

ExPAND

Expanding Primary Attributes for National Decision Making

By

Brian Low

B.Sc., University of Victoria, 1993

A Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

In the Department of Geography

© Brian Low, 2005

University of Victoria

All rights reserved. This thesis may not be reproduced in whole or in part, by
photocopy or other means, without the permission of the author.

Supervisor: Dr. Olaf Niemann

ABSTRACT

Canada is committed to report on the current state and condition of its forests in support of national commitments such as Montreal Process, Convention on Biodiversity, Kyoto Climate Accord and others. National reporting requires data to be extracted, collated and analyzed from multi-resolution, multi-scale data sets held by the Federal, 10 Provincial, and 3 Territorial jurisdictions. New methods and techniques for the integration, storage and analysis of data held by independent custodians over a distributed infrastructure are reviewed and applied for this project. Problems occur when required data are located in a distributed infrastructure. Accessing, analyzing, and reporting on this data has been time consuming, complicated and expensive.

The object of this study was to investigate how spatial operations can be efficiently undertaken in a distributed environment without copying the data to a centralized location for analysis. A study of tessellations, re-sampling approaches, geospatial Web services, and spatial data handling was conducted to create and develop an approach that will allow for the analysis of distributed data sets. The resultant system, Expanding Primary Attributes for National Decision making (ExPAND) allows for the distributed analysis and synthesis of independently collected and stored data held on a distributed network of data warehouses. ExPAND was developed and designed to perform over a formalized international standards based infrastructure. These standards and specifications are created and maintained by the *Open Geospatial Consortium* (OGC). Any infrastructure that adheres to OGC's Web Mapping Service specifications can benefit from ExPAND. This thesis draws on a subset of Canada's *National Forest Inventory* (NFI) over an OGC based architecture as a test and demonstration of ExPAND.

Supervisor: Dr. O. Niemann, (Department of Geography)

Table of Contents

Abstract	ii
Table of Contents	iii
List of Tables	iv
List of Figures	v
Acknowledgements	vii
Chapter 1 Introduction	1
Chapter 2 Tessellations	12
Chapter 3 Re-Sampling	23
Chapter 4 Geospatial Web Services	36
Chapter 5 Spatial Data Handling	50
Chapter 6 National Forest Inventory - Use Case Scenarios	72
Chapter 7 Conclusion	92
References	98
Appendix 1 Database Table Relationship	103
Appendix 2 Style Layer Descriptor Example	104

List of Tables

1	Results of rotations with 3 different re-sampling techniques	29
2	Results of re-sampling to a 10 meter pixel with 3 different techniques	31
3	URI example of the OGC map request	39
4	URL for WMS request to create 1500 x 1500 pixel image	46
5	Unique colours at differing image sizes	47
6	Cell size summaries between image sizes	48
7	Re-sampling database table	58
8	Summary Count database table	59
9	TILE_THEME_IDENTIFIER database table	60
10	CELL_VALUES table structure	61
11	CELL_VALUES database table example	62
12	THEME_DESCRIPTOR table structure	63
13	TILE_DESCRIPTION table structure	65
14	Cell Summaries table structure	66
15	Cell Summaries example table	66
16	NFI theme names	76
17	NFI photo plots in UTM Zone 9 and 10 within this study	79
18	Vegetation types from NFI Land Cover Classification System	81
19	Comparison of traditional GIS with SQL example for photo plot# 1389966	87
20	Comparison of traditional GIS with SQL example for photo plot# 1403731	88

List of Figures

1	Intersect diagram	7
2	Process description for ExPAND	9
3	The classification of tessellation on the database space	16
4	General tessellation taxonomy	17
5	Lambert Conformal representation of Canada	18
6	Transverse Mercator representation of Canada	19
7	Universal Transverse Mercator zones in Canada	20
8	Parent-Child Relationships and identification of children cells	21
9	Re-sampling for orientation	24
10	Re-sampling for cell size change	24
11	University of Victoria with the Cornett building clipped out	28
12	Results from a Web Mapping Service as identified in Table 3	39
13	Image overlay example	43
14	HH-Code index method	53
15	Data stores and Web services	56
16	Re-sampling polygons onto a tessellation	57
17	Two grids and corresponding CELL_VALUES subset	67
18	Themes / layers for a given 2 x 2 km NFI plot	75
19	Terrestrial Ecozones of Canada	77
20	British Columbia study area within UTM Zones 9 and 10	78
21	Subset of Photo plots within the Pacific Maritime and Montane Cordillera Ecozones	84

Acknowledgements

I wish to thank my committee members for their commitment to this project. This study was conducted with the help, direction and perseverance of Dr. Olaf Niemann, Dr. Mike Wulder, and Dr. Hugh Barclay.

Dr. Robin Quenet, of the Canadian Forest Service, provided the greatly needed moral support and scientific and domain knowledge for this study. Mr. Rick Morrison's expertise of relational databases and spatial warehousing was also a great asset to this research.

Acknowledgements are given to the *Canadian Forest Service* (CFS) and the management of the *Pacific Forestry Centre* (PFC) for providing the financial support for this research.

Chapter 1 Introduction

Introduction

Maps being published over the Web have become popular in the past several years. The Web has given maps a new medium to be made available. The stylised representations or cartographic representations of geographical features being published over the Web have been made easier through international standards bodies. One such standards based organization is the *Open Geospatial Consortium* (OGC). This consortium has developed and published a specification on Web Mapping. The *Web Mapping Service* (WMS) specification created by the OGC has defined how maps are requested and generated over the Web.

With the ease of Web mapping, many organizations are publishing their spatial data. Mapping Web sites or Web portals have become a familiar resource on the Web. Web mapping portals are found at Natural Resources Canada's National Atlas, Fisheries Canada's Mapster portal and British Columbia's Land Information Portal. These Web sites and portal examples are based on the OGC WMS specification. One of Canada's leading initiatives is the *National Forest Information System* (NFIS or NFIS Canada). NFIS is programme developed with the provinces, territories, and the federal government. Information on Canada's sustainable forest management practices are being published through OGC standards on NFIS.

These mapping portals have been successful in publishing spatial and non-spatial data to their user communities. However, each of these infrastructures lacked the capability to analyze the spatial data over a distributed infrastructure. Simple analysis such as intersection and statistical summaries are missing from these important Web resources.

The objective of this investigation is to develop techniques to analyze spatial data held at distributed locations. Analyzing spatial data will allow for simple questions related to geographic locations and areas to be answered. Questions such as finding the area of a given feature can be answered. Questions dealing with the amount of area within another

area can be resolved. These simple spatial questions are typical of questions used for Web based analysis.

The study will be conducted utilizing Canada's National Forest Information System as the demonstration architecture; however, the approaches from this study will be available to any architecture based on the specifications defined within this project.

Canada's Forest Background

Canada is a forested nation and its forests play a dominant role in the country's environment, economy and culture. With approximately 400 million hectare of forests, Canada is the steward to about 10% of the world's forests and 30% of the world's boreal forest with 20% of the world's fresh water (CFS, 2004a). Canada's forests are the backbone of an \$81.8 billion forest industry (CFS, 2004a). Canada's forests are not only important financially but are also key to understanding ecological issues. Forests produce oxygen and assist in the removal of carbon dioxide from the atmosphere. The forests play an important role in the global carbon cycle (Kurz *et al.*, 2002). Forests management activities can contribute to carbon sources and sinks. These management activities would have a direct impact on the net balance of greenhouse gas sources and sinks for Canada.

Achieving sustainable forest management is a significant challenge and Canada plays a leading international role in all aspects of forest sustainability. Canada relies on its forests for trade, recreation, quality of environment, tourism and spiritual values. Canadians have consistently stated their belief that the sustainable management of Canada's forest is fundamental to national economic, environmental, as well as quality of life issues (Montreal Process, 1997).

Canada plays a leadership role internationally to define and measure forest sustainability (CCFM 2003). Major initiatives include Montreal Process, Convention on Biodiversity, Kyoto Protocol (United Nations Framework Convention on Climate Change 1997), and others. Existing reporting includes the State of the Forest, State of the Environment,

Canada's Report on the Montreal Process and Criteria and Indicators of Sustainable Forest management in Canada.

The Canadian Council of Forest Ministers reports on Canada's sustainable forest managements through many means. One mechanism is through a framework on criteria and indicators of sustainable forest management within Canada. The framework contains criteria for (CCFM, 1997): conservation of biological diversity, maintenance and enhancements, conservation of soil and water resources, forest ecosystem contributions, and benefits of forests to society. The monitoring of Canada's forests is critical to reporting on the state of the forests.

The Canadian Forest Service of Natural Resources Canada, publishes authoritative reports on the status of its forests. Within many of these reports, statistics based on area, ownership, and forested types are given. These statistics are given for known geographical areas, such as provincial boundaries, ecological stratifications and economic, electoral and other units.

The compilation of national reports is complex due to the fact that 14 different jurisdictions are involved and each with its own inventories, methodologies and processes designed to meet internal business requirements. This variability has resulted in a variety of different approaches to accessing and compiling information for national reporting. For example, individual jurisdictions may submit precompiled information for consolidation to form a picture or a model of Canada. This compilation approach is used in the National Forestry Database Programme. This model suffers from many limitations. These limitations include: significant effort on the jurisdictions, interference with the normal business practices of the data contributors, reporting is periodic, and the information rapidly dated.

Canada's commitment to report on the state of the forests to many international and national bodies has further exacerbated reporting complexity due to the requirement to integrate forest resources information with other resources information such as protected

areas, soils, hydrology, biodiversity, carbon stocks accounting, and climate data. Access to spatial and non-spatial data held by many jurisdictions and departments within jurisdictions is required for analysis in order to fulfill reporting commitments.

The methods, procedures, and scale of capture of resources information varies widely depending on the jurisdiction. The information collected differs from each jurisdiction because of funding, internal business requirements, and the anticipated audience. Spatial data capture is undertaken at many different scales and resolutions depending on its end use. For example, broad scale maps may be created at between 1:7 million and 1:1 million, while detailed maps may be created 1:50,000, 1:20,000, and 1:5,000. To add further complications, a number of different projections are in used throughout the country. Jurisdictions capture and manage their data in many different projections. Some jurisdictions use *Universal Transverse Mercator*, UTM projections, while others use Lambert, Albers, or even customized projections that are specific to a region. Integrating spatial information to give a detailed description of the country has been difficult and expensive both in time and costs. Many problems occur when dealing with an area as large as Canada given the number of independent jurisdictions that are contributing non-standardized information. The National Forest Inventory was developed to alleviate the variability in developing a compilation report on the status of Canada's Forest. This inventory was designed around a common framework that included: common projection, common definitions, common descriptions, common data formats.

NFIS Background

The complexity and cost of accessing, integrating and reporting on national forestry statistics and sustainable forest management resulted in a business case being undertaken on behalf of the *Canadian Council of Forest Ministers (CCFM)* to evaluate the need for a national forest information system. The CCFM agreed on establishing an information infrastructure to answer on matters relating to sustainable forest management in Canada in August 2000 (CCFM, 2004).

In cooperation with the Canadian Forest Service and GeoConnections Canada, CCFM undertook the *National Forest Information System* (NFIS) initiative. The purpose of the NFIS initiative is to develop new tools and methodologies for data integration, analysis and reporting on sustainable forest management in Canada. NFIS Canada was developed as an infrastructure to demonstrate Canada's sustainable forest management practices.

The vision for NFIS was that NFIS would provide a framework to (CCFM, 2004, p. 2):

- Provide ready access to the most current, consistent and reliable forest resources information,
- Provide the transparent integration of information across jurisdictional boundaries,
- Provide consistency in reporting thereby avoiding different answers being given to the same question,
- Reduce costs through the sharing of information technology, and
- Eliminate duplication in reporting
- Present an accurate picture of Canadian forest practices,
- Make national and international statements on sustainable forest management practices.

NFIS has been developed to provide Canadians and the international community with authoritative information about the state of Canada's forests and how Canada is meeting and monitoring its sustainability goals. It is an initiative that began in 2000 and provides World Wide Web access to spatial and non-spatial data from holdings of the federal, provincial, and territorial and other government agencies. NFIS is based on an architecture that relies on standards based specifications to interoperate between distributed partners. Each jurisdiction and partner within NFIS stores and manages their data, but allows for the access of the data through adopted standards. Users of NFIS can discover, integrate, analyze and display current and authoritative, accurate information on Canada's vast forests. NFIS maintains several Web portals (<http://nfis.org>) that allow for the discovery and display of forestry data from partnering jurisdictions.

Among its responsibilities, CFS is charged with delivering national forest inventory information with the expressed purpose of assessing and monitoring the extent, state, and sustainable development of Canada's forests in a timely and accurate manner. This requirement is being addressed through the *National Forest Inventory* (NFI). The NFI employs a plot-based design consisting of permanent observational units located on a national grid. Repeated measurements are conducted over a set time frame and changes in these plots are summarized and reported over time. The inventory is designed to be able to provide quantitative information on Canada's forest and related resources. The types of question being asked (CFS, 2004b) include:

- What is the total forested area within a specified area?
- What is the area of forest types within a specified area?
- What is the area of forest type by age within a specified area?
- What is the area forest types by protection status within a specified area?

These questions are also compounded when time is introduced. The questions can be posed as with a time or date constraint. What is the total forested area within a specified area in a given year could be asked. These questions may seem simple, but are actually very difficult to answer due to the complex nature and large volumes of data coming from multiple jurisdictions.

The Study

The challenge to answering the questions outlined above requires a mechanism supporting spatial operands that match the prepositions of "by" or "within". This mechanism must be invoked against spatial data that are held on different servers over a distributed infrastructure. The prepositions "within" and "by" can be considered to be equivalent to the spatial operation of intersection. Figure 1 illustrates how a box is intersecting with a circle. Only the common areas between the box and the circle are considered to be intersecting. The area of which the circle is "within" the box is shaded in Figure 1.

This study is designed to investigate how certain this spatial operation can be implemented over a distributed network. The study and system is given the acronym

ExPAND (*Expanding Primary Attributes for National Decision Making*). ExPAND is developed as a mechanism, or a tool, to assist in the answering of questions based on common shapes and well known geometries over a distributed network.

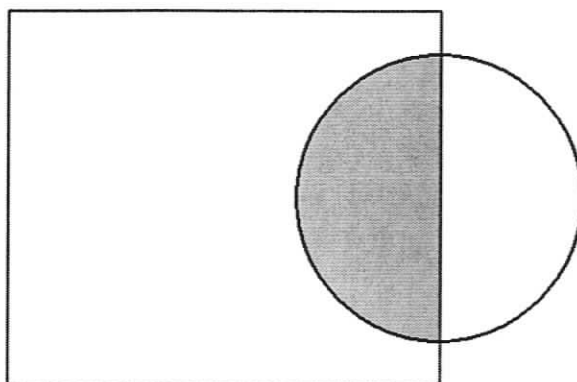


Figure 1. Intersect diagram.

The study was implemented on an application relating to forest inventory within a known geographical area. The question of “What is the area of unique forest types within a specified area?” is addressed in this study. The study looks at methods to address this and other similar questions that can be answered through the application of the spatial operation of “by” and “within” in a defined set of criteria. The spatial operations must be capable of being conducted where data are:

- held in a distributed environment,
- in the original projection, and
- stored in original formats.

In addition, the procedures must work with data from the National Forest Inventory and be invoked through the National Forest Information System.

An alternative to traditional *Geographical Information Systems* (GIS) vector spatial overlays was considered for this study. During this study, Web Services are looked at to answer the questions associated with the prepositions “within” and “by” in questions and

the use of the intersection / union spatial operations. Traditional GIS has limitations that can cause problems when analyzing distributed information holdings. These limitations include that of data transfers, accuracies, and version issues. The majority of forestry related data required for national reporting are held by provincial and territorial jurisdictions and other government agencies. The conventional approach, bringing the data to a centralized store for analysis, requires significant resources and knowledge to address different formats and projections. Topology problems can also arise from the conventional analysis approach if it is based on vector and non-raster data. Data received from partners may have topological errors that need to be corrected before being analyzed. This might be a result of differing tolerances when processing and capturing of data. Other issues that are important in this approach: the problem of stale data, version of data, and the requirement by the contributors to vet the information before it is used.

Adherence to *Open Geospatial Consortiums* (OGC) specifications along with business rules allows NFIS to circumvent these issues. NFIS uses OGC Web Mapping Service and Web Feature Service to access the authoritative and most current information published by the respective jurisdictional servers.

An examination of how *Web Mapping Services* (WMS) can be used for national reporting requiring spatial analysis was conducted in this study. As a component of NFIS, ExPAND will allow analysis of distributed raster data over WMS by re-sampling and storing the information to a common grid. Data needs to be compiled in a manner that will allow for the easy storage, access, and interrogation.

The information sets used in this investigation include Canada's National Forest Inventory and the Terrestrial Ecozones of Canada. The NFI is based on a systematic sample of permanent sample plots covering the entire extent of Canada (Gillis *et al.*, 2003). The permanent sample plots consist of two kilometer by two kilometer photo plots established on a 20 km by 20 km grid for a one percent sample of Canada's area. These plots are the primary data collection tools within the NFI project. The terrestrial

ecozones (15 nationally) provide a generalized ecological classification at a subcontinental scale based on abiotic and biotic factors (Marshall and Schut, 1999).

To conduct this study, several important investigations were required. Key areas of study include re-sampling methods, tessellations or common grids, distributed Web services, and data handling. For this study to be successful, the topic areas mentioned need to be examined and outlined. Figure 2 illustrates the processes needed to conduct the ExPAND study.

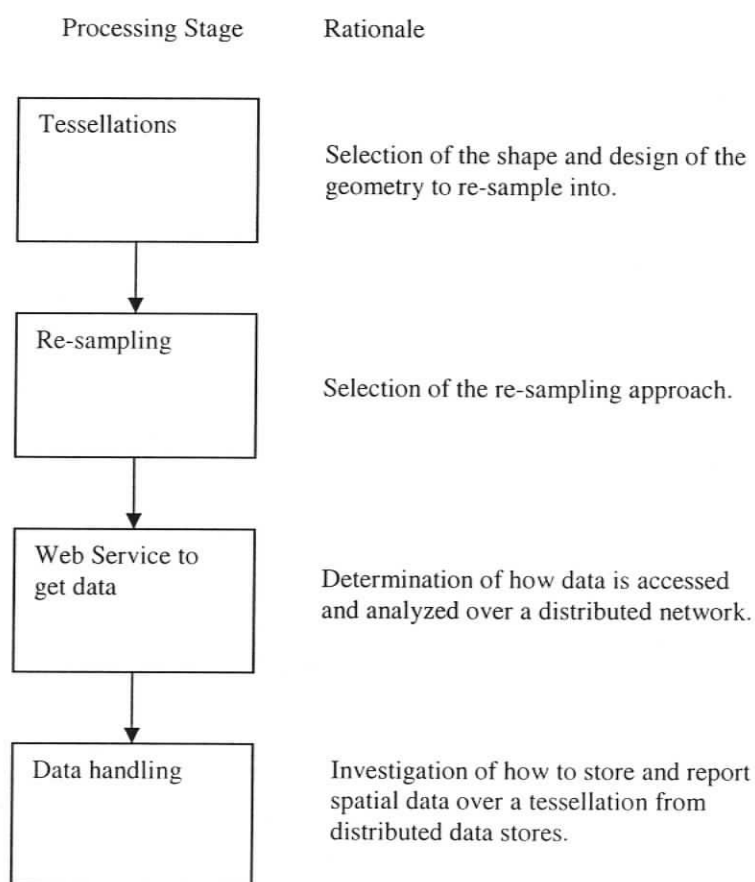


Figure 2. Process description for ExPAND.

The use of global tessellations provides a method of combining and storing data of regional and national scope. Global tessellations allow for the storage and representation of data over the entire Earth or large areas. Requirements for tessellations and cell divisions have been set out for national reporting and for this project. These include

common cell size (area) and a relationship to that of a parent cell. Hierarchical relationships are to be maintained along with information about identity. These fundamental requirements are described in Chapter 2 along with issues related to some of the typically used global tessellations.

In Chapter 3, a review of different techniques of re-sampling data is achieved with an emphasis on results of these re-sampling techniques. Critical requirements for re-sampling are identified and the selection of a re-sampling method is also described in Chapter 3. Re-sampling of the spatial data will allow for the comparison and interrogation of geometric features against others.

Data access is critical to this study. Web Services have provided a means to access spatial and non-spatial data through mechanisms that describe inputs and outputs. The OGC has introduced a set of specifications that allow for the communications of spatial data over the World Wide Web. Web Services and how the geospatial community has used Web Services for interoperability are summarized in Chapter 4.

Spatial data handling and storage of global/national data has been a challenge. Compatibility, version control, temporal changes and other issues have plagued the geospatial community. A process model that allows for distributed data to be re-sampled onto a common tessellation then stored within a relational database is described in Chapter 5. After being re-sampled, and stored in a relational database such as Oracle or PostGres, the spatial data can be analyzed using simple query languages like *Structured Query Languages* (SQL).

In Chapter 6, a subset of the National Forest Inventory is used to confirm the processes within this project. Several scenarios with the National Forest Inventory are described. These scenarios use ExPAND to demonstrate the practicalities of this study. ExPAND will be used as a mechanism to assist in the compilation of summary attributes for the National Forest Inventory programme.

In the concluding chapter, the entire project is summarized and future studies and investigations are discussed. The possible uses of this work and enhancements are outlined for future considerations.

Chapter 2 Tessellations

Introduction

Tessellations are a repeated geometric design that covers an area without gaps or overlaps. In this chapter, a discussion of the benefits of selected spatial tessellations is given. Requirements for tessellations that span large areas are becoming increasingly common. These tessellates are used for things as national reporting and statistics generating. Creating reports based on large areas based on multi-jurisdictional and multi-departmental information holdings have been time consuming and complicated. National reports such as the State of Canada's Forests 2004-2005 is an example of a report that is produced from multi-jurisdictions and multi-departments. Complications related to data management and data manipulation include: data processing procedures, edits, additions, merging, conversions, translations, re-projection, and etc.

Tessellates are used within this study as a common geographical framework to register or define values onto. The tessellations are used to record specific information about that values or attributes associated with the geometric shape. That is, for a given 5 x 5 meter area, the tessellate representing that region would contain information representing that 25 square meters.

Tessellation allows for the symmetrical coverage of the surface of a plane without overlapping or leaving gaps and provides for combining and indexing spatial data at regional and national scales. They allow for the storage, referencing, and representation of large size data holdings and allows for the efficient summary and analysis of these data for national reporting. Tessellates can be used to efficiently summarize large data gigabyte – terabyte holdings into smaller dataset.

