1
W17b x Sr												1		1	1
W17s x Ba												1		1	1
W17s x Sr												1		1	1

Table E.4. A binary matrix of the conjugate $\rho^{-(N+1, M+1)}$ relation, where 1 indicates exceeding concentrations of the water quality parameters in the corresponding water sites (rows) and cell areas (columns).

$\rho^{-(N+1,M+1)}$	c01	c02	c03	c04	c05	c06	c07	c08	c09-11	c12-13	c14	c15
W1 x I							1	1				
W4 x I										1	1	1
W4 x M										1	1	1
W6 x M				1						1		
W6 x T				1						1		
W7 x M	1	1										
W7 x T	1	1										
W8 x M		1										
W8 x T		1										
W9 x M							1					
W9 x T							1					
W11 x I						1			1	1		
W11 x M						1			1	1		
W12 x I					1	1						
W12 x M					1	1						
W12 x T					1	1						
W13 x I						1			1			1
W13 x M						1			1			1
W13 x T						1			1			1
W14_1 x I						1			1			1
W14_1 x M						1			1			1
W14_1 x T						1			1			1
w14_2 x I						1			1			1
w14_2 x M						1			1			1
w14_2 x T						1			1			1
W14b x I						1			1			1
W14b x M						1			1			1
W14b x T						1			1			1
W17b x I										1	1	1
W17b x M										1	1	1
W17s x I										1	1	1
W17s x M										1	1	1

Glossary

Algebraic topology	A branch of mathematics which uses relational algebra to study geometric structures and topological spaces.
Aperture	Refers to a tessellation process to indicate a cell ratio between different resolution levels within DGGS.
Backcloth structure	A geometric structure defined by at least two sets used to model the multidimensional relation between set elements, and to represent the structure where certain activity can take place.
Backcloth-traffic system	A multilevel structure which includes the mapping process of the traffic structure onto the backcloth structure.
Complexity	A concept under the complex system science to facilitate the study of systems and their components, in order to provide methods and frameworks for effective analysis and formalization of information.
Conjugate relation	A transposed direct relation of λ^{-1}. The conjugate relation is represented as a transposed incidence matrix, and is used to form a conjugate simplicial complex.
Connected component	Array of connected simplices in a consecutive chain that can share common properties or characteristics.
Cover set	A collection of elements such that each element can be a member of multiple sets at the next more generalized hierarchical level, or be covered by it.
Data driven analysis	A type of analytical process which uses collection of various information sources as the primary basis for understanding and decision making.
Direct relation	Implies the relationship between two sets via some defined condition λ to determine whether elements are related to each other. The direct relation is represented as an incidence matrix, and is used to form a direct simplicial complex.

Discrete Global Grid Systems (DGGS)	A hierarchical system of regular polygons used as a model for data integration, analysis and visualization.
Fragmentation	A term describing number of connected components in a simplicial complex. Large number of individual components implies high fragmentation and vice versa.
Geometric realization	Graphic representation or visualization of a simplicial complex via a polyhedron solid.
Incidence matrix	Binary matrix representation of a relation between sets' elements where row elements represent simplices and column elements represent vertices of simplices. The value of 1 indicates that elements are related and 0 indicates that they are not related.
OGC abstract specification	A list of criteria to specify the core requirements for the development of a standardized DGGS framework in order to claim the conformance.
Partition set	A collection of elements such that each element can be a member of only one set at the next more generalized hierarchical level, to resemble a three-based hierarchy.
Q-analysis	A technique which utilizes concepts of algebraic topology to model system structure via simplicial complexes, to study system connectivity and interaction between its components. Q-analysis ranks each simplex according to its q-dimension, and compares its connectivity with other simplices. Q-analysis is also known as Polyhedral Dynamics.
Q-connectivity	Implies connectivity between simplices that do not have any vertices in common, but are connected through the chain of other simplices.
Q-dimensionality	A measure to indicate a total number of related elements or vertices that define a simplex. Q-dimensionality is always one unit less than the determined vertex total. Q-dimensionality is equivalent to the q-value and q-level notions.

Q-nearness	Implies connectivity between simplices that directly share simplex vertices or a face. Q-nearness also implies q-connectivity, but not vice versa.
Regular polygon	A polygon with equal sides and equal internal angles.
Shared-face matrix	A matrix-based representation of the q-near simplices in a set.
Simplex	A geometric object that represents a single element of a row set via array of elements in the column set in the incidence matrix.
Simplex vertex	A single element of a column set in the incidence matrix that is used to form a simplex.
Simplicial complex	A topological representation of connectivity between system components via collection of simplices and use of polyhedron to form and visualize a connected multidimensional structure.
Slicing	A process of converting information to a binary format via some benchmark parameter θ.
Traffic structure	A geometric structure defined by at least two sets used to model the multidimensional relation between set elements, and to represent the activity or behaviour that can be attached to the backcloth structure.