Tessellations provide a quick and efficient way for storing and retrieving data. Analysis can be done without computationally intensive spatial operations. Each tessellate contains the location, the attributes, and the size of each unit. Tessellates can be compared with each other and summarized based on inherited information. A tessellate's neighbour can be queried and compared.

Tessellation

Song *et al.* (2002) describe the requirements for modern environmental monitoring and modeling. They state that partitioning the earth's surface into a global grid optimized for survey sampling and unbiased, spatially complete data collection of relevant environmental phenomena is required for global monitoring. Major scientific initiatives have recently emerged that attempt to monitor the state of the global environment and assess the condition of ecological resources over the entire earth. Brooks (1999) writes that the key issue is how to integrate both spatially and temporally disparate data for global reporting. Brooks used global grids to collect and access data for the scientific community.

An approach to this issue is to partition the earth into sampling or analysis units that form a hierarchy. Song *et al.* (2002) state that the ideal global partitioning system would consist of grid cells equal in surface area and identical in shape. It is easier to summarize and report on shapes that are of equal area and shape. Adherence to these two criteria makes sampling and analysis over the entire earth easier.

The division of a given space into non-overlapping regions can be considered a tessellation. Lee *et al.* (2000) describe the purpose of tessellation is to create "buckets" to contain data in a computer. Attributes can then be stored in the "bucket" or data record representing the tessellate.

Tessellations have many uses and benefits. Goodchild (1994) formulated a set of criteria for the evaluations of global grids. Global grids are:

1. Areal cells constitute a complete tiling of the globe, exhaustively covering the globe without overlapping.
2. Area cells have equal areas.
3. Areal cells have the same topology (adjacency, connectivity, relationships)..
4. Area cells are the same shape.

5. Areal cells are compact (should approximate a circle, the rounder the better, coordinates of vertices to be close together).
6. Edges of cells are straight in a projection.
7. The midpoint of the arc connecting two adjacent cell centers coincides with the midpoint of the edge between the two cells.
8. The points and areal cells of the various resolution grids that constitute the grid system form a hierarchy that displays a high degree of regularity.
9. A single areal cell contains only one grid reference point.
10. Grid reference points are maximally central within areal cells.
11. Grid reference points are equidistant from their neighbors.
12. Grid reference points and areal cells display regularities and other properties that allow them to be addressed in an efficient manner.
13. The grid system has a simple relationship to traditional latitude-longitude graticule (historically data is collected using latitude-longitude and thus makes it easier to re-sample).
14. The grid system contains grids of any arbitrarily defined spatial resolution.

The above requirements for an ideal tessellation are not easily met. Like many other problem sets in geomatics, it is necessary to decide on a balance between needs and criteria requirements. Many uses of tessellations may only fulfill a small list of these requirements. Kimerling *et al.* (1999) stated that partitioning the globe into a global grid with prescribed fineness results in distortions of area and or shape. Different partitioning results in different distortions. By using the criteria set out by Goodchild (1994), it is possible to balance the needs of the tessellation with that of an ideal grid, when considering the overall objectives of the grid.

There are many benefits to implementing tessellations as the principal data type. Relational databases are one way of storing tessellations. When tessellations are stored in a relational database, all traits associated with the tessellate are inherited. The strengths of relational databases for efficient summations, fast queries, structured queries, and transaction control can then be realized.

Version control of data can be handled and maintained for each tessellate by the database. Temporal models can be created with multiple attributes stacked on top of each other for a given tessellate. Tessellations with attribution can be representations of a feature's area, characteristics, and temporal identity. Comparisons of temporal data can be simplified when a standard tessellation is adopted and comparisons over time are reduced to identifying differences in attribution between versions. Langran (1993) describes how a temporal grid can be used to represent change over time. Langran (1993) uses a grid or tessellation that represents the space and time components and can be modeled to show changes over time.

Surfaces and gradients can also be created over a tessellation. These gradients can be created from sample points and expanded over the surface. Geo-statistical operations, such as Kriging, can be conducted on the tessellation to extrapolate for areas of missing data. Temperature, climate, precipitation, slope, elevation, and aspect can be modeled using appropriate attributes associated with tessellates.

Improved and faster numerical operations can be achieved through the use of tessellated data. Models can be created to take advantage of the nature of the data type. Spatial indices and unique identification allow the attribute to be easily retrieved. Spatial relationships can be conducted without the use of intensive geographical computations. Simple distance calculations can be conducted by walking through or traversing tessellates within a tessellation.

Classes of Tessellation

Lee *et al.* (2000) describe two basic forms of tessellation. The partitioning of space by following a feature is a *Feature-Primary Tessellation* (FPT), and the partitioning of space by an arbitrary decomposition is a *Space-Primary Tessellation* (SPT). Figure 3 illustrates the classification of tessellations used by Lee *et al.* Lee *et al.* further subdivides these two classes. In FPT, a further subdivision is formed with the classes of hierarchical or

non-hierarchical features. Lee's example of a hierarchical class in FPT is the jurisdictional boundaries created from the different levels of government. This can be demonstrated with an example of a municipality that falls under the jurisdiction of a city that is in a province, which is in a country.

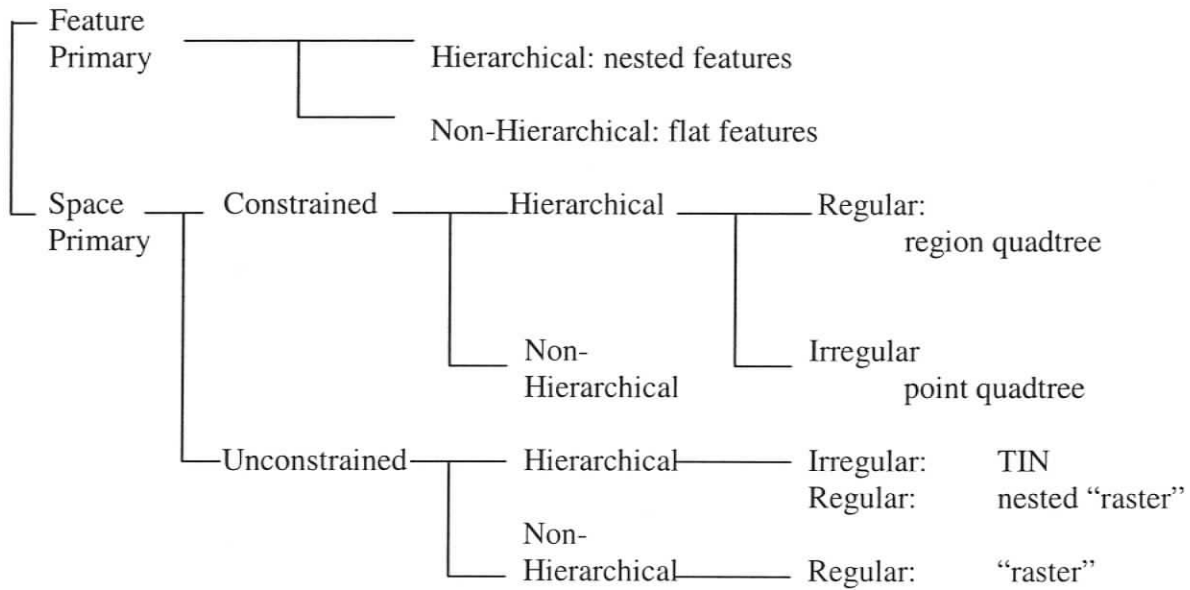


Figure 3. The classification of tessellation on the database space. Lee *et al.* (2000).

The SPT class may also be subdivided. SPT is subdivided into a constrained and unconstrained subclass. The constrained subclass is constructed from geographical features. Lee *et al.* (2000) identified that a *Triangular Irregular Network* (TIN) was an example of a constrained class. TINs do not represent real world features, but rather are constructed on real measurements. The unconstrained SPT subclass has cells that are independent of the location and shape of features.

Dutton (1994) created a general tessellation taxonomy that was later revised by Kimerling *et al.* (1999). At the top of this taxonomy (Figure 4) is the division of tessellating the sphere or spheroid and of map projection surface tessellation. Under the category of sphere, it is divided into three categories. The three categories are quadrilateral cells, platonic polyhedron, and Voroni tessellation.

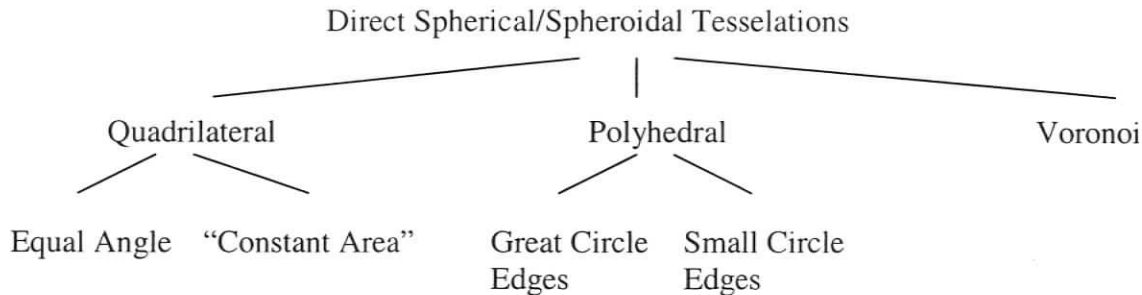


Figure 4. General tessellation taxonomy. Kimerling *et al.* (1999).

Equal Angle Quadrilaterals are equal latitudinal and longitudinal extents like that of the ETOPO 5 minute global digital elevation model. Constant Area Quadrilaterals are cells that are equal area and the latitudinal component is constant while the longitudinal is variable. Having constant longitudinal component and a variable latitudinal component is also a feasible approach. The strategy is to maintain a constant area by only varying one of the components of longitude or latitude. The polyhedral class is based on the faces being constructed from the edges of the great circle and the small circle.

Selecting the Tessellation

ExPAND required a tessellation be created to encompass the study area. Each tessellate must be of equal area with the study area. The shapes of each tessellate must be the same. ExPAND would use these tessellations as a means to compare geographical shapes easily.

After investigating the possible choices for tessellations as outlined by Lee (2000) and Kimerling (1999), several possibilities were considered. The choice of tessellations was also made more complex with the requirement that the tessellation must be compatible with that of a sampling mechanism developed by the Canadian Forest Service *National Forest Inventory* (NFI) Program. Wulder (2001) describes the primary data collection tool for the NFI project which is based on photo plots that are systematically placed on a 20 km grid. The photo plots are of 2 km by 2 km squares. This existing grid of photo plots must fit into the hierarchical nature of any tessellation method selected. To meet the

requirements of this study, a simple tessellation was selected. The tessellation of choice was based on the regular nested raster as described by Lee.

The choice of a projection onto which the tessellation was to be placed was investigated. Projections may limit the geographical extents represented by a map. Geometric distortions are introduced by map projections. The analysis and representation of information held in standard map projections is not easily accomplished. When reporting on an area the size of Canada or North America, portraying data on a typical “flat map” requires making a number of compromises. Map projections have a direct affect on shape, orientation, area and distance – these are decisive factors in choosing a projection as no one map projection can accurately represent all these factors.

Examples of distortions related to map projection can easily be found. Figure 5 illustrates the shape and orientation distortion in a Lambert Conformal projection (central meridian of -95.0 degrees and latitude of origin of 49.0 degrees) used for Canada.

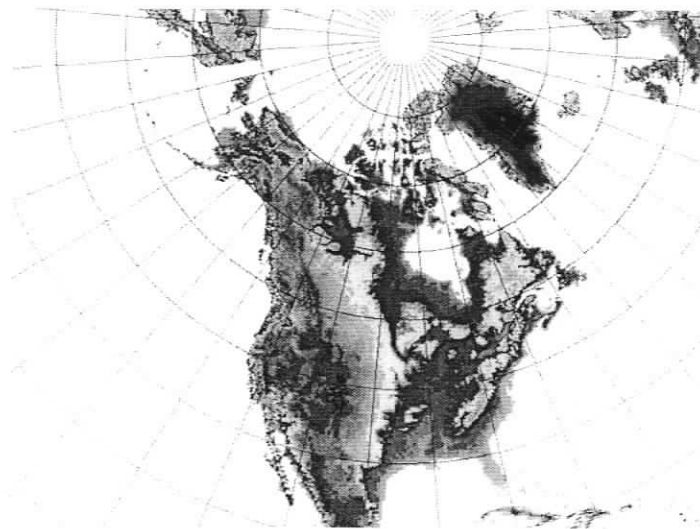


Figure 5. Lambert Conformal representation of Canada.

Figure 6 illustrates the shape, area and distance distortion of representing data in a Transverse Mercator projection true for central Canada. Where relatively small areas are

being analyzed, distortions related to the selection of map projections can be minimized. However, for large study areas the selection of projection can have a significant effect particularly with respect to surface area and shape.

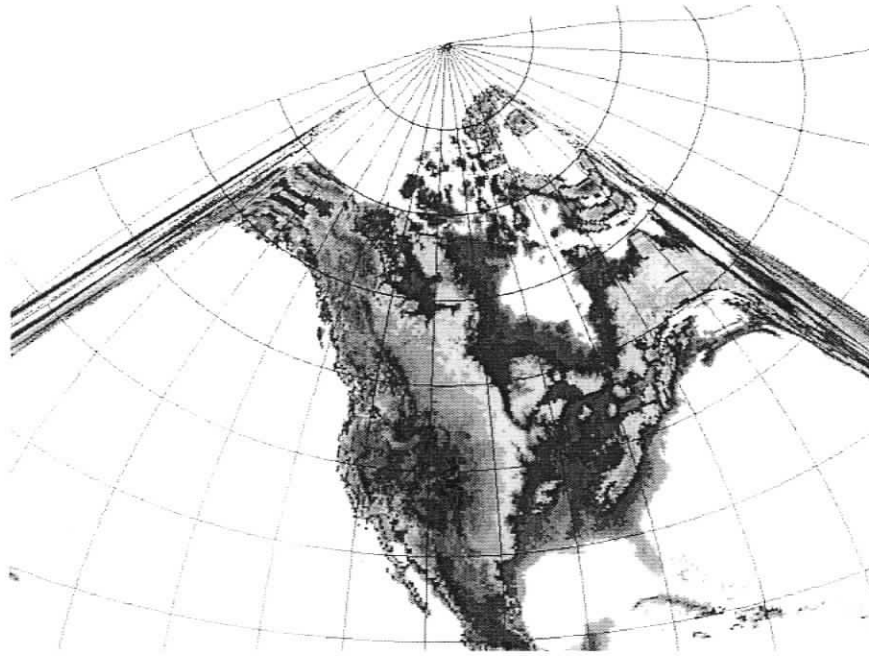


Figure 6. Transverse Mercator representation of Canada.

No projection can meet all the requirements of shape, area, distance, and orientation as well. The distortions in shape, distance, area and direction for the projection of choice must not impact the regular partitioning of the tessellation. The tessellation and projection must be able to allow for the partitioning of a surface area into non-overlapping working units or cells that can be associated with attributes.

The tessellation of choice was created on a regular nested cell structure of square boxes. The square cells were created on a Universal Transverse Mercator (UTM) projection. One factor in choosing the UTM projection was that the collected data's extents were defined in UTM coordinates. Having the tessellation based on a UTM projection allowed each cell of the tessellation to be the same size and shape. The UTM projection divides the Earth into zones that consist of six degrees of longitude. Each zone has a central

meridian upon which the projection coordinate system is based. UTM Zones 7 through 22 were used to represent Canada's land mass. Zone 7 has a central meridian of 141 degrees west and Zone 22 has central meridian of 51 degrees west. Figure 7 shows the UTM zones covering Canada.

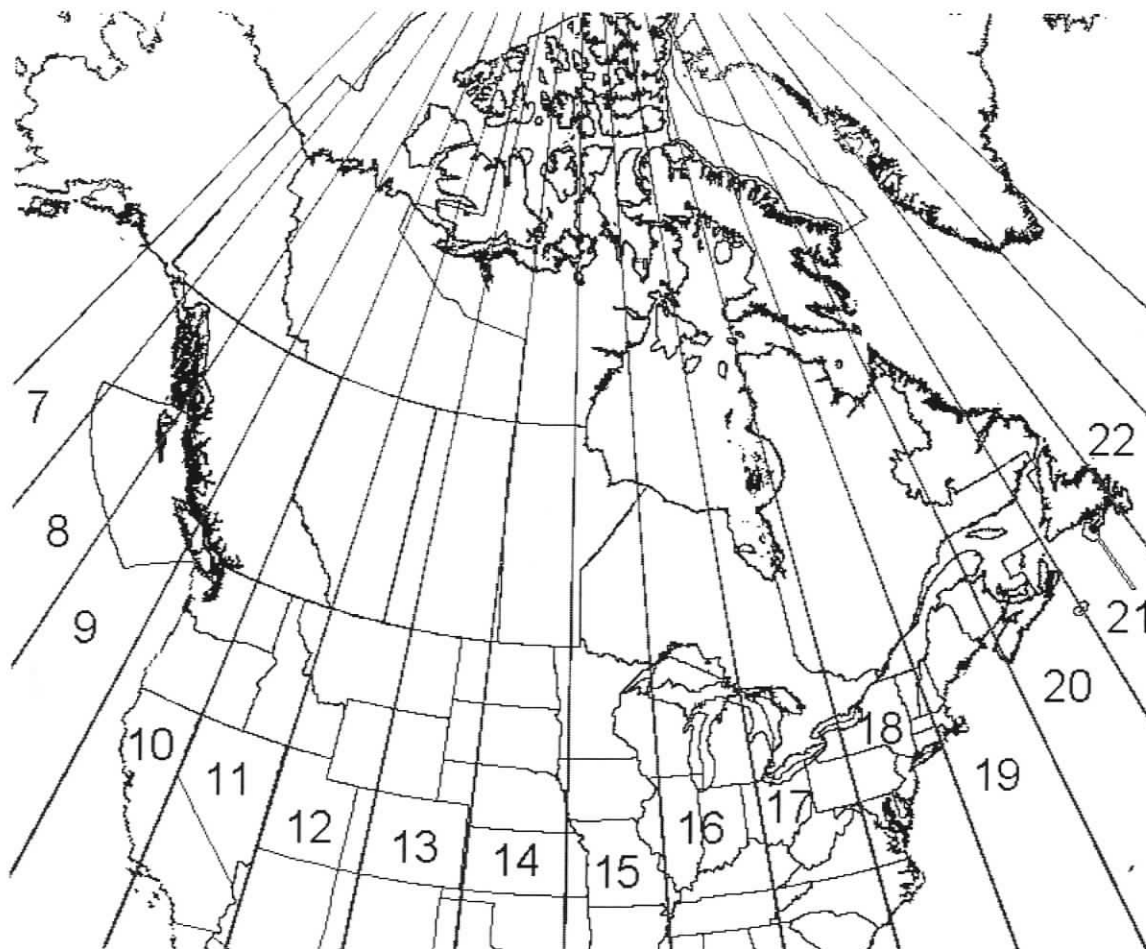


Figure 7. Universal Transverse Mercator zones in Canada.

This grid system meets many of the Goodchild (1994) criteria outlined earlier. The tessellations were created in a manner that would make them hierarchical, equal (with limited distortion) area, and common shape. Each UTM Zone is tessellated into square grids of equal area and shape. The combination of all the UTM tessellations gives a virtual National tessellation. In theory a virtual global tessellation could be created with

all the UTM zone tessellations. For this study an equal area within the study bounds is a requirement, and thus the UTM projection will meet this requirement within a localized area. The UTM projection system also provides a structured means of managing data through the use of UTM Zones.

Each parent cell can be broken down into children tessellations. This hierarchical parent-child relationship can go on until a cell resolution has been selected that is suitable for the required use. Figure 8 shows how a Parent Cell "A" of 2 km x 2 km that is subdivided into 4 1km x 1km children cells labeled A1, A2, A3, A4. Each of the children cells A1, A2, A3, A4 are cells that reference or know their relationship to that of the parent cell A. The parent-child relationships are stored in database that is described in Chapter 5. This mechanism of describing the parent-child relationship allows for relationships of differing shapes and topologies.

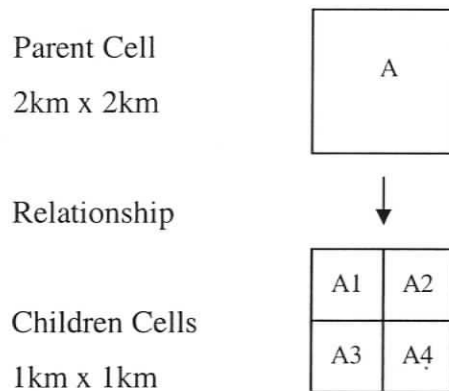


Figure 8. Parent-Child relationships and identification of children cells.

The smallest tessellation cell for this project is a 5 x 5 m cell. The parent cell is a 2 km x 2 km cell. The parent cell contain 400 x 400 (160,000) individual children cells. These 5 x 5 m cells are designed to intersect with the National Forest Inventory project 2 x 2 kilometer photo plots.

Wehe (1982) concluded that as cell size increased, the accuracy of maps and inventories produced decreased. Fellows and Ragan (1986) found that although the increase of cell size reduced accuracy, this reduction was dependent on the feature type and use of the data set. The 5 x 5 meter cell was selected to allow for the spatial coincidence between the sampling cells and the parent cell.

Summary

The use of global tessellations allows for the efficient storage, capture, management, analysis, representation and reporting of geographical space over a large area. The concept of tessellations and partitioning of space has been extensively researched including a list of requirements and criteria for areal cells developed by Goodchild (1994). Large area mapping has many limitations associated with projections; however, tessellations on the sphere or globe have been shown to provide a solution. The study area, Canada which covers a subset area of the globe, allowed the selection and application of a UTM based tessellation with hierarchical properties

The decision to select a UTM based tessellation was heavily influenced by the requirement to be compatible with the existing NFI sampling grid. The tessellation method of choice introduced a hierarchical nature that allowed children cells to inherit their relationship and location based on their parent cell. The parent cells in this case were the NFI 2 x 2 km photo plots are distributed over Canada on a regular 20 km systematic grid.

The chosen tessellation method has equal areal cells with common topologies, including common shape and a common orientation in their respective projections. Each grid areal cell has only one reference point. The tessellation is a hierarchical regular nested raster classification as described by Lee (2000).

Chapter 3 Re-Sampling

Introduction

In this chapter, the technique of re-sampling is discussed. Re-sampling is commonly used in the field of remote sensing, signal processing and GIS. Re-sampling allows data sets to be compared and to be synthesized together on a common reference grid. Re-sampling is necessary when pixel values are required in a destination grid at resolutions, positions or locations differing from the source input pixels. Operations on multiple data sets in remote sensing normally require the technique of re-sampling on the source data. Comparisons and examples of different techniques of re-sampling are given. The re-sampling technique selected for the thesis will be discussed.

Definition

Re-sampling is the process in which a grid is assigned values that have been interpolated from a differing input grid. Ehlers (1997) defines re-sampling as the actual assignment of new digital numbers to the calculated pixel coordinates. Fisher (1997) defines a pixel as the smallest element of an electronic image, and can only be subdivided by creating repetitive information. Cracknell (1998) states the generation of image intensity values at a regular square or rectangular grid of geographical coordinates, from the navigated grid, is an interpolation procedure which is usually described as re-sampling. Wolberg (1990) defines interpolation as the process of determining the values of a position lying between its samples. Interpolation is achieved by fitting a continuous function through the discrete input samples. A newly re-sampled grid would be derived from an interpolation of the input grid.

A collection of data sets that have been re-sampled to a common reference grid can be used in conjunction with each other. This allows operations such as arithmetic, relational, statistical, and analytical to be performed on the coincidental collection. Operations on data sets that were not referenced or re-sampled to a common grid will produce erroneous results. To compare data or to perform analysis, all data sets must be registered or coincide with each other.

Examples of Re-sampling

The technique of re-sampling is commonly used in remote sensing. There are many situations in which re-sampling would be required. Re-sampling and transformations are required when input data needs to be oriented in a different manner (Figure 9). Re-sampling is appropriate where it is necessary to change cell sizes (Figure 10). Blacknell (1989) describes the use of re-sampling in: image-image matching, image-map matching, correction of geometric distortion and radargrammetry.

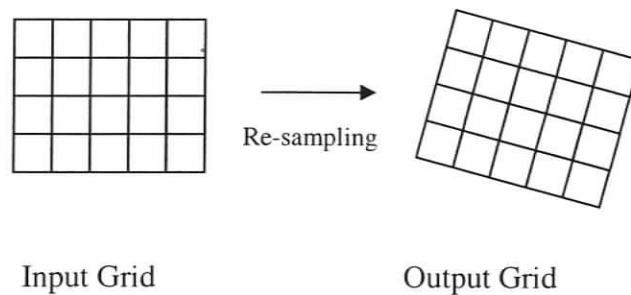


Figure 9. Re-sampling for orientation.

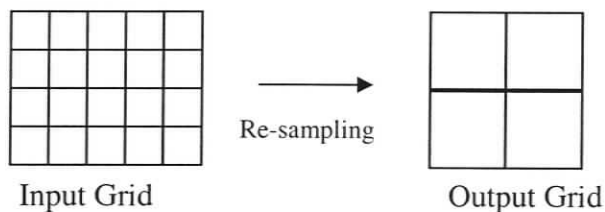


Figure 10. Re-sampling for cell size change.

Image to image matching requires the comparison of two images together. These two images will share some geographical area of interest. However, these two images may not be gathered or originate from the same source. With differing sources of images, many differences will be observed in the characteristics of the grids. Differences include the variations in pixel sizes and orientations of the grid (see Figures 15 and 16). These differences must be addressed for the data to be considered coincident. Having data that is coincidental makes analysis simpler.

Image to map matching is undertaken when remotely sensed raster data are compared to or used in conjunction with traditional GIS maps. Vector and raster map representations are stored in a projection system that most likely will not be compatible with raw

remotely sensed data. Re-sampling allows the remotely sensed image to be re-oriented and re-projected and the resulting grid is compatible with the map data. Having vector data available in a common reference allows process like feature identification to be performed.

Geometric errors can be found in many remotely sensed data sets. These errors can appear from sensor orbit, velocity, altitude, and attitude errors. Corrections for the Earth's curvature or shape must be addressed in some images through geometric corrections and ortho-corrections. Remotely sensed data that has been affected by the Earth's rotation must be addressed. Re-sampling is an essential part of image geometry and correction.

Methods of Re-sampling

All re-sampling methods require that the values produced are consistent with all the input information available. There are many interpolation methods available for re-sampling. These may include: nearest neighbour, linear interpolation, cubic interpolation, two-parameter cubic convolution, cubic splines, b-splines, windowed sinc function, Hann and Hamming Windows (Bernstein, 1983). In this chapter, descriptions of some of the common re-sampling methods are given.

The simplest of the interpolations is the nearest neighbour, also known as the point shift algorithm. The nearest neighbour algorithm decides the value of a pixel by taking the closest input pixel from the calculated location. Each interpolated value is assigned the value of the nearest sample point in the input grid. The value of the output grid pixel at location x, y is value of the input grid pixel at point location of u, v . The formula for nearest neighbour is: $I(x, y) = I(u, v)$ (Bernstein, 1983). The term, "I", refers to the pixel intensity or value. With this method, no new output values are created. All values in the output grid are found in the original data. Due to the nature of the interpolation, a shift in the pixel value may occur up to $\frac{1}{2}$ the pixel. This approach produces a blocky or step like appearance at times.

The bilinear interpolation method requires four neighbouring pixel (a 2 x 2 group) values to calculate the new output value. The output value is a weighted average of the 4 closest input cells. This method is based on two linear equations. Linear interpolations are considered first degree equations. They pass a straight line between every two known points.

Wolberg (1990) derives the linear interpolation:

“Given an interval (x_0, x_1) and function values f_0 and f_1 for the endpoints, the interpolation polynomial is $f(x) = a_1x + a_0$ where a_0 and a_1 are determined by solving equation 1:

$$[f_0 f_1] = [a_1 a_0] \begin{bmatrix} x_0 & x_1 \\ 1 & 1 \end{bmatrix} \quad (3-1)$$

This gives rise to the following interpolating polynomial:

$$f(x) = f_0 + \left(\frac{x - x_0}{x_1 - x_0} \right) (f_1 - f_0) \quad (3-2)$$

The cubic convolution interpolation method requires the use of sixteen of the neighbouring pixels (a 4 x 4 group) to calculate the new output value. Based on the sixteen nearest input cells, a weighted average based on distance is calculated as the new digital number. This process fits a curve through values of the sixteen nearest input cells. Bernstein (1983) describes the use of the 16 neighbouring values to compute the output intensity with equation 3:

$$I(x,y) = \sum_{m,n} a_{m,n} I(u+m,v+n), \quad -1 \leq m, n \leq 2 \quad (3-3)$$

This approach is an improvement over the bilinear approach, but may be considerably slower to process. The cubic convolution method tends to produce a smoother result in

images. The cubic convolution uses a larger set of points to calculate the new value. The extra points allow for more continuous result in the values of the resample.

The cubic spline interpolation is a piecewise continuous third degree polynomial. For N sample size points, there will be $N-1$ cubic polynomials. Each polynomial will pass through the sample points. The sample points are also known as control points. The polynomial for a given location (k^{th}) would be defined as:

$$f_k(x) = a_3(x-x_k)^3 + a_2(x-x_k)^2 + a_1(x-x_k) + a_0 \quad (3-4)$$

The coefficients are defined in terms of the data points and their first and second derivatives (Wolberg 1990).

Example Re-sampling Results

A sample image grid was created from a 1:20,000 (092B.043) scale digital ortho-photo. This ortho-photo covers the southern tip of Vancouver Island including with the University of Victoria (Figure 11). A clipped grid of the Cornett building, where the department of Geography has offices, was created from the original data set (Figure 11). The clipped grid of the Cornett building is composed of 130 rows by 110 columns with a total of 14300 pixels. Each pixel represents a 1 x 1 m area. It is composed of 201 unique different gray values or digital numbers. This tile of 14300 pixels was used as an input to the re-sampling exercises below.

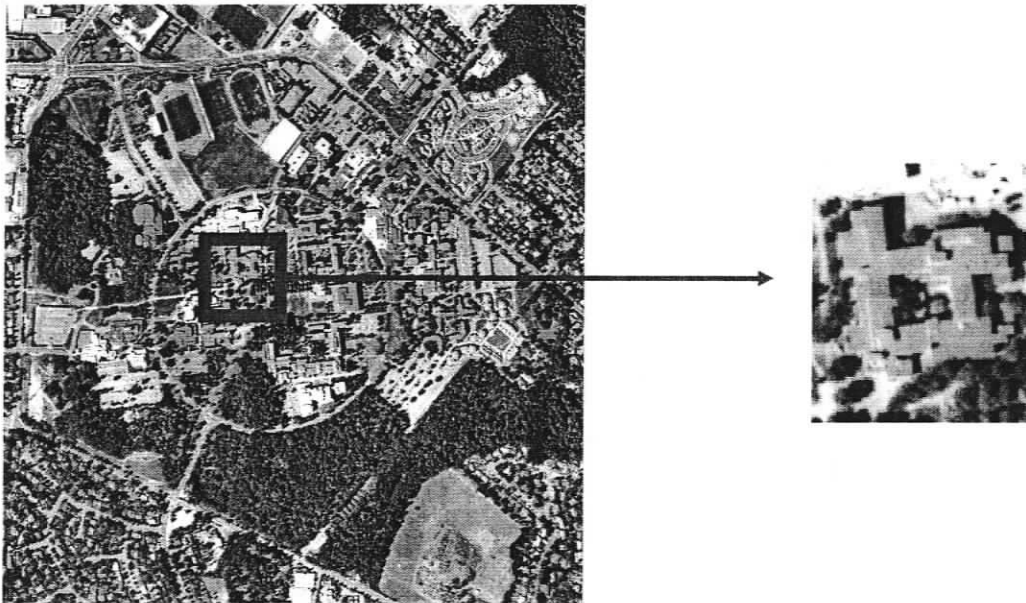
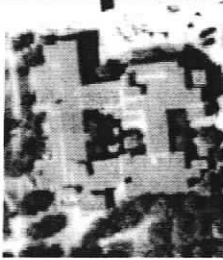
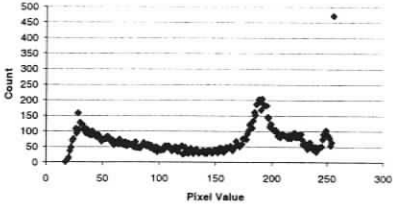
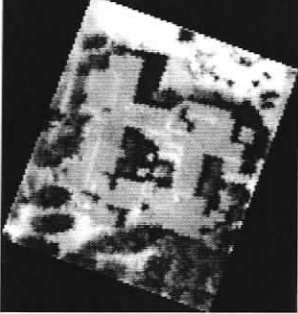
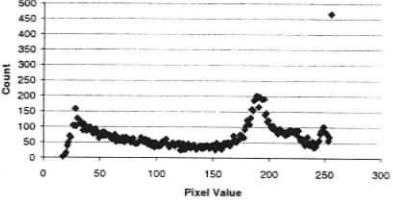
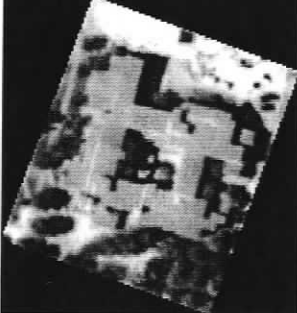
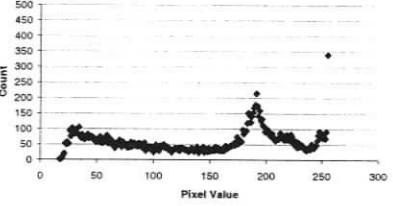
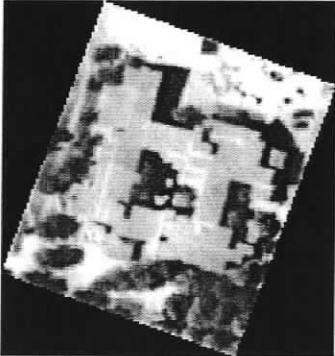
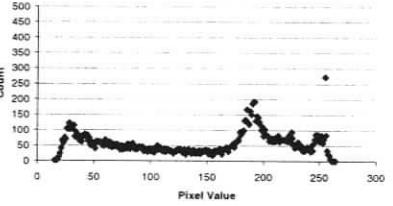


Figure 11. University of Victoria with the Cornett building clipped out.

Several commercial GIS software packages implement the three most common re-sampling algorithms, namely nearest neighbour, bilinear, and cubic interpolations (Anon. 1991). The re-sampling of grids for these tests was processed with the software package by ESRI. The ESRI GRID suite was used with the command RESAMPLE.

Table 1 shows the different results from the three re-sampling approaches with a rotation operation. The original image was rotated by 20 degrees using a ESRI's Arc/Info GRID suites. This function allowed the user to select one of the more commonly described re-sampling approaches.

Table 1. Results of rotations with 3 different re-sampling techniques.

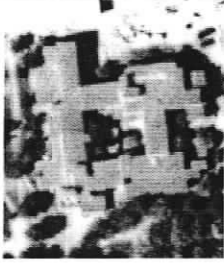
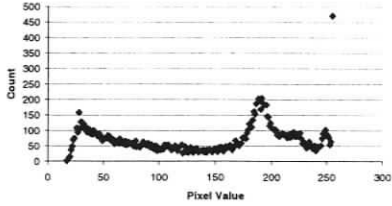
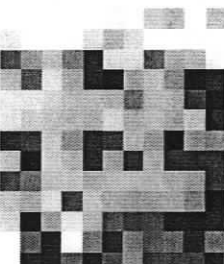
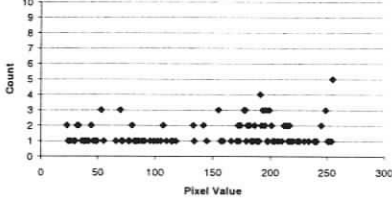
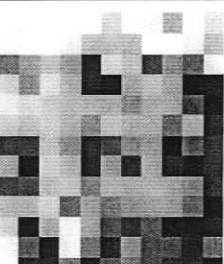
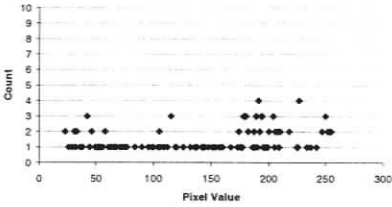
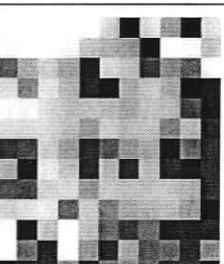
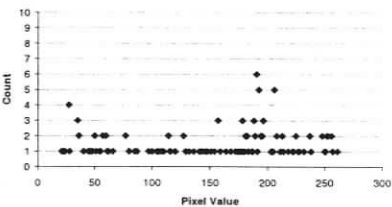
Image	Description	Distribution of Pixel Values
	<p>Original Image</p> <p>Number of unique Values: 201 Minimum Value: 17 Maximum Value: 255 Mean Value: 145.782 Standard Deviation: 73.533</p>	
	<p>Rotate 20° – <u>Nearest Neighbour Interpolation</u></p> <p>Number of unique Values: 201 Minimum Value: 17 Maximum Value: 255 Mean Value: 145.746 Standard Deviation: 73.507</p>	
	<p>Rotate 20° – <u>Bilinear Interpolation</u></p> <p>Number of unique Values: 238 Minimum Value: 17 Maximum Value: 255 Mean Value: 145.769 Standard Deviation: 72.501</p>	
	<p>Rotate 20° – <u>Cubic Interpolation</u></p> <p>Number of unique Values: 248 Minimum Value: 15 Maximum Value: 264 Mean Value: 145.772 Standard Deviation: 74.127</p>	

In rotating the image by 20 degrees, the three approaches gave differing results. The nearest neighbour interpolation resulted in a jagged edge or steps along the building edges. The bilinear interpolation shows a smoother rendering of the building, but may reduce image detail. The cubic interpolation has given a smoother image than both the nearest neighbour and bilinear operations.

New pixel unique values were calculated with both the bilinear and cubic interpolations. In the bilinear approach a total of 238 unique values were used to compose the image. In the cubic approach, there were 248 unique digital values. The original image had only 201 unique values. The nearest neighbour interpolation resulted in the same 201 unique values as the interpolation method suggested.

Table 2 shows the results from the three re-sampling approaches when the pixel sizes of the grids were changed to a 10 meter pixel. Instead having an image that is comprised of 130 x 110 pixels, grids are composed of 13 x 10 cells are generated. The resulting images have a total of 130 cells.

Table 2. Results of re-sampling to a 10 meter pixel with 3 different techniques.

Image	Description	Distribution of Pixel Values
	<p>Original Image</p> <p>Number of unique Values: 201 Minimum Value: 17 Maximum Value: 255 Mean Value: 145.782 Standard Deviation: 73.533</p>	
	<p>Re-Sampled pixel size 10meter - <u>Nearest Neighbour Interpolation</u></p> <p>Number of unique Values: 97 Minimum Value: 23 Maximum Value: 255 Mean Value: 151.846 Standard Deviation: 70.580</p>	
	<p>Re-Sampled pixel size 10meter - <u>Bilinear Interpolation</u></p> <p>Number of unique Values: 98 Minimum Value: 23 Maximum Value: 255 Mean Value: 149.692 Standard Deviation: 68.461</p>	
	<p>Re-Sampled pixel size 10meter - <u>Cubic Interpolation</u></p> <p>Number of unique Values: 97 Minimum Value: 21 Maximum Value: 261 Mean Value: 150.336 Standard Deviation: 71.079</p>	

It is noted that in re-sampling the image to a 10 meter pixel from a 1 meter pixel, the nearest neighbour approach resulted with 97 unique values. The original image of 130 x 110 had 201 unique values. This difference in unique values is due to the fact that the

nearest neighbour method only uses the pixel values closest to the output pixel. The remaining 104 unique values were never sampled or seen with this technique.

Comparisons

Many factors are used in selecting the method of interpolation for re-sampling. These factors include the complexity of the algorithms and the type and amount of data. The mathematical and computational requirements of each algorithm together with the desired output characteristics are the determining factors in selecting the most appropriate method.

The choice of re-sampling technique is also dependent on the type of input data. Those algorithms that average surrounding pixels may be useful with input data that is raw or unclassified. Classified grids such as estimations of biomass per cell would not work well with methods that average surrounding pixels. Averaging of pixels for a classified data set may result in a product that has values outside of the classification. Procedures such as the nearest neighbour do not result in new values or classifications. Bilinear, and cubic convolution interpolations result in output values not found in the input source. In both examples of the bilinear, and cubic interpolations, new unique values were found that were not present in the original data set. The generation of new unique values can confound further analyses. Analyses based on these newly introduced values may skew or influence results.

Smoothing of values can be achieved with the use of the bilinear and cubic functions. Sharp boundaries such as building walls or streets may become blurred because of the averaging process. Cubic interpolations normally result in a sharper result than that of the nearest neighbour and do not have the over-smoothing factor of a bilinear approach. The nearest neighbour approach results in a step by step look on straight edges. This jagged look is not found in the other two interpolation methods.

Computational time is a consideration when processing large images. In the previous examples, the input data source was relatively small. The input image was composed 130

x 110 pixels. The difference in computational time between the different re-sampling techniques for both rotation and resizing was unnoticeable. However, if a larger data set were used as input, a noticeable difference would be seen. The cubic convolution would result in the slowest of the three techniques. The fastest would be that of the nearest neighbour.

Discussion

ExpAND requires a method of re-sampling the data to a common framework. Having all datasets re-sampled to a common framework will allow users for the spatial operations to be conducted easily. The sources of data for ExpAND will come from distributed data stores across a network of servers. These servers will hold spatial data in differing projections and formats. The re-sampling process will allow for the capture of spatial features over a common framework. Having the capability to re-sample is an essential component for the comparison, analysis and portrayal of data.

For this project, several key factors require consideration when selecting a re-sampling technique. It is critical that no new values are to be created. The values being re-sampled can be considered categorical values that represent discrete classes. That is no new pixel values are created during the re-sampling process. The domain of values found in the re-sampled grid must be equivalent to or smaller than that of the domain of values from the original grid. Thus the use of a bilinear or cubic convolution interpolation is not acceptable.

The nearest neighbour solution allows for no new values created during the process, and thus meets the 'no new value' criterion. However, a second requirement of the project is that the sampling process actually represents the largest area of a given sample population (weighted by area). The largest area for a given discrete classification within the sample population is used as the resulting value. For example, if a sampled area has 80 ha of Douglas Fir as one classification and another area of 20 ha of Balsam classification, the re-sampling approach will result in a classification of type Douglas Fir for that sampled area.

To meet the criteria, a weighted area sampling approach is used. This is achieved in a similar manner to that of the nearest neighbour. But instead of finding the closest or nearest sample point in the sample grid, the process sums all the unique occurrences of values in the source grid and returns the value with the largest occurrence. Since each sample grid is of equal area within the sampled area, the weight applied to each cell is equivalent. With equivalent weights applied to the area of each cell, a modal average can be applied.

The value assigned to the sample is the largest occurrence (mode) of input values that fall on the sample grid. The mode describes a typical sample measurement in terms of the most common outcome (Agresti & Finlay, 1999). The mode is the category or interval with the highest frequency. The value assigned to an output grid x,y would be assigned the value of the largest occurrence within the input grid location within x,y .

This mode re-sampling approach has similar characteristics to that of the nearest neighbour. With this method, no new output values are created. All values in the output grid will be found in the original data. However, there is a chance that values from the source data maybe missed in the re-sampled output. A classification class that is never the dominant class in the region being re-sampled will not show up.

Summary

It is important to recognize the results of a sampling exercise. Different sampling methodologies result in different outcomes. Many sampling approaches result in data values that are not in the original domain set. The selection of appropriate sampling methods is dependent on the data sets and application. This study is primarily interested in discrete categorical data sets that have been stratified into known classes or themes. Proportional values or ratio based values are not appropriate, but can be placed into categorical classes for this analysis.

The weighted area approach to re-sampling will be used for this study due to the fact that it allows for the systematic sampling of data based on the frequency of occurrence within given area. No new classifications can be created from this process.

Chapter 4 GeoSpatial Web Services

Introduction

This study required access to spatial data from differing projections in a distributed environment. The use of Web Services is essential and key to accessing distributed data holdings. In this chapter, a discussion on the key role of Geospatial Web Services in a distributed geospatial reporting systems is outlined. Web Services are mechanisms that allow services to be called or invoked over the Web. Traditionally, the Web has been used as a means to find or to retrieve information. In the past several years, the Web has been able to support a number of different services to users.

Web Services

Web Services can be defined as (Kreger, 2001, p. 6) “an interface that describes a collection of operations that are network accessible through standardized *Extensible Markup Language* (XML) messaging. A Web service is described using a standard, a formal XML notion, called its service description. It covers all the detail necessary to interact with the service, including message formats (that detail the operations), transport protocols and location. The interface hides the implementation details of the service, allowing it to be used independently of the hardware or software platform on which it is implemented and also independently of the programming language in which it is written. This allows and encourages Web Services-based applications to be loosely coupled, component-oriented, cross-technology implementations.”

Web Services allow users access to technology or functionality over the web. Processes and functionality can now exist at distributed locations. The computing industry has been challenged with accessing data and operations at distributed locations for many years and Web Services are now meeting this need. Examples of Web Services can be found throughout the Web. Companies such as Google and Altavista use Web Services to allow third parties to access their search engines.

Web Services are being introduced into the Geospatial community through a distributed processing environment that allows for the manual and automated change detection of

remotely sensed data from distributed databases. Lampropoulos *et al.* (2003) developed and implemented different detection methodologies as Web services and tools.

Lampropoulos, uses change detection Web Services along with data stores within the Canadian Geospatial Data Infrastructure for precision agriculture. Registration and fusion algorithms are combined with access to data over the Web for this application. The Web service described in Lampropoulos *et al.* paper, allows for the integration of the latest statistical thresholding technologies for enhancing adaptive change detection algorithms (Lampropoulos, 2003).

The *Open Geospatial Consortium* (OGC) is a “Not for Profit” organization that develops specifications for geospatial and location based services. This consortium is composed of academics, companies, and government agencies interested in furthering geospatial standards. Its mission is to deliver spatial Web Service interface specifications, that are openly available for use.

Coppock and Rhind (1991) document the history of GIS. Coppock and Rhind place a timeline of computer-based GIS starting as early as the late 1960’s. They discuss the work of Tomlinson in 1966 in creating the *Canadian Geographic Information System* (CGIS). GIS was used heavily in resource management starting in the late 1970’s. Many great advances in information technology have made GIS technologies common tools for resource management. However, these advances in technologies have not provided a solution for vendor specific incompatibilities. Different formats and incompatible data types have been always time consuming and expensive and access to, and integration of spatial information holdings has presented a problem that is not easily addressed.

Web Mapping Service

OGC adopted Web Services technology in the late 1990’s. In 1997, OGC put in place a specification program to test Web Mapping and the interoperability experiment and program test-bed for Web mapping was created. These interoperability tests lead to the OGC endorsed *Web Mapping Service* (WMS) specification.

The WMS specification describes a Web service that fulfills three main functions. A WMS would be able to produce a map, answer basic queries about the content of the map and describe the map product or query response from the service.

WMS uses a picture or image to represent geographical elements and features. These images can be created as JPG, PNG, GIF, and other formats. The image created by a WMS service provides a representation of the data associated to a geographical location for a given subject or theme. The image created by the WMS service is a digital map. Later specifications of WMS allow for the user to control how the map is rendered. Cartographic symbolization was added to allow a user to control how features were rendered on the map.

To produce a digital map from a WMS service, all parameters or variables necessary to create the map must be passed to the service. These parameters include the exact coordinates of the portion of the Earth to be mapped, the layer or theme of the map, the size of the map, the format of the map, the projection of the map, and others variables. The WMS specifications allowed for a common way of describing and sending these parameters to the service. The WMS specification uses Uniform Resource Identifier (URI) to pass parameters to the service. Berners-Lee *et al.* (1998, p.1) define a URI as “a compact string of characters for identifying an abstract or physical source”.

Table 3 has an example of a URI that illustrates a request to a Web Mapping service. This URI passes the necessary parameters to a Web service that generated an image shown in Figure 12.

Table 3. URI example for an OGC map request.

Example URI	https://ca.nfis.org/cubewerx/cubeserv/cubeserv.cgi?VERSION=1.1.3&REQUEST=GetMap&SRS=EPSG:42304&BBOX=-2030482.788875392,399041.17865288,-2028836.695124608,400069.98724712&WIDTH=560&HEIGHT=350&LAYERS=BC_ORTHO:Common:NFIS-Common&STYLES=&FORMAT=image/gif&BGCOLOR=0xFFFFF&TRANSPARENT=FALSE&EXCEPTIONS=text/html&QUALITY=QUICKEST
-------------	---

This URI is requesting a map image that is created from an orthophotography layer in a Lambert Conformal projection at the given bounding location co-ordinates. Other parameters such as version control, background color, transparency control and error reporting are also passed to the service.

**Figure 12.** Results from a Web Mapping Service as identified in Table 3.

Web Mapping Services can render traditional GIS and Remote Sensing data as simple images that can be retrieved via the internet. Points, lines, polygons, complex elements, rasters, and grids can be rendered with WMS technologies. Some implementations of WMS also allow for the re-projection of spatial data in real time. This functionality allows users to request maps in projections other than the source data's original projection.

The WMS technology is in fact a mechanism to rasterize data. Similar to that of Raster GIS products, vector data is converted onto a regular grid. The WMS technology allows for the conversion of vector data to raster data in quick manner through an interactive method. Spatial data is rendered to image formats such as GIF, PNG, and JPEG. These formats can be thought of as simple raster data formats.

Given the results from a WMS request, spatial data can be rendered into raster images quickly regardless of formats and projections. The service supports the capability to quickly access large spatial warehouses in a common manner. This common request interface and a common description of results have made this service extremely valuable in the Web geospatial community. WMS has enabled competing products access to common datastores or warehouses. A particular vendor's client products can now use another vendor's server product. For example, ESRI has developed and distributes an interoperability extension for their products that will allow their products to read from standard WMS resources (Anon, 2004). This has been achieved successfully using this simple Web service for publishing and sharing spatial data.

Style Layer Descriptors

The WMS servers create a representation of the spatial data as an image. The image is created by a procedure that applies a theme or rule to the spatial geometry. A number of OGC WMS servers allow for *Style Layer Descriptors* (SLD) to describe how the service will render the geometry. SLD's are XML documents that describes rules and filters for creating choropleth maps based on attributes of the spatial or non-spatial data. For example, if a user wanted to create a graphical representation of a land cover map with

each vegetation polygon of type TC (Treed Coniferous) coloured as green, the following SLD could be applied as shown in Example 1.

```
<?xml version="1.0" encoding="utf-8" standalone="no" ?>
  <StyledLayerDescriptor version="1.0.0"
    xmlns="http://www.opengis.net/sld"
    xmlns:ogc="http://www.opengis.net/ogc"
    xmlns:xlink="http://www.w3.org/1999/xlink"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.opengis.net/sld
      http://schemas.opengis.net/sld/1.0.0/StyledLayerDescriptor.xsd">
    <Name>EXPAND NFI</Name>
    <Title>SLD for MapServer OGC Web Services</Title>
    <Abstract>This is an SLD to for NFI EXPAND LandCover</Abstract>
    <NamedLayer>
      <Name>nfi_lc_9</Name>
      <UserStyle>
        <Name>nfi_lc_9</Name>
        <Title>nfi_lc_9</Title>
        <IsDefault>1</IsDefault>
        <FeatureTypeStyle>
          <Rule>
            <ogc:Filter>
              <ogc:PropertyIsEqualTo>
                <ogc:PropertyName>VEG_TYPE</ogc:PropertyName>
                <ogc:Literal>TC</ogc:Literal>
              </ogc:PropertyIsEqualTo>
            </ogc:Filter>
            <PolygonSymbolizer>
              <Fill>
                <CssParameter name="fill">#00ff00</CssParameter>
              </Fill>
            </PolygonSymbolizer>
          </Rule>
        </FeatureTypeStyle>
      </UserStyle>
    </NamedLayer>
  </StyledLayerDescriptor>
```

Example 1. Style Layer Descriptor for forest cover.

With user defined SLD's and default themes, a WMS service can produce choropleth or themed maps and allows users to create customized legends or theme for spatial data. Cartographic choices, such as colours, line thickness and patterns can be controlled and simple analytical capabilities can also be added to the SLD's. For example, a SLD can

filter a spatial data set to only display those features that are of a specific domain. Example 1 shows the SLD's descriptions for polygons that have vegetation type (VEG_TYPE) as TC (Treed Polygon, Coniferous).

This simple method of analytical filtering can be used for model generalization. Brassel and Weibel (1988) and Joao (1998) describe model generalization as a filtering process. Analytical Generalization is never used for display purposes, but strictly for data reduction in order to obtain subsets of an original database for data analysis. Cartographic Generalization is used for graphical display, and it aims to improve the visual effectiveness and readability of a map (Joao 1998).

GIS and GIS users have in the past balanced both cartographic and model generalization to satisfy both display and analysis purposes. The WMS service has been created to handle both of these generalization mechanisms. However, the communities of practice using WMS have targeted the cartographic model recently. The use of WMS as a tool for analytical generalization has been supported with the capabilities of filters within the SLD definitions. This project uses the functionality of generalization in WMS for analysis.

The combination of WMS and SLD has allowed for the simple creation of images that represent spatial data. Many spatial operations have been implemented as Web based services. The National Forest Information System (NFIS Canada) project office at the Pacific Forestry Centre of the Canadian Forest Service has developed a number of Web based spatial operations based on the WMS specification. These operations rely on the standard mechanism of requesting and receiving data from standards based WMS service.

Distributed Spatial Analysis Architecture

This study required the ability to conduct spatial operations on data held on multiple datasets coming from different data stores or servers accessed through WMS. A set of spatial operations based on GIS and image processing was developed by the NFIS Project Office under the *Distributed Spatial Analysis Architecture* (DSAA). The DSAA

framework allows for the analysis of images created by WMS services and includes such functions as buffering, area calculations, median and mode filters, and overlay operations.

The data used in this study include provincial forest cover data and Agriculture Canada's Ecozones. These spatial data are stored on distributed spatial warehouses supporting WMS capabilities. An image overlay capability was required to calculate the intersection of the source data to that of the chosen global grid.

The image overlay service takes two images as inputs and calculates a new image that depicts the unique combinations of source images. This service also reports on the combinations of values of the source images and the number of occurrences of unique combinations. Each pixel in the source images is assigned a unique signature encoded Red, Green and Blue and an Alpha value. A summary report is generated giving a pixels or picture elements count for each unique resultant encoded value.

Figure 13 illustrates the image overlay service. It compares the pixel values of the two input sources for all pixel locations. The service scans through both input images and finds all unique combinations of both input sources and recreates a new image with unique colors for each combination. A resulting XML report gives the number of occurrences or pixels that have the same unique input value combination.



Figure 13. Image overlay example.

In Figure 20, the inputs are 3 x 3 pixel images. The first image has 6 pixels that are dark grey and the second image has four pixels that are lighter grey. The overlay service scans

through the 3 x 3 array and finds all unique combinations of colours from the input images. The resulting image represents three areas. The three areas are the dark gray, the lighter grey and the lightest grey areas. The light grey areas are those areas that were only light grey in the inputs, the dark grey areas are the areas that were originally dark grey from the input source and the lightest area is where the two dark and light greys were found. This overlay process is similar to that of many GIS overlay / intersection operations. This overlay process is described as the Cross-tabulation method (Klinkenberg, 2005). A Crosstab (Cross-tabulation) combines two grids into one resulting grid. Every possible combination of values is supported. If the source grids each have three classes, the possible combinations = 3×3 , or 9 classes. The result of the crosstab will contain a maximum of 9 classes.

Other DSAA operations have similar input requirements with a number of the services based on classic set theory and logic. The functions developed for the DSAA framework takes inputs from WMS images with other parameters needed for the functions being passed to the service. Functions such as buffering, area summaries; mode and mean filters operate in a similar manner as the overlay service.

Discussion

The bounding area and the pixel count requested from a WMS service is critical to the accurate representation of features in the source data. For this study, it was found that for a given area, having a larger image size or pixel count gave more accurate results due to the fact that each pixel represented a smaller area. Where the area represented by each pixel was large there was loss of small features and consequently the WMS service did not render the image accurately.

It is important to note that there are various WMS image output formats that dictate the extent to which the images are amenable to further analysis. WMS formats such as JPEGs are suitable for cartographic or web display and not suitable for further analysis. JPEG is an image compression standard written by the Joint Photographic Experts Group. This format and many like it are “lossy” formats where the decompressed images are not

the same as the source image. JPEGs are designed to exploit the restrictions of the human eye. For example, small color changes are perceived less accurately than small changes in brightness (Anon., 1999). JPEGs and many other formats with these characteristics are not suitable for the purposes of analysis within the DSAA framework.

For this study, the use of the Portable Network Graphic (PNG) format was selected. The PNG format offers a compression mechanism without losing image information (losslessly). Restoring and resaving an image does not degrade its quality, unlike standard JPEG and, in addition, PNGs supports 48-bit true colour and 16 bit gray scale, thus allowing for a larger palette of colours or variations of gray.

In order to apply the NFIS Distributed Spatial Analysis Architecture, it is essential that the inputs are choropleth map or images. A choropleth map is a map that uses different colors, patterns, and symbols to show different values over a map or space. For example, in order to map leading tree species from a forest cover layer (spatial coverage), each unique leading species would be given a unique colour. A map representation created from the WMS service must generate a choropleth to be used successfully in DSAA.

The SLD service is generally used to generate choropleth maps. However, SLD's have limitations and can be computationally intensive operations when applied to native spatial datasets. In the sampling process, a tessellation or grid of 400 x 400 five meter cells is created. Each of the 160,000 cells is assigned a unique identifier starting from 1 to 160,000. Each cell is given a colour representation from the hexadecimal RGB range of (00,00,01) to (02,71,00). To create a SLD that contained 160,000 style rules would require the service to read and apply these rules during each request. To create a less computationally expensive process, the tessellation was created as a MapInfo MID/MIF file with paint brush and fill colours set to the unique identifier. The MID/MIF format allowed for the creation of colour polygons directly applied to the geometry.

These MID/MIF files are served through a WMS server based on the University of Minnesota's MapServer project. MapServer is an OpenSource development environment

for building spatially enabled Internet applications. MapServer supports many of the OGC web services specifications and these capabilities are employed through WMS to publish the tessellations as a choropleth map.

Image requests against WMS service of differing sizes were conducted to understand the sampling inherited rasterization and sampling of spatial data required by the WMS operations. The test was conducted with the 400 x 400 grid cell that was published through the UMN Mapserver product. Each of the unique cells within the 400 x 400 grid was assigned a unique color. Requests for images at differing sizes but covering the same location or extent were conducted. Table 4 has a URL (a form of a URI) that was used to generate the results for a 1500 x 1500 pixel image within the bounding rectangle of (UTM Zone 7) x=652029 y=6933423 by x=654129 y=6935523. The total image requested by the request would represent an area of 441 ha (2100 m by 2100 m).

Table 4. URL for WMS request to create a 1500 x 1500 pixel image.

URL to request for GetMap	http://gis.2y.net/cgi-bin/expandmap.cgi?VERSION=1.1.1&REQUEST=GetMap&SRS=EPSG:26910&BBOX=652029.000000,6933423.000000,654129.000000,6935523.000000&WIDTH=1500&HEIGHT=1500&LAYERS=expand&FORMAT=image/png&EXCEPTIONS=text/html&QUALITY=QUICKEST
---------------------------	---

The expected number of colours is equal to the number of cells plus the white space around the grid area. The source data had 160,000 cells that were uniquely coloured with the surround being a separate colour. This results in 160,001 unique colours. Table 5 shows the results of multiple requests at differing sizes. It is also important to note that the 100 x 100 image is not large enough to incorporate 160,001 unique colours, but only a maximum of 10,001 colours.

Table 5. Unique colours at differing image sizes.

Size of image requested in pixels	Number of unique colours in resulting image
100x100	9315
500x500	160001
1000x1000	160001
1500 x 1500	160001
2000 x 2000	160001

These results showed that an image that is too small would not be able to contain the number of unique cells in the source 400 x 400 grid. However, having an image of 500 x 500 pixels, resulted in the proper number of unique colours, but the representation of each cell size was not always consistent. Each cell represents a 5x5 meter square that has an area of 0.0025 ha. Table 6 shows the discrepancy between calculated cell sizes from different image sizes. This shows that requested images that are too small did not represent the 0.0025 ha area cell size well. The images that were requested at a larger size created represented the original cell size better.

The cell sizes were calculated with a DSAA service known as the Pixel Area Service. This service calculates the Area in hectares (ha) of each unique colour combination within an image. The Pixel Area service takes the WMS map requests and calculates the area based on the projection, bounding area, and the resulting image. The apparent discrepancy in the 2100 x 2100 result showing a smaller standard deviation than the 3000 x 3000 result is explained by the fact that the 2100 x 2100 sample grids and total extent of the sample area are exactly congruent.

Table 6. Cell size summaries between image sizes.

Size of image requested in pixels	Minimum cell size (ha)	Maximum cell size (ha)	Average cell size (ha)	Standard deviation of cell sizes (ha)
100x100	0.044	0.088	0.044454	0.004444
500x500	0.002	0.011	0.002819	0.00125
1000x1000	0.002	0.005	0.002772	0.000692
1500 x 1500	0.002	0.004	0.002332	0.000474
2000 x 2000	0.002	0.003	0.002583	0.000493
3000 x 3000	0.002	0.003	0.002272	0.000445
2100 x 2100	0.003	0.003	0.003	0.000000

From the results in Table 4, the images that were created smaller than 1500 x 1500 had average calculated cell size larger than the source 0.0025 hectares. Images created at 1500 x 1500 to 3000 x 3000 had average cell sizes that were closer to that of the 0.0025 ha area. The largest cells size found in the 1000 x 1000 images were 0.005 ha and the images smaller than that found even larger cells.

When the request for the image is a 2100 by 2100 pixel image (one meter per pixel) the resulting cell sizes are uniform throughout the image. That is, the sizes of each cell are identical throughout the grid. However, our Pixel Area Service calculations shows that the area that each cell represented an area of 0.003 rather than the 0.0025 ha. This error could also be a result of the limitations of significant figures and possible rounding errors in the calculations of area through the Pixel Area Service of DSAA.

Summary

Web Services are essential to accessing and processing datasets held in a distributed server environment. Conforming to international standards and adopted specifications, such as those of the Open Geospatial Consortium's Web Mapping Service and Style Layer Descriptors, data can be requested in a standard way. These Web Services can

then be used as inputs to other Web Services that support analysis of the data held in raster formats. Caution must be taken in choosing the image sizes for analysis. Procedures applied to sampling, generalization, filtering and sampling are critical to preparing the data for spatial analysis. Requesting the correctly formatted raster data representation and ensuring that sufficiently large image area or pixel count will determine the accuracy of the analytical results.

Chapter 5 Spatial Data Handling

Introduction

The organization, storage, and analysis of data is vital when dealing with large volumes of spatial and non-spatial data and where fast query response time is required. ExPAND requires data from distributed locations to be gathered and analyzed together. The handling of this data from gathering, re-sampling, storing, archiving, and analyzing is essential. This chapter describes a method of decomposing spatial data into small cells that fall within a tessellation. The cells are stored in a relational database that provides an uncomplicated view of the data. By taking advantages of the many benefits of relational databases and *Structured Query Language* (SQL), spatial data can be easily added, updated, and accessed.

Egenhofer and Herring (1991a) reviewed the fundamental concepts necessary for the analysis of spatial data within spatial databases. Queries in a spatial database are based on the relationships among spatial objects. They outlined a formalized categorization of binary topological relations between regions, lines, and points. They also defined the need for the incorporation of spatial relations over geometric domains in spatial query language. Experimental query language supports queries with relationships. The spatial relationships that are described for the SQL Extensions (Egenhofer and Herring, 1991a, p2) include:

- adjacent,
- contains,
- contains point,
- enclosed by,
- intersect,
- near, and
- self intersect.

Application of the spatial operation of intersect is the prime focus of this study. Intersect can be thought of as the operation used for the keyword “within” or “by”. This study investigated the development of a simple solution to answer the questions related to the

intersection of two spatial objects. Standard SQL is used throughout the study without the extensions that Egenhofer and Herring (1991a) describe.

Source Spatial Warehouse

The NFI is in the process of loading all thematic data into the National Forest Information System. This project accessed a subset of the data that has been loaded to the NFIS spatial warehouse. The core NFI data themes are stored on a relational database with spatial capabilities. The spatial warehousing infrastructure for the NFIS initiative is based on the commercial relational database, Oracle. Along with Oracle, a spatial component the CubeWerx's Cubestor suite is the mechanism used for storing spatial data within Oracle. Cubestor is a spatial cartridge that allows for the efficient storage, retrieval, management, and indexing of spatial information in a relational database.

Large data warehouses rely on spatial indexing to allow for the quick access of information. Spatial indexing is one of the key factors involved in advances in geographical spatial warehouses. Langran (1993) describes indexes as references to logical or physical addresses that provide access to stored information. Spatial indexes are commonly partitions of the data within a geographical region. Closely located geographical features share a common spatial index or portion of an index. This use of indexing and tree technologies has seen companies such as Oracle, IBM, ESRI, and Canada's own CubeWerx develop products that address the large volumes of data that are involved in today's information management fields. Many such mechanisms exist such as Btrees, Rtrees, Quadrees etc. These index methods all share common characteristics, they keep information about the shape and position of features within the infrastructure.

The *Canadian Hydrographic Services* (CHS) required a means to store and manage large volumes of sounding data. No commercial product was available to handle the Canadian Government's requirements. HH-Codes or Helical Hyperspatial Codes were developed to meet this need. This indexing method allowed for the storage and retrieval of information based on a non-Euclidean geometry of cubes (Varma, 2000).

HH-Codes are a bit representation of 1 to n dimensional space (Varma, 2000). HH-Codes can index features in one, two, three or even four dimensional feature space. When sorted, HH-Codes cluster locations close to each other. In one dimension they cluster in a similar manner to binary trees. In two dimensions they cluster similar to quadtrees. A two dimensional HH-Code is the interlacing of two one dimensional HH-Codes. In three dimensions they cluster similar to quadtrees and is the interlacing of three one dimensional HH-Codes.

The HH-Codes is based on cubes that are subdivided by quarters. Each subdivided quarter contains the parent identification as part of its own index. This allows each component a reference to the cube to which it is a child. Figure 14 illustrates a two dimensional HH-Code representation, where the top right square is indexed with the code of 11. The lower right box is indexed with a code of 01 and the top right as 10 and the lower right as 00. Each box is given a unique code based on a binary representation. Each child within a box or subdivision is given the parents unique index code. The first partition or subdivision of the top right box would have all 10 as the prefix HH-code. All further subdivision would carry that 10 prefix. The children of the top right box would be 1000, 1001, 1010, 1011. This continues on until an arbitrary minimum box size has been reached. The major benefit of this mechanism is that it allows simple traversing of the data structure to find an element given the parent prefixes. It is easy to find or compare two elements to determine if they are in close proximity by looking at the prefixes of their HH-Code.

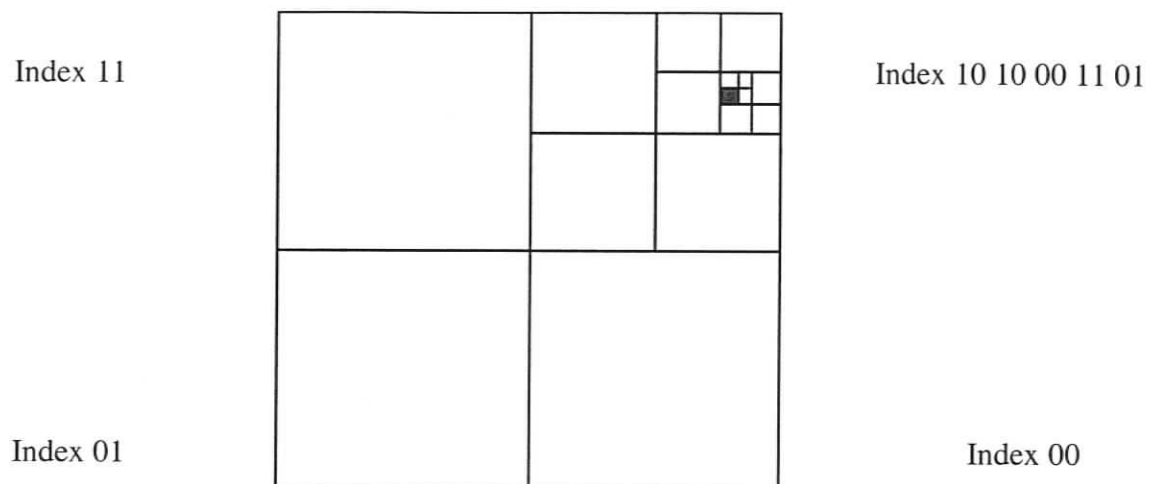


Figure 14. HH-Code index method.

HH-Codes allow Cubestor to index, retrieve, and manage vast volumes of spatial data within a relational database. The technology has allowed users to store raw geographical features or geometry efficiently and effectively. This spatial index optimizes performance used in spatial operations. Finding shapes or geometries are the most common uses of the indices. This index method is similar to those of regular database management systems uses of indexes when searching tables for information.

All data themes for this project (land cover, ownership, land use, protection status, and ecozones) were loaded into their separate respective tables in Cubestor or onto other OGC compliant WMS data stores. Individual tables were generated for each UTM Zone by each NFI theme. Even though the individual theme within each UTM zone are stored in one table, individual NFI plots and NFI version numbers can still be used to filter the data. Filter mechanisms are available with the use of the Style Layer Descriptor, OGC Filter Specifications or with the CubeWrx's CubeXplor client's WHERE clause.

Some reference data, such as the sampling cells that were generated as MID/MIFs are stored on a WMS data store outside of Oracle. These data stores are based on the open source package developed by the University of Minnesota (UMN) MapServer.

Data Access

All data stores used for this study are accessible through Web services as described in Chapter 4. The principle Web service used was the Open Geospatial Consortium *Web Mapping Service* 1.1.1 (WMS). All spatial data used in the study were located on data stores that are accessible through standard WMS requests.

The sampling grids (5 x 5 m cells) were published through the UMN Mapserver software. This sampling grid was created as a MapInfo MID/MIF file as the format allowed for the assignment of colours and pen strokes. Pen strokes are used to define the symbolization and representation of the geographical shape. Since there are 160,000 unique cells for each NFI photo plot, it was more efficient to colour the cells within the native GIS format than to use SLD technology.

Other data sets were accessed from WMS servers from across Canada. The EcoZones of Canada are published from Agriculture and Agri-Food Canada in Ottawa. The EcoZones provide a generalized framework containing 15 broad physiographical and ecologically similar areas (Wiken, 1986).

Data Projection

All source spatial data are stored in its native projection. In this study, the National Forest Inventory photo plot data was provided in UTM. The horizontal datum of the projection for the given data was North American Datum of 1983 (NAD83). The GRS80 ellipsoid with a semi-major axis of 6378137.0 m and a semi-minor axis of 6356752.3 m (Bugayevskiy and Snyder, 2000) was used.

Other national and regional datasets that are used in this project are in their native projections, including Lambert, and Albers. The WMS Web service as described in Chapter 4, allows data to be stored in a native projection but published in the projections defined by the WMS request if the appropriate transformation is available.

All re-projection and re-sampling errors introduced by the WMS Web service were considered to be out of scope for this project. It is recognized that vendors of WMS services rely on different mechanisms to render images from source spatial data and do not necessarily implement their re-sampling and re-projection software in the same manner. Differences or errors resulting from re-sampling and re-projection were not investigated.

Sampling and Storage

The majority of the source spatial data used in this investigation was stored and loaded in CubeStor/Oracle. A configuration of the UMN MapServer product publishes the tessellation required for the sampling process or “gridding”. The user defined grid size as described in Chapter 3 was used to sample the NFI data within a NFI Plot. Sampling was achieved with the use of the Distributed Spatial Analysis Architecture’s Overlay service. During the process of sampling, the colour value or Digital Number of the interrogation was stored in an Oracle table along with the NFI Plot Number/Identifier and cell identifier.

Representing spatial data in a raster form provides many advantages. Cells have a fixed number of neighbours and all cells are of constant size and shape (Gatrell, 1991). These characteristics are key to the reporting and analysis functions of this system as they provide properties that make some spatial analysis inherently simpler and easier. Overlaying of maps and logical operations are made less complicated when data is placed on the common cell / raster layout. Many of these operations can be achieved through simple summaries in a relational database. Knowing the cell size and shape allows for interrogation of the database and for summaries of areas. A query can be posed against the database to rapidly summarize the number of cells that met specific criteria. The number of cells meeting the specific criteria is used to calculate the total land area for that particular class.

The re-sampling process involved overlaying the input NFI Photo plot data on a common grid / tessellation. The NFI photo plots are located across Canada and are 2000 x 2000 meters in size. The re-sampling process was developed based on individual photo plots.

The tessellation grid chosen was a grid of 5 x 5 meter cells that lined up with the 2000 x 2000 meter cells resulting in each NFI photo plot having a 400 x 400 (5 meter) cell matrix that was used as the basis for sampling each of the coverages.

Sampling was accomplished with Web Services as described on Chapter 4. The data was accessed through WMS and analyzed with a Web image overlay service. This resulted in large XML files capturing all occurrences of values within each 5 meter grid. The XML document was sorted and the largest occurrences of a unique value within a given 5 meter grid were recorded. The resulting values were stored in a table within a database to show values generated from the re-sampling.

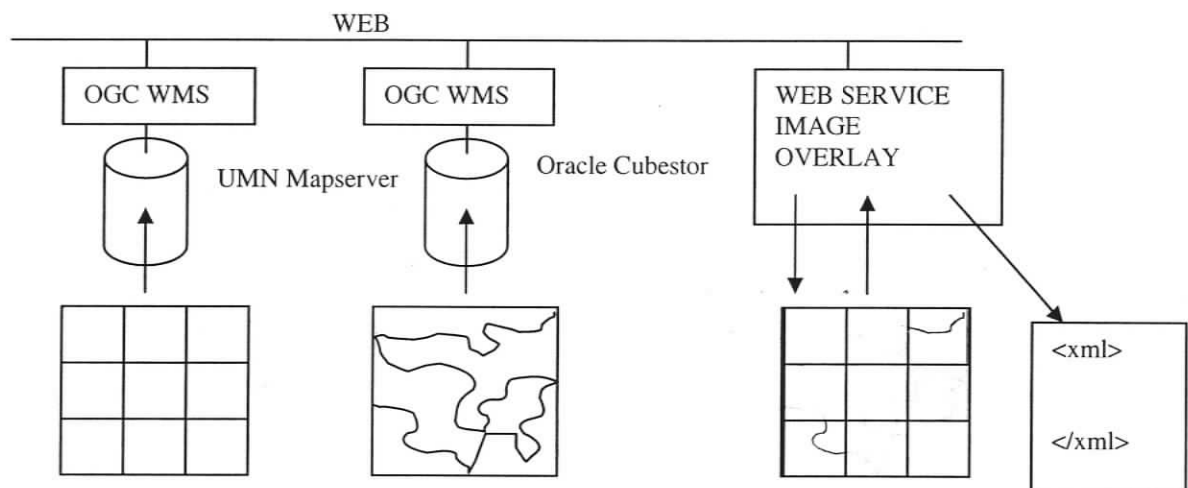


Figure 15. Data stores and Web services.

Figure 15 illustrates the re-sampling process data flow. The Web Service for image overlay was given the instruction to fetch the correct NFI theme data from Oracle Cubestor via WMS service. At the same time the image overlay process fetched the grid or tessellation cells (5 x 5 m) for that same location from the other input sources. Both requests were made through the OGC WMS GetMap request. All necessary parameter requests to the data stores are identical, specifically the projection, size of image, and location were identical between requests. The only real differences between the GetMap

requests was the information or layers to be returned. The input data to the image overlay service can originate from any OGC WMS compliant server. The Web service for image overlay generates an XML document with all combinations of the intersection. This document can then be stored within the database framework described below.

Raw sample data was stored in a large table (CELL_VALUES, Table 7) with references to the NFI Plot Identifier, the cell number within the plot, the particular theme upon which the sample was based and the sampled value or pointer (key) to the attribute file. The term themes refer to the NFI layers of: Forest Cover, Protection and Land Use. Appendix 1 provides a detailed entity relationship diagram of the relational database structure for this project.

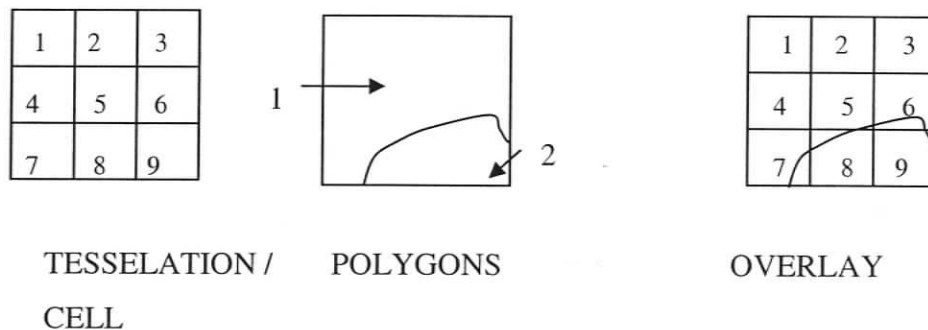


Figure 16. Re-sampling polygons onto a tessellation.

A systematic sampling of a data request through WMS GetMap was accomplished with a program that samples a given NFI Photo plot by a user defined cell size. All sampled data were entered in the CELL_VALUES table. Figure 16 illustrates the re-sampling process on a polygonal area. The image was overlaid with a tessellation grid of nine cells. The nine tessellation cells were assigned a value based to the largest polygon within the individual cell. In this example, cell one was assigned the polygon identified as 1, cell two was assigned polygon number 1 and cell number eight was assigned polygon id 2. Cell nine was assigned the polygon number 2. The CELL_VALUES table

contains information on these nine tessellation cells and the polygon that occupied the largest area of the cell. Table 7 shows how Figure 16 was portrayed in the CELL_VALUES database table. The column THEME_VALUE holds the polygon identifiers from Figure 16.

Table 7. Re-sampling database table.

CELL_NUM	THEME_VALUE
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	2
9	2

Langran (1993) describes several different cases of GIS mechanisms to describe dimensional dominance. This re-sampling to a common grid is considered a temporal grid. The temporal grid is the simplest geographic data structure to provide a suitable vehicle to introduce methods of clustering for dimensional dominance (Langran, 1993). This approach provides for a snapshot of the grid over time. She further states that the approach is also used for detecting changes between sequent Landsat images. Each image is stored separately and values are compared or clustered over time. The CELL_VALUES table is in fact, a form of storing a spatial grid within a relational database. In this case the CELL_VALUES table hold multiple grids from differing themes, space and time.

The CELL_VALUES table contains one record for each tessellation cell for each theme. In the case of a tessellation of 160,000 cells, the CELL_VALUES table have 160,000 records describing the cell contents. Obviously this table or set of tables will grow

dramatically after analyzing several tessellation units and themes. To expedite the queries and summaries to this table a new table containing summaries by tessellation collections is created.

CELL_SUMMARIES is a table that captures a summary for a given collection of tessellation cells. The unique condition of a cell value is recorded along with the occurrence of such value within the table. That is the number of times a value occurred in the CELL_VALUES table is recorded in this table. From Table 8 and Figure 3, the CELL_SUMMARIES shows that there were seven occurrences of polygon 1, and two occurrence of polygon 2. Table 8 shows the summary table and the number of reduced records as compared with Table 7.

Table 8. Summary Count database table.

THEME_VALUE	COUNT
1	7
2	2

From Table 8, it is possible to infer the area of a given theme (polygon identifier from Figure 5) value by the number of tessellation cells found since each tessellation cell is of equal area and of similar shape. The total area a given theme can be determined by multiplying the tessellation cell count by the area each cell represents. From Table 8, if the theme was described as using a cell size of 5 x 5 m (25 square meters) a quick calculation of the counts by the cell area gives the total area of a specified theme value. Theme value of covers had 50 square meters and theme value of 2 covers 25 square meters.

The intersection of multiple themes can be stored back into the CELL_VALUES table and summary statistics of that theme can also be stored in the CELL_SUMMARIES table. The THEME_DESCRIPTION table stores information about how an individual theme is derived and what the source of the theme.

The table TILE_THEME_IDENTIFIER holds information about the combinations of themes to tiles. In this study, a tile is considered to be a NFI 2 km by 2 km photo plot. The column TILE_THEME_ID is a key that uniquely identifies a particular theme (THEME_ID) with that of a NFI Photo plot (TILE_COLLECTION_ID). Table 9 gives an example of how the TILE_THEME_IDENTIFIER table stores information about a TILE and its corresponding THEME. The example in Table 9 shows that two combinations are recorded for the same plot (1534491). TILE_THEME_ID of 54 has information about THEME_ID 1 and TILE_THEME_ID 55 has information about THEME_ID 2.

Table 9. TILE_THEME_IDENTIFIER database table.

TILE_THEME_ID	TILE_COLLECTION_ID	THEME_ID
54	1534491	1
55	1534491	2

Another table, CELL_DESCRIPTION is created to store information about the tessellation cell. Information such as area size, width, and height of the cells are stored here. This table also records the number of rows and columns of the cells are required.

Table 10. CELL_VALUES table structure.

COLUMN NAME	COLUMN TYPE	EXAMPLE	DESCRIPTION
TILE_THEME_ID	VAR_CHAR	54	Key to Tile & Theme combination that describes this combination
CELL_NUM	INT	1	Sample cell location /grid reference as defined by cell description table.
THEME_VALUE	VAR_CHAR	62-C	Value of theme as described by theme description table.

The CELL_VALUES table (as described in Table 10) is constructed as each point cell is sampled. Table 11 shows a sample table of the NFI Photo plot number 1534491 that has been sampled for forest cover and ownership. Plot number 1534491 contains three different classified forest cover types. Three separate polygons with identifiers of 092B052_2994, 092B052_2998 and 092B052_2989 are recorded. Two unique ownership classes were found on this plot: 62-C and 62-N. Examination of Table 13 shows that cells # 81 and 82 both contain the same forest cover pointer and the same ownership descriptor. Cell numbers 83, 84 and 85 have common ownership but all have different forest cover pointers.

Simple SQL queries can be constructed to interrogate the CELL_VALUES table. These queries could summarize all the cells that contained the THEME_VALUE 62-C within the collection name of 1534491. More complicated queries could be constructed to find the number of cells that have 62-C in 1534491 under theme 54 that are no longer 62-C under another theme 56. Theme 54 would be the ownership status in 2003 and theme 56 would have been the ownership status in 2004. Many simple SQL queries can be created to give the results of temporal change, and of intersection of themed values.

Table 11. CELL_VALUES database table example.

TILE_THEME_ID	CELL_NUM	THEME_VALUE
54	81	62-C
54	82	62-C
54	83	62-N
54	84	62-N
54	85	62-N
...
55	81	092B052_2994
55	82	092B052_2994
55	83	092B052_2994
55	84	092B052_2998
55	85	092B052_2989
...
56	81	62-N

Example 2 gives a SQL expression that will return the number of occurrences of unique theme values from the CELL_VALUES table that was from the tessellation collection 54. The results of the query are similar to Table 8. This database design can be efficiently queried by taking advantage of the powerful SQL language.

```
SELECT theme_value, COUNT(*) FROM cell_values WHERE tile_theme_id = 54
GROUP BY theme_value
```

Example 2. A SQL statement to generate a simple summary.

Basic metadata about the THEME_ID and THEME_VALUE column is defined by the THEME_DESCRIPTION table. The Theme Value column can be either the actual value of the theme (categorical or ratio attribute) or the column can be used as a pointer (key field) to another database. Definitions of the theme are stored in the

THEME_DESCRIPTION table. Information concerning the theme and how it was generated along with the source information is captured in this table. Imbedded in this table is the database table to which a THEME_VALUE can be joined. Table 6 outlines the descriptions and information needed for the THEME_DESCRIPTION table.

Theme can be thought of as layers of a map. A theme can be restricted by attributes or space and/or time. A theme could be comprised of data that represents an ownership status in the year 2003. However, another theme could be the ownership status for a completely different year. Themes can also be created from subsets of existing themes. It is possible to create a theme based on variations of the attributes being rendered and sampled on the grid. This allows for a simple approach to temporal warehousing of spatial objects.

Table 12. THEME_DESCRIPTION table structure.

COLUMN NAME	COLUMN TYPE	EXAMPLE	DESCRIPTION
THEME_ID	INT	54	Unique Theme Identifier
THEME_NAME	VAR_CHAR	Ownership 2003	The name of the theme
THEME_DATE	VAR_CHAR	2003-05-06	Date the theme was sampled / generated
TILE_CLASS_ID	INT	5	Key to cell class description table. Describes the size of the individual sample cells and number of rows/cols etc.
THEME_VAL_TYP E	VAR_CHAR	K	The possible value types can be F Float I Integer C Character

			K Key to another table { } Set contained in bracket
THEME_VAL_MIN	VAR_CHAR		Value minimum range (integer or float type)
THEME_VAL_MAX	VAR_CHAR		Value maximum range (integer or float type)
THEME_NO_DATA_VAL	VAR_CHAR		NULL character
DESCRIPTION	VAR_CHAR	Ownership sample on 2003 for entire province	Description of the theme
THEME_SRC	VAR_CHAR		Source of THEME
THEME_EXT_DB	VAR_CHAR		External database connection string
THEME_URL_SRC	VAR_CHAR	https://ca.nfis.org/cubewerx/cubeserv/	WMS get map request

The sample cells used in this process are described in Table 13. The cell structure is described by an area or by the cell's width and height in meters. It also describes the number of rows and columns. Another table is used to describe the individual cell relationships to other cell classes. This table describes the relationship of parent and cell classes. Given a 25 meter cell class it allows the identification of which cells from a 5 meter cell class would be the children. The CELL_RELATIONSHIP table lists all the parent-child relationships for all cell classes.

Table 13. TILE_DESCRIPTION table structure.

COLUMN NAME	COLUMN TYPE	EXAMPLE	DESCRIPTION
TILE_CLASS_ID	INT	1	Tile Identifier
TILE_AREA	FLOAT	4000000	Area of tile (square meters)
CELL_AREA	FLOAT	25	Area of one cell (square meters)
CELL_SIZE_X	INT	5	Width of cell in meters
CELL_SIZE_Y	INT	5	Height of cell in meters
NO_ROWS	INT	400	Number of cells in tile
NO_COLS	INT	400	Number of columns in tile

A summary (CELL_SUMMARIES) table allows for the rapid access to frequency distributions within a given NFI Plot. Table 14 gives the column descriptions for such a table construction. The CELL_SUMMARIES table contains the key that links back to collection name identifier, along with the theme identifier. The table also includes the theme value and the summary count. The unique characteristics/attributes can be thought of as the intersection of all attributes within a common spatial geometry. This unique summary table can be considered to be similar to that of an intersection or union between multiple data sets.

Table 14. Cell Summaries database structure.

COLUMN NAME	COLUMN TYPE	EXAMPLE	DESCRIPTION
TILE_THEME_ID	INT	54	A theme describing the combination / intersection of multiple source theme data.
THEME_VALUE	VAR_CHAR	62-C	Unique occurrence of value from theme as described by theme description table.
VALUE_COUNT	INT	550	Number of cells that meet the criteria of the unique_theme_value

Table 15 gives an example of the Cell_Summaries table. This table is used to summarize the occurrences of values within the CELL_VALUES table. The example illustrates that for TILE_THEME_ID of 54 (NFI Plot 1534491 and Ownership) there are 550 cells that have values of 62-C. This summary table allows a user to quickly analyse or summarize areas for a given collection or for a given theme value. This table greatly reduces the number of records that Table 11 (Cell_Values example) would have.

Table 15. Cell Summaries example table.

TILE_THEME_ID	THEME_VALUE	VALUE_COUNT
54	62-C	550
54	62-N	380
54	62	300
57	NFI_8545	380
57	NFI_8545	550
57	NFI_9000	805

Spatial Operations

Spatial operations can be applied to the CELL_VALUES table with simple SQL queries. Figure 17 illustrates two spatial grids at the sample location and their corresponding records within the CELL_VALUES table. Tile Theme 1 is a numeric theme with values that are either 4 or 5. Tile Theme 2 is based on the characters A or B. The corresponding CELL_VALUES table shows the values of each theme for each cell. This table is used to record the values and the relative location of each cell within the spatial grid.

Tile Theme 1		Tile Theme 2		Cell_Values Table		
4	4	A	A	Tile_Theme_ID	Cell_Num	Theme_Value
4	5	B	B	1	1	4
				1	2	4
				1	3	4
				1	4	5
				2	1	A
				2	2	A
				2	3	B
				2	4	B

Figure 17. Two grids and the corresponding CELL_VALUES subset.

The power of SQL language allows for the creation of simple queries that result in sophisticated spatial-like operations. The partial CELL_VALUES table from Figure 17, can be used to count the number of unique occurrences. For example, Theme 1 occurs within Theme 2. The cross product or intersection of these two themes can be achieved with a SQL statement shown in Example 3. This SQL statement returns the combinations of unique pairs of themed values within two given separate themes in the same area.

```

SELECT DISTINCT t1.theme_value, t2.theme_value, count(*)
  FROM theme_values t1, theme_values t2
 WHERE t1.tile_theme_id = 1
 AND t2.tile_theme_id = 2
 AND t1.cell_num = t2.cell_num
 GROUP BY t1.theme_value, t2.theme_value;

```

Example 3. A SQL query to produce the “Cross Product” / Overlay results.

The SQL statement in Example 2 can easily be chained to other queries thus allowing for results from multiple themes. This capability allows the simultaneous processing of the intersections of multiple themes. The result of this query would return three rows with the combinations of the themed values with their occurrences. The combination (4,A) occurs twice, while (4,B) and (5,B) occur once each. Since each of the themes stored within the CELL_VALUES tables is described in the THEME_DESCRIPTION table the cell size is known. This in turn allows the area of each cell to be known and provides the required information to calculate the area that the unique combination of 4 and A occurrences. This simple SQL query can be used in a spatial context allowing for the intersection of spatial data without having complex and expensive GIS software.

Advantages and Disadvantages of Matrix Data

There are many advantages to employing this matrix data structure. The structure behaves in a similar manner to two-dimensional matrices where each pixel occupies specific locations and is given a value. It is a convenient and simple structure and is well defined mathematically. Egenhofer and Herring (1991) describe the many benefits of the two-dimensional matrices. They indicate that the matrices are mathematically well defined, simple to comprehend and simple to implement and are supported by many high level programming languages. However, they list several key disadvantages to this data structure. These disadvantages include (Egenhofer and Herring, 1991):

Detail – the matrix is represented in the same way, and no advantage is taken from existence of larger areas covered by the same type of pixel;

Abstraction – to get less detailed representations, all the data have to be considered and no different levels of abstraction can be achieved without checking all details;

Storage Capacity - potential waste of storage space;

Size of data sets – transmission of raster data is limited by the bandwidth of the transmission channel and matrix representation is an impediment when transferring collections of raster at high rate; and

Inefficiency - the whole image has to be kept in main memory which can be critical for very large images.

Many of Egenhofer and Herring's disadvantages were indeed apparent in this study. A number of approaches to addressing and resolving these limitations are addressed. The use of relational databases and large spatial data stores allows for the storage of matrix-like data. It is not necessary to store the entire image while processing a small sub area within a database. The size of data sets being transferred has been addressed by using Web Services thereby requesting only the exact area of interest. This study provides examples of the integration, analysis and synthesis of multiple data sets from large areas.

Discussion

The CELL_VALUES table needs to be optimized and indexed properly to reduce query times. During this study many different adaptations of the CELL_VALUES tables were investigated. The first adaptation of the CELL_VALUES table had the TILE_COLLECTION_ID and the THEME_ID column as part of the CELL_VALUES table. This made queries statement easier to create. It was possible to use the SQL WHERE clause and filter by TILE_COLLECTION_ID. However, the total query runtime resulted in a dramatically slower performance than when the CELL_VALUES table that did not contain TILE_COLLECTION_ID and THEME_ID.

A normalization exercise was needed to be undertaken on the CELL_VALUES table. A second adaptation of the CELL_VALUES table resulted in having a unique key that represented the combination of TILE_COLLECTION_ID and THEME_ID. The new

key (TILE_THEME_ID) was used as an identifier for the new table (TILE_THEME_IDENTIFIER) that stored information about the tiles and themes relationships. This latter table resulted in total query runtimes of over 10 times the speed of the first CELL_VALUES schema. The first test resulted in a runtime of 3.65 minutes and the second test resulted in 0.20 minutes. The tests were done with 480,000 records in the CELL_VALUES tables.

The CELL_VALUES table contain one record for each cell that has been sampled. Using a plot area of 2 x 2 km with a cell size of 5 x 5 meter results in 160,000 records. This is repeated for each theme that is added. This table can grow large after several sampling processes occur. A number of approaches to optimizing the response were considered. The first approach was to only have one parent or collection name per table. That is, only store the cell values for one given plot in a CELL_VALUES table. This results in a separate CELL_VALUES table for each collection or parent cell and requires a significant effort to manage of all the tables.

Further summary tables can be generated from the CELL_VALUES table or from the CELL_SUMMARIES table to allow for efficient access to specific criteria. Reporting by NFI Plot Identifier, provincial or territorial jurisdictions, or even reporting by ECOZONES can be optimized with further summary tables.

Several other tables are required to capture metadata about the theme description, cell descriptions, current theme and the relationships of parent and child cells. These tables store information that is required to properly interrogate the summary tables.

Summary

The handling of the geographically aware spatial data has been reduced to relational tables of sampled areas. This easily implemented approach reduces the requirement for heavy GIS-like spatial operations on databases, and expensive GIS services. Operations such as queries and overlays (see Chapter 4) can be achieved through straight forward SQL like operations on the CELL_VALUES table.

Data Handling is critical to the analysis of the sampled data. The data is stored in a manner that makes it easily retrievable, and easily stored. Differing themes, versions, and differing time series of sampled data for a given location is easily discovered, compared, and stored. The hierarchical nature of the tessellations is captured in the schema of the database. Each child cell has information that allows for the relationship of neighbours and of parent. The data stored within the database provides a design that allows the sampled data to be used for further analysis.

Chapter 6 National Forest Inventory - Use Case Scenarios

Background

ExPAND can be used as a tool to compile and analyze information from multiple data sources in a given or common location. ExPAND was created to allow for the access and analysis of data over a distributed network. To demonstrate these capabilities, several scenarios using Canada's National Forest Inventory were explored.

The Canadian Forest Service has published in the past several authoritative statements on the distribution of the forest resources. Authoritative statements and reports generated from the government rely on data to be processed in a common and methodical manner. Common questions and statements are made of Canada's forests. These include (CFS, 2004b):

- What is the total forest area within a given ecological or political area intersected by an ecological area?
- What is the area of forest type within a given ecological or political area intersected by ecological area?
- What is the area of forest types that is protected within a given ecological area?
- What is the area of age classes by protection status?

Many of these questions require compilations and statistical analysis to be conducted at reporting units. These reporting units are divisions of political provincial and territorial boundaries by that of ecological classification boundaries.

CanFI

To meet the needs of forest managers and policy makers, the CFS created *Canada's Forest Inventory* (CanFI) (CFS, 2004a). These compilations of forest statistics were compiled in 1981, 1986, 1991, and 2001. These compilations were based on provincial and territorial operational databases and inventories. Each jurisdiction submitted their best forest resource inventory as sources to the CanFI. The data was then aggregated to large tiles or mapsheets, provincial and national levels. These datasets were then used for

further analysis and reporting. The latest release of CanFI 2001, was created from the aggregation of 57 different inventory sources (Gillis *et al.* 2005).

Gillis *et al.* (2003) discuss limitations within the CanFI approach. The nature and rate of changes to resources along with forecasts and projections are some of the limitations of CanFI. Other limitations include that of inconsistencies in data capture and data management. Discrepancies in measurements and lack of standards were also limitations of CanFI. By addressing these issues, a new inventory was designed by the CFS and the Canadian Forest Inventory Committee. The new inventory is known as the National Forest Inventory.

National Forest Inventory

The purposes of NFI are to assess and monitor the extent, state and sustainability of Canada's forests in a timely and accurate manner (Gillis *et al.* 2005). NFI will provide data for Criteria and Indicators to monitor sustainable development and will provide information for climate change initiatives. International requirements from the Global Forest Resources Assessments can be addressed with data from NFI (Gillis *et al.* 2005). Indicators for forest health and socio-economic status can be correlated and compiled using the new inventory.

The new NFI is based on systematic sampling grid that covers the entire country. Ground plots and photo plots will be used to estimate area of attributes for species composition, volumes and density. These sample plots will be re-measured over a ten year cycle. The new inventory will allow for the accurate collection and storage of timely information on the extent and state of Canada's forest.

Ground plots, photo plots and remotely sensed data are used to capture basic attributes or values for criteria and indicators. Core to the design of this inventory are the following elements (NRCAN, 1999):

- A network of sampling points across the population
- Stratification of the sampling points by terrestrial ecozones

- Estimation of most area attributes from remote-sensing sources (photo plots) on a primary (large) sample
- Estimation of species diversity, wood volumes, and other desired data from (small) ground based subsample
- Estimation of changes from repeated measurements of all samples.

The inventory is based on a set of national standards that will allow for the collection of data in a consistent manner. By following these standards, a baseline of forest resources over time is created. To provide reliable area statistics, a minimum of 1% of Canada's land mass is to be surveyed (Gillis *et al.* 2005). Photo plots are the primary data collection tool for the National Forest Inventory. These photo plots are based on a 20 km systematic grid with plots being 2 by 2 km (Gillis *et al.* 2005). All spatial information given about a plot must be stored and archived for future use and analysis.

As part of the NFI initiative, photo plots will be systematically collected with conventional photo interpretations using medium scaled aerial photography, existing operational inventories and satellite imagery when necessary. Features are delineated and interpreted according to land cover classes and characteristics. Remotely sensed products are used where traditional aerial photographs are not suitable or available. Remotely sensed data will also be used to assess: change, the need to revisit a plot, and forest condition (Gillis *et al.* 2005).

Attributes estimated from aerial photographs include: area, land cover, forest type, age and volume of trees, disturbance activities, land-use changes (reforestation, afforestation and deforestation), mortality, access and human influence, and soil erosion (Gillis *et al.* 2003). Land-cover and biomass products are expected from remote sensing products where conventional aerial photographs are not available. The Earth Observation for Sustainable Development (EOSD) of Forests is one of these products (Wulder *et al.* 2003). EOSD is designed to provide complete coverage of the forested areas of Canada with satellite data at regular intervals to produce land-cover products. A goal of EOSD is to provide land cover products to capture change in forest conditions over time to support

national and international reporting requirements. The photo plots will be systematically re-measured to allow for the analysis of change over time. A re-measurement cycle over ten years will allow for the analysis over time.

NFI Data

The NFI project office receives all spatial data as either vector ESRI Shape Files or as raster GeoTIFF files. For each 2 x 2 km photo plot, the NFI Project Office receives a vector representation of each of the themes. The vector themes collected for the NFI project are: Land Cover, Ownership, Land Use, and Protection. Figure 18 illustrates how multiple themes are overlapped on the same plot area. Satellite classifications are received as GeoTIFF files. All spatial data is stored within the NFIS project office's spatial data warehouses. These data stores can be accessed through standard Web services as described in Chapter 4.

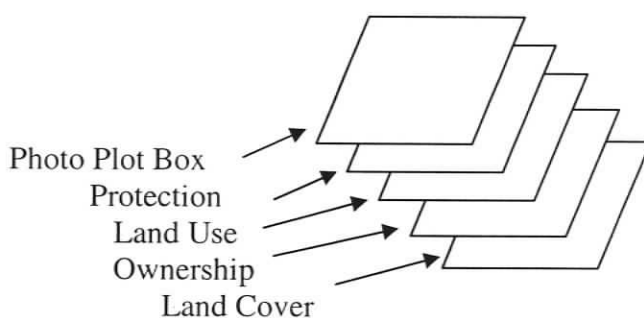


Figure 18. Themes / layers for a given 2 x 2 km NFI plot.

Table 16 lists the National Forest Inventory themes along with their appropriate theme codes used for this study. These theme codes would be used in conjunction with date and version for the THEME_DESCRIPTION table in Chapter 5. Descriptions of how the theme is generated are based on the NFI theme, date, version, and the sampling size. Other metadata is required to properly describe the theme being used for analysis.

Table 16. NFI theme names.

Theme	Theme Code
Land Cover	Lc
Ownership	Ow
Land Use	Lu
Protection Status	Ps
Box (plot boundary)	Pb

Along with the NFI themes described above, Agriculture and Agri-foods Canada's Ecozones and Natural Resources Canada's Provincial and Territorial Boundaries are necessary for reporting. Figure 19 illustrates Canada's 15 Terrestrial Ecozones. The Ecozones are an ecological mosaic of Canada on a sub-continental scale. These units are large and generally characterized by interactive abiotic and biotic factors (Wiken, 1986). These 15 units are the: Arctic Cordillera, Northern Arctic, Southern Arctic, Taiga Plains, Taiga Shield, Hudson Plains, Boreal Shield, Mixed Wood Plains, Atlantic Maritime, Taiga Cordillera, Boreal Cordillera, Pacific maritime, Montane Cordillera, Boreal Plains, and the Prairies.

Terrestrial Ecozones of Canada

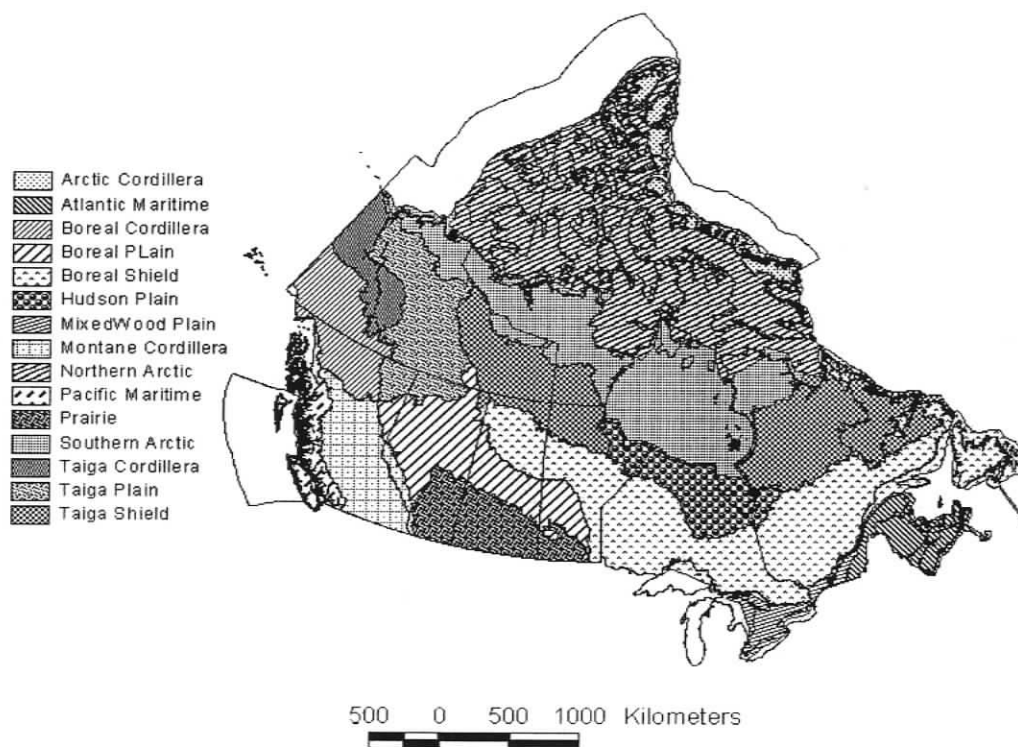


Figure 19. Terrestrial Ecozones of Canada. (GIS Source data: Ecological Stratification Working Group, 1995)

The Terrestrial Ecozones of Canada are published by Agriculture Agri-Food Canada in Ottawa. This data is accessible through standard Web Mapping Services that can be reached over the Web.

Compilation Requirements

The NFI Photo Plot data is compiled to create an amalgamation of layer attribution from differing sources of data onto the Land Cover polygons. Compilation will result in attribution based on averages or per hectare values for each Land Cover polygon within a NFI Photo Plot (CFS, 2004c). These attributes will be polygon-level. These attributes will be based on trees, layer origin layer treatment and layer disturbance. Attributes to be compiled may include that of volume, biomass, crown closure, site age, site height, site

index, afforestation size, area originating from planting, area originating from harvested, area regenerated from seeding, area that has significant fires, and others (CFS, 2004c).

Many of these compilations and statistical results will require the analysis of layers or sources of information from other datasets. ExPAND can be used to assist in these compilation requirements.

Use Case Scenario 1

In compiling data in accordance to the NFI requirements, a situation may occur when a 2 x 2 km photo plot may differ in map projection than that of data coming from another source. Source data sets may come as a national coverage, or may come as a separate 2 x 2 km photo plot in a UTM projection. For example, data from Agriculture and Agri-foods Canada may be in a Lambert projection with a central meridian in Ontario, but the NFI photo plot may be at the edge of UTM Zone 9. The techniques and processes laid out in this study will allow for the compilation and analysis of these two data sets. The requirement to analyze spatial data from differing map projection has been identified.

To demonstrate this scenario, a subset of the NFI photo plots was analysed. The subset used for this project included data from UTM Zones 9 and 10. Figure 20 illustrates the area where the NFI Photo Plots were chosen from between UTM Zones 9 and 10 and within the Pacific Maritime and the Montane Cordillera Ecozones.

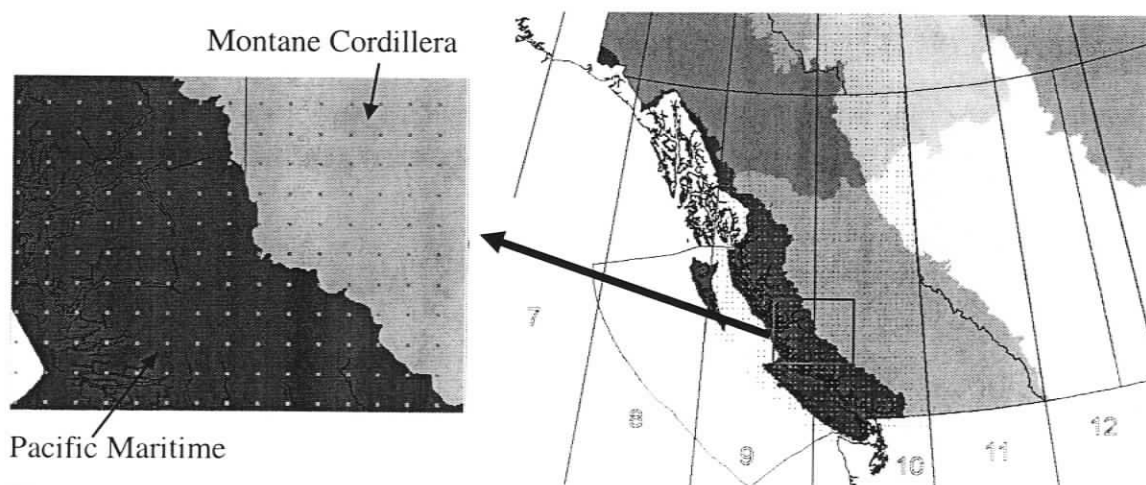


Figure 20. Photo plots within UTM Zones 9 and 10.

The eight selected NFI photo plots are listed in Table 17. These plots were specifically selected because the area they encompass span two UTM zones and two Ecozones. Two Ecozones was selected to demonstrate the project working with a seamless national coverage. Summaries were generated for each of the eight specific NFI photo plots.

Table 17. NFI photo plots in UTM Zone 9 and 10 within this scenario.

UTM Zone 9	UTM Zone 10
1389966	1389971
1396846	1396851
1403726	1403731
1410606	1410611

Process

Each of the eight 2 x 2 km plots had a 400 x 400 cell 5 m x 5 m grid file created. Each of the 160,000 5 meter cells was uniquely coloured. This grid file is used as the tessellation cell for re-sampling source data. The source data for this analysis was that of the Ecozones of Canada and that of the Land Cover and Ownership.

The Ecozones of Canada was sampled through the process and stored in a relational database table. Each photo plot was represented by the individual cells that comprised the parent tile. The database contained 160,000 records for each theme for each NFI photo plot tile.

The NFI Land Cover is based on the existing or current land cover. This classification system is hierarchical in nature. At the top of the hierarchy are two classes. The classes are vegetated or non-vegetated. At the second level, the classes are: vegetated treed, vegetated non-treed, non-vegetated land and non-vegetated water. The remaining hierarchies include that of cover type, landscape position and to density and sparseness (NRCAN 2004). At the lowest level of the hierarchical structure are the finer

attributions. Level six of the hierarchy would contain information concerning: genus, species, variety, height, and age.

ExPAND can report and analyze data from any classification scheme. For this study, the sampling of the Land Cover theme was based on the fourth level of the NFI Land Cover Classification System. The Vegetation type signifies the distinct type of vegetation or non-vegetated condition of the landbase within the polygon (NRCAN 2004). The classification has added a water code to this level to insure all polygons are classified and have a value.

There are 17 classes for the Vegetation Type classification at the fourth hierarchy level. Table 18 describes the classes used for this study. Each of the classes was assigned a unique colour. Each polygon was associated with a colour based on that polygon's vegetation type. The colour was assigned to the polygon based on a XML style layer descriptor. Appendix 2 has an example of a style layer descriptor for NFI plot 1389971. Basically, for each polygon, a defined rule colours the polygon based on the vegetation type. This creates a choropleth map as described in Chapter 4 that can be used for analysis with DSAA.

Table 18. Vegetation types from NFI Land Cover Classification System (Fourth Level). (NRCAN, 2004)

Classes	Description
TC	Treed polygons: Coniferous
TB	Treed polygons: Broadleaf
TM	Treed polygons: Mixed
ST	Vegetated, non-treed polygons: Shrub tall
SL	Vegetated, non-treed polygons: Shrub low
HE	Vegetated, non-treed polygons: Herb
HF	Vegetated, non-treed polygons: Herb Forb
HG	Vegetated, non-treed polygons: Herb Graminoid
BY	Vegetated, non-treed polygons: Bryoid
BM	Vegetated, non-treed polygons: Bryoid Moss
BL	Vegetated, non-treed polygons: Bryoid Lichen
SI	Non-vegetated polygons: Snow/Ice
RO	Non-vegetated polygons: Rock/Rubble
EL	Non-vegetated polygons: Exposed Land
WA	Water polygons
MI	Unreported/sample overlaps another jurisdiction
SA	Unreported/data is missing

During the re-sampling process the choropleth map is re-sampled against the 400 x 400 5 meter grid. A single colour is assigned to each of the 160,000 cells. The colours are then translated back to the original vegetation type classes. The cell information is then stored with the correct vegetation type classes in the CELL_VALUES table described in Chapter 5.

Example 4 shows how a simple SQL statement can give the results of the cross product of the Ecozones of Canada with that of the Vegetation Type from NFI Photo Plot number 1389966. The count results from the query are multiplied by a factor of 25 because the description of the theme tells us that the cells used for these themes are 25 square meters in area.

```
select distinct t1.theme_value, t2.theme_value, count(*)*25 as area
from expand.cell_values t1, expand.cell_values t2
where t1.tile_theme_id = 10
and t2.tile_theme_id =11
and t1.cell_num = t2.cell_num
group by t1.theme_value, t2.theme_value;
```

6 rows retrieved.

theme_value	theme_value	area
-----	-----	-----
Pacific Maritime	RO	66275
Pacific Maritime	SL	415200
Pacific Maritime	TB	163150
Pacific Maritime	TC	2888825
Pacific Maritime	TM	360050
Pacific Maritime	WA	106500

Example 4. Simple SQL Query for “Cross Product” of NFI Photo Plot 1389966 Ecozones and Land Cover.

Example 4 gives the results of the question “What is the area of Land Cover BY Ecozones for plot 1389966”. The question posed with the word “By” is mimicked into the query highlighted in Example 4. The results are the area in square meters for each Land Cover classification within the Pacific Maritime for the NFI Photo Plot of 1389966.

In Example 5, the same query was conducted on a different NFI Photo plot. The request was completed on TILE_THEME_ID 5 and 14. These TILE_THEME_ID’s are associated with the NFI Photo Plot of 1403731. The TILE_THEME_ID 5 is a key that references the plot 1403731 with that of Agriculture Canada’s Ecozones and ID 14 is that

same plot but the data is based on the NFI Land Cover Classification Scheme for Vegetation Type. The first column from the results is the Ecozone Classification and the second column for theme_value is the Vegetation Type. This Vegetation Type is the fourth level of the NFI Land Cover Classification (NRCAN, 2004). The code RO is defined as rock and rubble. The area column of the result will add up to exactly 400,000 square meters or 40 hectares. The selection statement is based on counting the number of records that fulfill the requirements. The resulting area is the number of records that match the requirements multiplied by the size that each cell represents. In this case, all cells are based on a 5 x 5 meter square. Each cell would represent 25 square meters. The count of records that fulfill the query is then multiplied by the size of the cell (25 square meters).

In this example, we have shown that data in one projection can be analyzed and compared with data in another projection. Data stored in one projection and published through OGC WMS services allow for the data to be analyzed in another well defined and known projection. This allows for the re-sampling to the common tessellation. The systematic re-sampling of the tessellation from the source data sets and the population of the relational database has allowed for the intersection and spatial analysis of the data from multiple map projections. With Web Services and the outlined spatial data handling procedures, one can analyze spatial data regardless of map projections.

Use Case Scenario 2

In addressing concerns regarding compilations for the National Forest Inventory, geographical units were created for reporting. Reporting units with the NFI consisted of the Ecozones and jurisdictional boundaries. An NFI Unit is defined as an Ecozones within the boundaries of a province or a territory (Gillis *et al.* 2005). With these requirements, one needs to know when a 2 x 2 km photo plot is intersected by an Ecozone or by a provincial or a territorial boundary. Figure 21 illustrates how NFI Photo Plots can fall onto two differing Ecozones. The requirement to recognize when a NFI Photo Plot is divided by two Ecozones is understood and necessary for compilation and analysis by NFI reporting units.

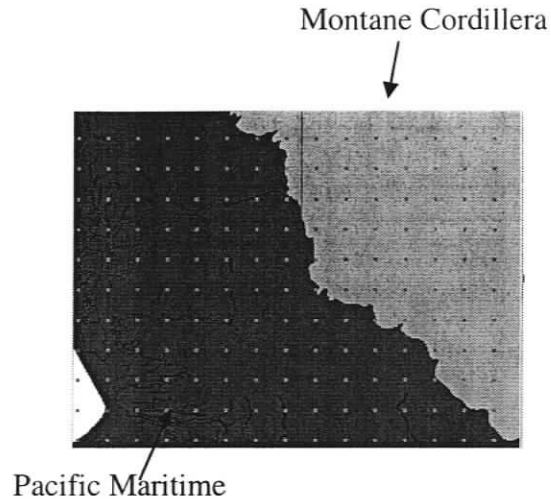


Figure 21. Subset of Photo plots within the Pacific Maritime and Montane Cordillera Ecozones

To demonstrate how ExPAND can be used to address this requirement, the same eight photo plots used in scenario 1 are analysed. This analysis will address how much of one source classification falls within another classification. For this scenario we are interested in knowing how much of the NFI Photo Plot's Land Class classification falls within the two Ecozones. This intersection or cross product can be calculated using the ExPAND techniques described. The same processes and data handling procedures are conducted.

Example 5 shows after the selection is accomplished a resulting table is given with information about Ecozones and Land Cover intersections. The resulting table details the breakdown of NFI Photo Plot 1403731 by Land Cover "By" Ecozones.

From the results of Example 5, we can tell that the Rock / Rubble (RO) classification existed in both the Montane Cordillera and the Pacific Maritime. Knowing the actual amount within each Ecozone of an individual classification will allow for the compilation and statistical analysis of those individual NFI Units. ExPAND will be a tool that will allow for the intersection or conditional selection of areas to a given Ecozone as described by the reporting units. This demonstration shows how ExPAND can be used to report on attributes within a geographical unit.

```

select distinct t1.theme_value, t2.theme_value, count(*)*25 as area
from expand.cell_values t1, expand.cell_values t2
where t1.tile_theme_id = 5
and t2.tile_theme_id = 14
and t1.cell_num = t2.cell_num
group by t1.theme_value, t2.theme_value;

```

2 rows retrieved.

theme_value	theme_value	area
Montane Cordillera	RO	3601950
Pacific Maritime	RO	398050

Example 5. Simple SQL Query for “Cross Product” of NFI Photo Plot 1403731 Ecozones and Land Cover.

With this example, the requirement to allocate estimations or compilation attributes based on intersection of secondary source information is addressed. This demonstration clearly shows how ExPAND can report on the amount of area of the intersection or cross product of two data sources. This capability is critical in processing the compilations required for the National Forest Inventory.

Considerations

For future work concerning the National Forest Inventory, a consideration of what value or attribute that is sampled should be given. Instead of creating a choropleth map based on the land cover classification level 4, one could create the colour mapping based on the unique polygon identifier. This will insure that all attributions and classification hierarchies are represented within ExPAND and any further analysis of the summarized data.

Instead of having one classification stored in the database, one would have the polygon identifier stored. This identifier will be a link or key back to the full description of the polygon. The analysis will not be just done on one level of classification but by any attribute linked or associated to that original polygon. The results of the intersections or

cross products would be reported by polygon identifiers rather than a classification. This would also allow for data types based on ratio to be stored and analyzed.

Comparisons with Traditional GIS

A comparison with traditional (Arc/Info) vector GIS system was conducted. This test involved downloading the data to a local workstation. The Canadian Ecozones was downloaded from an Agriculture Canada Web site. The data originally came in as an Arc/Info Export (E00) file in geographic coordinates. The data was then re-projected with double precision into a UTM Zone 10 projection. The new data set was built for polygon topology with a tolerance of 0.1 meters.

The NFI data sets for landcover were converted from an ESRI Shape File to an Arc/Info coverage. The Vegetation Type classes were added to the Arc/Info PAT file. This dataset was then intersected with the ecozones (Arc/Info command: INTERSECT bc_pp_1389966_pb utm_eco_9 eco_1389966 POLY). This Arc/Info command computes the geometric intersection of the two input data sets (NFI Photo Plot 1389966's Land Cover in UTM Zone 9 and utm_eco_9-Canadian Ecozones in UTM Zone 9) to create a new data set for comparisons.

The newly created data set contained Ecozone classifications and vegetation type (land cover classifications). Table 19 shows the areas for the given vegetation types from Example 1 and from that of traditional vector GIS operations.

Table 19. Comparison of traditional GIS with SQL example for photo plot # 1389966 (square meters).

Vegetation Type	Area from SQL Example 1	Area from Traditional GIS	Difference
RO-Rock Rubble	66275	66053	-222
SI-Snow Ice	415200	413394	-1806
TB-Treed Broadleaf	163150	162946	-204
TC-Treed Coniferous	2888825	2889631	806
TM-Treed Mixed	360050	361041	991
WA-Water	106500	106932	432
Total	4000000	4000000	0

Table 19 shows that the area summaries for vegetation types from the SQL example did not differ more than 0.5% from that of a traditional vector GIS. The largest difference came from the Vegetation Type of Snow and Ice. The Snow and Ice class had a difference of 1805.0938 square meters. That difference would be roughly 72 tessellation cells that were incorrectly classified out of approximately 16535 pixels for that class. The same test was conducted under Arc/Info on the NFI Photo Plot 1403731 in UTM Zone 10 with that of a re-projected Ecozones coverage. The resulting areas are shown in Table 20 with that of the results from the SQL Example 2.

Table 20. Comparison of traditional GIS with SQL example for photo plot # 1403731 (square meters).

Ecozone	Area from SQL Example 2	Area from Traditional GIS	Difference
Pacific Maritime	398050	398252	-202
Montane Crodillera	3601950	3601746	202
Total	4000000	4000000.0000	0

The differences in the areas of the two Ecozones classes are 202 square meters (or 0.0202 hectares). The total differences of the SQL Example with that of the Traditional GIS is 0. The total area of the SQL Example and the traditional vector approach resulted in the area of the plot area being the same.

The EXPAND process requires the tessellation of cells within a photo plot. This tessellation is defined before re-sampling and is designed to be hierarchical. The tessellations added together will always form the total area of the parent tessellation. In this case the 400 x 400 5 m cells will always add up to the 4,000,000 square meters (or 400 ha).

Errors

Some of the discrepancies can be found through commission errors while assigning the cell one value. The “rasterization” or “gridded” data approach has inherited errors. Reasons for the discrepancy of areas for the vegetation classes can be attributed to the size of the grid cells / pixel. During this study the smallest cell size or pixel area was a 5 x 5 m (25 square meters).

Wehde (1982) writes of the grid cell size in relation to errors in maps and inventories. He stated that as cell sizes were allowed to increase, the accuracies of maps and inventories

produced by computer processing decreased. Likewise, sample statistics (mean, mode and variance) of the interboundary distance at each cell size were found to decrease systematically with the increases in cell size. This suggests that a smaller tessellation cell would produce better results if the source data had a high enough precision and accuracy. The study used a 5 x 5 m cell for the smallest cell. The 5 x 5 m cell required 160,000 records in the database. A future study might look at creating 2 x 2 m cells or even 1 x 1 meter cells. This however, would result in a substantially larger database (1000000 to 4000000 records). The larger database would result in slower summaries and queries.

The differences shown on Table 19 and Table 20 can be accounted by other means also. The actual representation of the data through Web services (WMS) has inherit errors. These errors are a result of several processes within the WMS mechanisms.

Web Mapping Services possible errors include that of:

- Rasterization of polygonal features (creating choropleth maps)
- Cartographic generalization (Joao, 1998)
- Analytical generalization (Joao, 1998)
- Re-sampling errors (rotation)
- Re-projection errors

Congalton (1997) explores the consequences of vector-to-raster conversion. A summary of issues involved in the conversion was identified. The three issues that he summarized from were polygon fill, cut cells, and cell size. The polygon fill criteria were that a polygons region was rasterized with the correct classification value. The cut cells criteria were that one classification for a cell is assigned when the cell actually contains more than one class. The cell size is a balance between the size of the raster file and the quality of the conversion. Piwowar (1987) suggests that the optimum grid cell size should be one-fourth the size of the minimum polygon. In this study the size of the grid cells from the WMS is derived from the request for the image. The request for the image has the size of the image plus the information about the image's bounding coordinates. These together can be used to calculate the pixel size for a given WMS request.

The *National Center for Geographic Information and Analysis* (NCGIA) published a technical report in 1989 describing *Taxonomy of Error in Spatial Databases*. This report is a general overview of the concepts, methods, and models developed for evaluating spatial database accuracy. This report describes the area estimation problems, using dot grid density as a means to estimate area, along with other issues pertaining to area estimations. Generalization errors, choroplethic mapping, and data symbolization are also described. All these errors would be evident in the study conducted here. Further research into how much each of these errors played in the overall error in area estimation is unknown and out of scope for this project.

The sampling process used the “gridding” of the data can also introduce errors. The method of re-sampling for this study was achieved with the largest areal value. The weighted area sampling approach used for this study, forced the value of the cell to the polygon with the largest area within the tessellation cell. That is the largest continuous polygon within the 5 x 5 m cell will be assigned the re-sampled value for the entire cell.

Another approach for re-sampling on a weighted area may have been to assign the value of the largest collective area. The implementation used for this study was to assign the tessellation grid the value of the largest continuous polygon or colour. The largest polygon may not always be the largest summed area within the tessellation. Many small sliver polygons can actually add up to be larger than a continuous single polygon. This was not accounted for in this re-sampling implementation.

Summary

The intersection or cross product of any spatial layer can now be composed and run without the use of traditional expensive GIS products. The spatial operation “By” or “Within” can be achieved through simple SQL statements. This study shows that the spatial operations can be conducted with Web Mapping Services and Geospatial Web Services.

The spatial operations “By” or “Within” are functional equivalent with that of the ESRI Product Arc/Info’s command “IDENTITY, INTERSECT, and UNION”. The intersect function computes the intersection of two coverages (Anon, 1991). The SQL query mimics the Arc/Info functions for spatial intersections or the Cross-Tabulation of the input data sets. These spatial operations can now be achieved through simple SQL statements that can be implemented easily on any relational database.

The differences shown between the traditional GIS and the SQL approach was less than 1%. The area summaries and statistics generated from this study are within acceptable tolerance set out by the National Forest Inventory Project Office and the NFIS Project Office.

Further work and research could be conducted to reduce the area differences and to reduce other errors. Errors are introduced and compounded in many ways. The representation of the data through the Web Mapping Service to re-sampling and storing the tessellation grids are prone to errors. These errors need to be quantified in further studies.

Chapter 7 Conclusion

The objective of this thesis was to develop an efficient mechanism to analyze spatial data over a network of distributed data stores. This was achieved by utilizing existing international standards developed by the Open Geospatial Consortium. Research into methodologies and storage mechanisms along with spatial data handling was required to reach the objectives stated. The study included an example process on a subset of the National Forest Inventory.

In order to meet Canada's commitments to the international and regional forestry and environmental communities in a cost effective manner the Canadian Council of Forest Ministers undertook the development of a framework supporting the sharing and interoperability of their respective forest and land related resources information holdings. NFIS was developed to provide territories, provinces, and the federal government with secure Web-based access to their authoritative information holdings. The framework relies on the OGC interoperability and standards based services and applications for the access, integration, analysis, synthesis and reporting of information held on the partnering jurisdictions NFIS servers.

An examination of Canada's commitments, for example those of the Montreal Process, immediately identifies the requirement for spatial analyses of the information for reporting such indicators as area by forest type within protected areas.

The work undertaken in this study was designed to demonstrate the feasibility and benefits of being able to undertake spatial analysis operations on data held on a network of distributed servers.

While the emergence and widespread adoption of OGC standards and services have provided unparalleled access to spatial information holdings services development has primarily addressed the seamless integration and viewing of data rather than analytical capabilities. Conventional GIS spatial analysis has generally required that the various data sets be downloaded to a common repository where the information very quickly

becomes dated and may no longer reflect the authoritative information holdings of the information providers. Additionally, there are real costs to the data custodians to service information requests.

This study was designed to investigate approaches and procedures that would allow for the spatial analysis of data held on a distributed network of servers. The conceptual model was to design an approach that supported Web-based spatial analysis of custodial information holdings housed by independent jurisdictions in their own environment but linked to a distributed network supporting OGC protocols and standards. The work undertaken concentrated on developing approaches and procedures addressing the spatial functions described as “by” or “within”. The equivalent operation would be that of the intersection operation. The system developed is entitled ExPAND, *Expanding Primary Attributes for National Decision Making*.

Investigations included testing various re-sampling and spatial tessellation techniques, developing and testing OGC compliant Web Services for data access and spatial analysis and also procedures for information handling and storage in a distributed server architecture.

Retention of data integrity is essential in selecting a re-sampling method with a requirement that no new classes are creating during re-sampling. The method chosen for re-sampling was based on a weighted area allocation, which met the re-sampling requirement. This method shares many similarities to nearest neighbour, but is based on the largest portion or area feature within the sampled area as opposed to the nearest feature. No new classes were created.

Tessellation allows for the symmetrical coverage of an area without overlaps or gaps and is used as a common base or geometry within which re-sampling can be undertaken. Its application allowed for comparisons of re-sampled classes using a common base. A tessellation was constructed based on 5 x 5 meter squares grids in a UTM projection. Re-sampling was undertaken based on these localized tessellations.

EXPAND requires that the geographical data be published through a common interface or Web service. The Web service known as the OGC Web Mapping Service provides a standardized request protocol for accessing geographical data. Further components of WMS such as creating choropleth representations based on SLDs allowed the data to be analyzed from a distributed network of servers. Other Web services such as the Distributed Spatial Analysis Architecture's image overlay service were used as a mechanism to re-sample the data to a common tessellation.

Effective and efficient spatial data handling was identified as being critical to the success of the project. The use of relational databases, optimized handling, and storage of spatial data allowed for the access and analysis of the re-sampled data. Simple Structure Query Language statements were created to interrogate the database to answer questions relevant to the data. . The tessellation cells and the hierarchical relationships between cells was adopted and incorporated into the data model within the relational databases. This spatial data handling method incorporated relational databases, along with that of Web services to allow for the systematic sampling and reporting of a given area.

Study results based on a sub-set of the 2 x 2 km National Forest Inventory photo plots clearly demonstrated that the approach and procedures developed satisfied the original study requirements including the following:

- The method must allow for data to be held in a distributed architecture.
- The method must allow data to be held in the original projection.
- The method must allow data to be stored in original formats.
- The method must work in conjunction with the NFIS.
- The method must work with data from the National Forest Inventory.

EXPAND supports the intersection spatial operation within a properly formed SQL script. These scripts allow for the compilation and summaries of spatial data with relative ease and can be easily extended to allow for multiple datasets to be intersected together. Queries, analysis, and summaries were generated based on the ecological and land cover

types described in the National Forest Inventory. The results of the spatial operations were compared to traditional GIS analysis. These comparisons showed area differences of between 0.1 and 1.0 percent.

The capacity of NFIS to support the territorial, provincial and national reporting and compilations on the state and condition of Canada's forests and land related resources and sustainable forest management is being significantly enhanced through the development of ExPAND.

These capabilities are moving Canada to a position where it can provide authoritative reporting of information such as (Omule *et al.* 2004):

- What is the total forested area within a known area?
- What is the area of forest types within a known area?
- What is the area of forest type by age within a known area?
- What is the area forest types by protection status within a known area?

These questions are posed and responded to in publications and reports such as the State of Canada's Forests and the State of the Environment. The research undertaken under this thesis can be used to answer these questions.

Future Study and Development

Further research is required to reduce the errors found in the summary area reports. The source of the errors are described in Congalton's (1997) paper on consequences of Vector-to-Raster, Wehde's (1982) work on the cell size, and the compilation publication by the National Center for Geographic Information and Analysis titled "Taxonomy of Errors in Spatial Databases" addresses many of the issues and errors.

Further normalization exercises on the relational database schemas used by ExPAND may reduce query response time. Currently, the large volume of records negatively impacts query times. Each record represents one tessellation cell. Better indexing and management of these tables may result in faster responses to queries.

The Distributed Spatial Analysis Architecture's overlay process and the re-sampling processes need to be incorporated into a single Web service. Currently there are many steps involved before the individual tessellation cells are loaded in the database. Merging of the separate services into a single service will reduce the processing time, and thus reduce the time to process.

On resolution of the above issues ExPAND will be moved from a pilot development to a fully operational implementation in support of reporting on sustainable forest resources management in Canada.

Future Applications

ExPAND is based on open standards that have been created to access Web Mapping in a standardized method. ExPAND allows for the easy access, re-sampling, storage, and analysis of spatial data from any WMS compliant server. ExPAND can be implemented on practically any relational database. Limitations to the operational use of ExPAND lie in the size of the smallest tessellation cell and the size of the database being used. Within these limitations ExPAND can be used as an analysis tool on any infrastructure that relies on interoperability through common Web services such as WMS.

ExPAND's quick method of spatial intersections can be used as a tool for reporting on the state and condition of Canada's forests. The methodology can also be used for similar reporting requirements where data is located at distributed locations using WMS as a publication mechanism. This approach allows jurisdictions to maintain and manage their information. The custodians of the data have full control of the access to that data and information. This architectural approach reduces the chance of analyzing stale or out-of-date data and can significantly reduce the resources required to service requests for data downloads.

ExPAND can be used as a tool for validation work, time series analysis, wildlife, terrain modeling, and other applications. The validation of differing scales and resolution mapping products can be achieved through ExPAND. Differing mapping products with

compatible classifications can be validated and compared. ExPAND can be used as a calibration tool when high resolution data sources are compared with that of lower resolution data sources in the same geographical location. The higher resolution data can be used to calibrate and verify lower resolution data with the reporting and analytical capabilities of ExPAND. The simple intersection operation described as an SQL statement can be used as a means to answer many of the questions that are asked. Many other applications can be created around the ExPAND concepts and designs.

References

- Agresti, A. & Finlay, B. (1999). *Statistical Methods for the Social Sciences* Third Edition. Upper Saddle River, New Jersey: Prentice Hall.
- Anon. (1991). *ARC Command Reference*. Environmental Systems Research Institute, Inc.
- Anon. (1998). *ESRI Shapefile Technical Description*. Environmental Systems Research Institute, Inc.
- Anon. (1999). *JPEG image compression FAQ, part 1/2*. Retrieved September 2003, from <http://www.faqs.org/faqs/jpeg-faq/part1/>.
- Anon. (2004). *ArcGIS OGC Interoperability Add-on User Guide*. Environmental Systems Research Institute, Inc.
- Avery, T. & Berlin, G. (1985). *Fundamentals of Remote Sensing and Airphoto Interpretations* Fifth Edition. Don Mills, Ontario: Maxwell Macmillan Company.
- Berners-Lee, T., Fielding, R., Irvine, U. & Masinter, L. (1998). *Uniform Resource Identifiers (URI): Generic Syntax RFC:2396*. Retrieved October 21, 2005, from <http://www.ietf.org/rfc/rfc2396.txt>.
- Blacknell, D. (1989). *The Resampling of Synthetic Aperture Radar Images*. In S. Brooks (Eds.), *Mathematics in Remote Sensing*, Oxford, England: Clarendon Press, 173-188.
- Bernstein, R. (1983). *Image Geometry and Correction*. Chapter 21 In R. N. Colwell, (Eds.), *The Manual of Remote Sensing*, Bethesda, MD: American Society of Photogrammetry, 1:873-922.
- Brooks, D.R. (1981). *Grid Systems for Earth Radiation Budget Experiment Applications*. (NASA Technical Memorandum 83233. 40p.).
- Bugayevskiy, L.M., & Snyder, J.P. (2000). *Map Projection A Reference Manual*. Philadelphia, PA.: Taylor & Francis.
- CCFM (1997). *Criteria and Indicators of Sustainable Forest Management In Canada*. (Canadian Council of Forest Ministers, Ottawa, Ontario. 137p.).
- CCFM (2003). *National Forest Strategy 1998-2003 Sustainable Forests*. (Canadian Council of Forest Ministers, Ottawa, Ontario. 47p.).
- CCFM (2004). *National Forest Information System (CCFM-NFIS): An Overview* (Canadian Council of Forest Ministers, Victoria, British Columbia. 7p.).

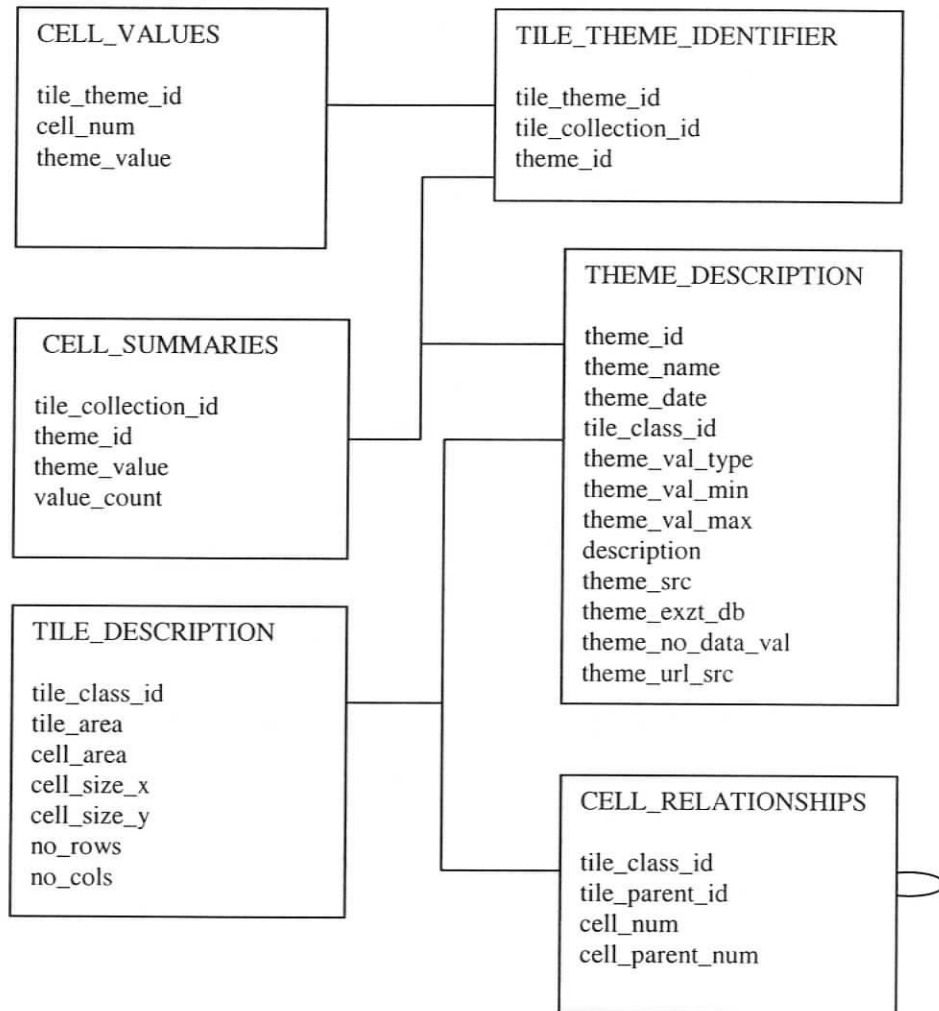
- CFS Canadian Forest Service (2004a). State of Canada's Forests 2003-2004. (Natural Resources Canada, Ottawa, Ontario. 93p.).
- CFS Canadian Forest Service (2004b). Canada's National Forest Inventory Design Overview (Natural Resources Canada, Victoria, British Columbia.13p.).
- CFS Canadian Forest Service (2004c). Canada's National Forest Inventory National Standards for Photo Plots Compilation Procedures (Natural Resources Canada, Victoria, British Columbia. 14p.).
- Congalton, R. G. (1997). Exploring and Evaluating the Consequences of Vector-to-Raster and Raster-to-Vector Conversion. *Photogrammetric Engineering & Remote Sensing*, 63(4), 425-434.
- Coppock, J.T., & Rhind, D.W. (1991). The History of GIS. In D. Maguire & M. Goodchild & D. Rhind (Eds.), *Geographical Information Systems*. (pp.21-43) Essex:Longman Scientific & Technical, Longman Group UK Ltd.
- Cracknell, A.P. (1998). Synergy in remote sensing – what's in a pixel? *International Journal of Remote Sensing*, 19(11),2025-2047.
- De La Beaujardiere, J. (NASA) Editor, (2001). Web Map Service Implementation Specification, Version 1.1.0, OpenGIS Project Document 01-047r2.
- Doyle, A. (International Interfaces, Inc.) Editor, (2000). Web Map Server Interface Implementation Specification, Version 1.0.0, OpenGIS Project Document 00-028.
- Egenhofer, M. J. (1991). Extending SQL for Graphical Display. *Cartography and Geographic Information Systems*, 18(4), 230-245.
- Egenhofer, M. J. (1994). Spatial SQL: a query and presentation language. *IEEE Transaction on Knowledge and Data Engineering*, 6(1), 86-95.
- Egenhofer, M. J. & Herring, J. R. (1991a). Categorizing Binary Topological Relations Between Regions, Lines, and Points in Geographic Databases. (Department of Surveying Engineering, University of Maine, Orono. Tech Report 28p.).
- Egenhofer, M. J. & Herring, J.R. (1991b). High-level Spatial Data Structure for GIS. In D. Maguire & M. Goodchild & D. Rhind (Eds.). *Geographical Information Systems*. (pp.227-237) Essex:Longman Scientific & Technical, Longman Group UK Ltd.
- Ehlers, M. (1997). Rectification and Registration. In J.Star and J. Estes and K. McGwire (Eds.), *Integration of Geographic Information and Remote Sensing*, Cambridge, United Kingdom:Cambridge University Press, 13-36.

- Fisher, P. (1997). The pixel: a snare and a delusion. *International Journal of Remote Sensing*, 18(3), 679-685.
- Gatrell, A.C. (1991). Concepts of Space and Geographical Data. In D. Maguire & M. Goodchild & D. Rhind (Eds.), *Geographical Information Systems*. (pp.269-297) Essex:Longman Scientific & Technical, Longman Group UK Ltd.
- Gillis, M. (2001). Canada's National Forest Inventory (responding to current Information needs). *Environmental Monitoring and Assessment*, 67, 121-129.
- Gillis, M., Gay, S., Clarke, D. & Power, K. (2003). Canada's National Forest Inventory: What can it tell us about old growth? *The Forestry Chronicle*, 79(3), 421-427.
- Gillis, M., Omule, A., & Brierley, T. (2005). Monitoring Canada's forests: The National Forest Inventory. *The Forestry Chronicle* 81(2), 214-221.
- Goodchild, M.F. (1994). Geographical grid models for environmental monitoring and analysis across the globe. *Proceedings of GIS/LIS '94 Conference*, Pohenix, Arizona.
- Graham, S., Simeonov, S., Boubez, T., Davis, D., Daniels, G., Nakamura, Y. & Neyama, R.(2002). *Building Web Services with Java*. Indianapolis, Indiana: Sams Publishing, 8-9.
- Joao, Elsa Maria (1998). *Causes and Consequences of Map Generalisation*. London, England: Taylor & Francis Ltd.
- Kimerling, A. J. et al. (1999). Comparing Geometrical Properties of Global Grids. *Cartography and Geographic Information Science*, 26(4), 271-288.
- Klinkenberg, Brian (2005) Crosstab and Reclass. Retrieved October 21, 2005, from <http://www.geog.ubc.ca/courses/klink/java/reclass.html>.
- Kreger, Heather (2001). *Web Services Conceptual Architectures (WSCA 1.0)*. Retrieved September 2003, from <http://www-306.ibm.com/software/solutions/webservices/pdf/WSCA.pdf>
- Kurz, W., Apps, M., Bandfield, E. and Stinson, G. (2002). Forest carbon accounting at the operational scale. *The Forestry Chronicle*, 78(5), 672-679.
- Lalond, W. (CubeWerx Inc.) Editor, (2002). *Styled Layer Descriptor Implementation Specification, Version 1.0.0*, OpenGIS Project Document 02-070.

- Lampropoulos, G., Gigli, G., Bardas, A., Armenakis, C. & Low, B. (2003). "Spatio-temporal Change Detection and Web-Based Distributed Computing Tools", Proc. of SPIE's 48th Annual Meeting 2003 (AM03), Vol 5205, San Diego, CA, 3-8 August 2003.
- Lampropoulos, G., Li, Y., Badas, J., Chan, H., McNairn, H. & Low, B. (2003). "Web-Based Distributed Processing Tools for Crop Classification Using CGDI Databases", Proc. of SPIE's 48th Annual Meeting 2003 (AM03), Vol. 5203, San Diego, CA, 3-8 August 2003.
- Langran, G. (1993). *Time in Geographic Information Systems*. London, England: Taylor & Francis.
- Lee, Y.C., Li, Z.L. & Li, Y.L. (2000). Taxonomy of space tessellation. *ISPRS Journal of Photogrammetry & Remote Sensing*, 55 139-149.
- Marshall, I.B. & Schut, P. H. (1999). *A National Ecological Framework For CANADA – Overview*. A cooperative product by Ecosystems Science Directorate, Environment Canada and Research Branch, Agriculture and Agri-Food Canada.
- Mather, P. M. (1987). *Computer Processing of Remotely-Sensed Images: An Introduction*. Chichester, England: John Wiley & Sons.
- Mattikalli, N.M., Deverux, B.J. & Richards, K.S. (1995). Integration of Remotely Sensed Satellite Images with a Geographical Information System. *Computer & Geosciences*, 21(8), 947-956.
- Montreal Process Liaison Office (1997). *Canada's Report on the Montreal Process Criteria and Indicators for the Conservation and Sustainable Management Of Temperate and Boreal Forests*. (The Montreal Process Liaison Office, Ottawa, Ontario. 142p.).
- Natural Resources Canada. (1999). *A plot-based national forest inventory design for Canada*. (Canadian Forest Service, Pacific Forestry Centre, Victoria, British Columbia. 70p.).
- Natural Resources Canada. (2004). *Canada's National Forest Inventory National Standard for Photo Plots Data Dictionary*. (Canadian Forest Service, Pacific Forestry Centre, Victoria, Working paper. 35p.).
- Okabe, A., Boots, B., Sugihara, K. & Chiu, S. (2000). *Spatial Tessellations Concepts and Application of Voronoi Diagrams Second Edition*. West Sussex, England.: John Wiley & Sons Ltd.

- Piwowar, J. (1987). Conversion between Vector and Raster Format Data for Geographic Information Systems Applications, Thesis for Masters of Arts, Waterloo, Ontario, Canada.
- Song, L., Kimerling, A. & Shar K. (2002). Developing an Equal Area Global Grid By Small Circle Subdivision. In M. Goodchild and A.J. Kimerlings (Eds.), Discreet Global Grids. Santa Barbara, CA. USA; National Center for Geographic Information & Analysis.
- Unser, M., Albroubi, A. & Eden M. (1995). Enlargement or Reduction of Digital Images with Minimum Loss of Information. IEEE Transactions on Image Processing, 4(3), 247-258.
- Varma, H. (2000). Applying Spatio/Temporal Concepts to Correlative Data Analysis. In Wright & Bartlett (Eds.), Marine and Coastal Geographic Information Systems. (pp75-94) London; Philadelphia, Taylor & Francis Ltd.
- Veregin, H. (1989). Taxonomy of Error in Spatial Databases National Centre for Geographic Information and Analysis. University of California, Santa Barbara (Technical Report 89-12).
- Weibel, R., (1992). Models and experiments for adaptive computer-assisted terrain Generalization, Cartography and Geographic Information Systems, 19(3), 133-53.
- Weibel, R. & Heller, M. (1991). Digital Terrain Modelling. In D. Maguire & M. Goodchild & D. Rhind (Eds.). Geographical Information Systems. (pp.269-297) Essex:Longman Scientific & Technical, Longman Group UK Ltd.
- Whede, M. (1982). Grid cell size in relation to errors in map and inventories produced by Computerized processing, Photogrammetric Engineering and Remote Sensing, 48(8), 1289-98.
- Wiken, E.B. (1986). Terrestrial EcoZones of Canada. Ecological Land Classification Series No. 19. (Lands Directorate, Environment Canada. 26p and map).
- Wolberg, G. (1990). Digital Image Warping. Los Alamitos, California: IEEE Computer Society Press.
- Wulder, M., Dechka, J., Gillis, M., Luther, J., Hall, R., Beaudoin, A. & Franklin, S. (2003). Operational mapping of the land cover of the forested area of Canada with Landsat data: EOSD Land cover program. The Forestry Chronicle, 79(6), 1075-1083.

Appendix 1
Database Table Relationship



Appendix 2 Style Layer Descriptor Example

```

<?xml version="1.0" encoding="utf-8" standalone="no"?>
<StyledLayerDescriptor version="1.0.0"
xmlns="http://www.opengis.net/sld"
xmlns:ogc="http://www.opengis.net/ogc"
xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.opengis.net/sld
http://schemas.opengis.net/sld/1.0.0/StyledLayerDescriptor.xsd">
<Name>ExPAND NFI</Name>
<Title>SLD for MapServer OGC Web Services</Title>
<Abstract>This is an SLD to for NFI ExPAND LandCover</Abstract>
<NamedLayer>
  <Name>nfi_lc_10</Name>
  <UserStyle>
    <Name>nfi_lc_10</Name>
    <Title>nfi_lc_10</Title>
    <IsDefault>1</IsDefault>
    <FeatureTypeStyle>
<Rule>
  <ogc:Filter>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>POLY_ID</ogc:PropertyName>
      <ogc:Literal>1389971_00000650</ogc:Literal>
    </ogc:PropertyIsEqualTo>
  </ogc:Filter>
  <PolygonSymbolizer>
  <Fill>
    <CssParameter name="fill">#00ff00</CssParameter>
  </Fill>
  </PolygonSymbolizer>
</Rule>
<Rule>
  <ogc:Filter>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>POLY_ID</ogc:PropertyName>
      <ogc:Literal>1389971_00000654</ogc:Literal>
    </ogc:PropertyIsEqualTo>
  </ogc:Filter>
  <PolygonSymbolizer>
  <Fill>
    <CssParameter name="fill">#900000</CssParameter>
  </Fill>
  </PolygonSymbolizer>
</Rule>
<Rule>
  <ogc:Filter>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>POLY_ID</ogc:PropertyName>
      <ogc:Literal>1389971_00000661</ogc:Literal>
    </ogc:PropertyIsEqualTo>
  </ogc:Filter>
  <PolygonSymbolizer>
  <Fill>

```

```
<CssParameter name="fill">#00ff00</CssParameter>
</Fill>
</PolygonSymbolizer>
</Rule>
<Rule>
  <ogc:Filter>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>POLY_ID</ogc:PropertyName>
      <ogc:Literal>1389971_00000714</ogc:Literal>
    </ogc:PropertyIsEqualTo>
  </ogc:Filter>
  <PolygonSymbolizer>
    <Fill>
      <CssParameter name="fill">#00ffff</CssParameter>
    </Fill>
  </PolygonSymbolizer>
</Rule>
<Rule>
  <ogc:Filter>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>POLY_ID</ogc:PropertyName>
      <ogc:Literal>1389971_00000716</ogc:Literal>
    </ogc:PropertyIsEqualTo>
  </ogc:Filter>
  <PolygonSymbolizer>
    <Fill>
      <CssParameter name="fill">#00ff00</CssParameter>
    </Fill>
  </PolygonSymbolizer>
</Rule>
<Rule>
  <ogc:Filter>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>POLY_ID</ogc:PropertyName>
      <ogc:Literal>1389971_00000722</ogc:Literal>
    </ogc:PropertyIsEqualTo>
  </ogc:Filter>
  <PolygonSymbolizer>
    <Fill>
      <CssParameter name="fill">#00ff00</CssParameter>
    </Fill>
  </PolygonSymbolizer>
</Rule>
<Rule>
  <ogc:Filter>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>POLY_ID</ogc:PropertyName>
      <ogc:Literal>1389971_00000731</ogc:Literal>
    </ogc:PropertyIsEqualTo>
  </ogc:Filter>
  <PolygonSymbolizer>
    <Fill>
      <CssParameter name="fill">#900000</CssParameter>
    </Fill>
  </PolygonSymbolizer>
</Rule>
<Rule>
```

```
<ogc:Filter>
  <ogc:PropertyIsEqualTo>
    <ogc:PropertyName>POLY_ID</ogc:PropertyName>
    <ogc:Literal>1389971_00000741</ogc:Literal>
  </ogc:PropertyIsEqualTo>
</ogc:Filter>
<PolygonSymbolizer>
  <Fill>
    <CssParameter name="fill">#00ff00</CssParameter>
  </Fill>
</PolygonSymbolizer>
</Rule>
<Rule>
  <ogc:Filter>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>POLY_ID</ogc:PropertyName>
      <ogc:Literal>1389971_00000748</ogc:Literal>
    </ogc:PropertyIsEqualTo>
  </ogc:Filter>
  <PolygonSymbolizer>
    <Fill>
      <CssParameter name="fill">#00ff00</CssParameter>
    </Fill>
  </PolygonSymbolizer>
</Rule>
<Rule>
  <ogc:Filter>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>POLY_ID</ogc:PropertyName>
      <ogc:Literal>1389971_00000769</ogc:Literal>
    </ogc:PropertyIsEqualTo>
  </ogc:Filter>
  <PolygonSymbolizer>
    <Fill>
      <CssParameter name="fill">#008000</CssParameter>
    </Fill>
  </PolygonSymbolizer>
</Rule>
<Rule>
  <ogc:Filter>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>POLY_ID</ogc:PropertyName>
      <ogc:Literal>1389971_00000779</ogc:Literal>
    </ogc:PropertyIsEqualTo>
  </ogc:Filter>
  <PolygonSymbolizer>
    <Fill>
      <CssParameter name="fill">#008000</CssParameter>
    </Fill>
  </PolygonSymbolizer>
</Rule>
<Rule>
  <ogc:Filter>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>POLY_ID</ogc:PropertyName>
      <ogc:Literal>1389971_00000790</ogc:Literal>
    </ogc:PropertyIsEqualTo>
```

```

</ogc:Filter>
<PolygonSymbolizer>
<Fill>
  <CssParameter name="fill">#00ff00</CssParameter>
</Fill>
</PolygonSymbolizer>
</Rule>
<Rule>
<ogc:Filter>
  <ogc:PropertyIsEqualTo>
    <ogc:PropertyName>POLY_ID</ogc:PropertyName>
    <ogc:Literal>1389971_00000799</ogc:Literal>
  </ogc:PropertyIsEqualTo>
</ogc:Filter>
<PolygonSymbolizer>
<Fill>
  <CssParameter name="fill">#00ff00</CssParameter>
</Fill>
</PolygonSymbolizer>
</Rule>
<Rule>
<ogc:Filter>
  <ogc:PropertyIsEqualTo>
    <ogc:PropertyName>POLY_ID</ogc:PropertyName>
    <ogc:Literal>1389971_00000800</ogc:Literal>
  </ogc:PropertyIsEqualTo>
</ogc:Filter>
<PolygonSymbolizer>
<Fill>
  <CssParameter name="fill">#00ffff</CssParameter>
</Fill>
</PolygonSymbolizer>
</Rule>
<Rule>
<ogc:Filter>
  <ogc:PropertyIsEqualTo>
    <ogc:PropertyName>POLY_ID</ogc:PropertyName>
    <ogc:Literal>1389971_00000801</ogc:Literal>
  </ogc:PropertyIsEqualTo>
</ogc:Filter>
<PolygonSymbolizer>
<Fill>
  <CssParameter name="fill">#008000</CssParameter>
</Fill>
</PolygonSymbolizer>
</Rule>
<Rule>
<ogc:Filter>
  <ogc:PropertyIsEqualTo>
    <ogc:PropertyName>POLY_ID</ogc:PropertyName>
    <ogc:Literal>1389971_00000802</ogc:Literal>
  </ogc:PropertyIsEqualTo>
</ogc:Filter>
<PolygonSymbolizer>
<Fill>
  <CssParameter name="fill">#00ff00</CssParameter>
</Fill>

```

```

    </PolygonSymbolizer>
  </Rule>
  <Rule>
    <ogc:Filter>
      <ogc:PropertyIsEqualTo>
        <ogc:PropertyName>POLY_ID</ogc:PropertyName>
        <ogc:Literal>1389971_00000805</ogc:Literal>
      </ogc:PropertyIsEqualTo>
    </ogc:Filter>
    <PolygonSymbolizer>
      <Fill>
        <CssParameter name="fill">#008000</CssParameter>
      </Fill>
    </PolygonSymbolizer>
  </Rule>
  <Rule>
    <ogc:Filter>
      <ogc:PropertyIsEqualTo>
        <ogc:PropertyName>POLY_ID</ogc:PropertyName>
        <ogc:Literal>1389971_00000808</ogc:Literal>
      </ogc:PropertyIsEqualTo>
    </ogc:Filter>
    <PolygonSymbolizer>
      <Fill>
        <CssParameter name="fill">#008000</CssParameter>
      </Fill>
    </PolygonSymbolizer>
  </Rule>
  <Rule>
    <ogc:Filter>
      <ogc:PropertyIsEqualTo>
        <ogc:PropertyName>POLY_ID</ogc:PropertyName>
        <ogc:Literal>1389971_00000817</ogc:Literal>
      </ogc:PropertyIsEqualTo>
    </ogc:Filter>
    <PolygonSymbolizer>
      <Fill>
        <CssParameter name="fill">#00ff00</CssParameter>
      </Fill>
    </PolygonSymbolizer>
  </Rule>
  <Rule>
    <ogc:Filter>
      <ogc:PropertyIsEqualTo>
        <ogc:PropertyName>POLY_ID</ogc:PropertyName>
        <ogc:Literal>1389971_00000847</ogc:Literal>
      </ogc:PropertyIsEqualTo>
    </ogc:Filter>
    <PolygonSymbolizer>
      <Fill>
        <CssParameter name="fill">#00ff00</CssParameter>
      </Fill>
    </PolygonSymbolizer>
  </Rule>
</FeatureTypeStyle>
</UserStyle>
</NamedLayer>

```

```
</StyledLayerDescriptor>
```